Challenges and perspectives of hate speech research

Strippel, Christian (Ed.); Paasch-Colberg, Sünje (Ed.); Emmer, Martin (Ed.); Trebbe, Joachim (Ed.)

Erstveröffentlichung / Primary Publication
Sammelwerk / collection

This book is the result of a conference that could not take place. It is a collection of 26 texts that address and discuss the latest developments in international hate speech research from a wide range of disciplinary perspectives. This includes case studies from Brazil, Lebanon, Poland, Nigeria, and India, theoretical introductions to the concepts of hate speech, dangerous speech, incivility, toxicity, extreme speech, and dark participation, as well as reflections on methodological challenges such as scraping, annotation, datafication, implicity, explainability, and machine learning. As such, it provides a much-needed forum for cross-national and cross-disciplinary conversations in what is currently a very vibrant field of research.

https://doi.org/10.48541/dcr.v12.0
digitalcommunicationresearch.de

ISSN 2198-7610
ISBN 978-3-945681-12-1

Strippel et al.
Challenges and Perspectives of Hate Speech Research
Volume 12

Christian Strippel
Sünje Paasch-Colberg
Martin Emmer
Joachim Trebbe

Challenges and Perspectives of Hate Speech Research

Digital Communication Research.de

12

Digital Communication Research

Edited by Martin Emmer, Ulrike Klinger, Merja Mahrt, Christina Schumann, Monika Taddicken & Martin Welker

Volume 12

Christian Strippel, Sünje Paasch-Colberg,
Martin Emmer & Joachim Trebbe

Challenges and Perspectives of Hate Speech Research

Editorial office of *Digital Communication Research*
Roland Toth, M.A.
Freie Universität Berlin
Institute for Media and Communication Studies
Garystrasse 55
14195 Berlin, Germany
info@digitalcommunicationresearch.de

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available at http://dnb.d-nb.de.

ISSN 2198-7610
ISBN 978-3-945681-12-1

Persistent long-term archiving of this book is carried out with the help of the Social Science Open Access Repository (SSOAR) and the university library of Freie Universität Berlin (Refubium).

DOI 10.48541/dcr.v12.0

A printed version of this book can be ordered with the publisher Böhland & Schremmer Verlag Berlin: www.boehland-schremmer-verlag.de

Berlin, 2023
digitalcommunicationresearch.de

Table of Contents

III. METHODOLOGICAL PERSPECTIVES: OPERATIONALIZATION, AUTOMATION AND DATA

Digital
Communication
Research.de

Recommended citation: Paasch-Colberg, S., Strippel, C., Emmer, M., & Trebbe, J. (2023). Sharing is caring: Addressing shared issues and challenges in hate speech research. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 11–22). Digital Communication Research. https://doi.org/10.48541/dcr.v12.1

Abstract: This book is the result of a conference that could not take place. It is a collection of 26 texts that address and discuss the latest developments in international hate speech research from a wide range of disciplinary perspectives. This includes case studies from Brazil, Lebanon, Poland, Nigeria, and India, theoretical introductions to the concepts of hate speech, dangerous speech, incivility, toxicity, extreme speech, and dark participation, as well as reflections on methodological challenges such as scraping, annotation, datafication, implicity, explainability, and machine learning. As such, it provides a much-needed forum for cross-national and cross-disciplinary conversations in what is currently a very vibrant field of research.

DOI 10.48541/dcr.v12.1

Sünje Paasch-Colberg, Christian Strippel,
Martin Emmer & Joachim Trebbe

Sharing is Caring

Addressing shared issues and challenges in hate speech research

1 Introduction

This book is in some way an unplanned outcome of a research project that we worked on in the past five years.[1] When we started in October 2017, online hate speech had been an increasingly important issue in both public and academia for quite some time already. However, our project coincided with a socially and politically turbulent time, which challenged hate speech research and called for an increased exchange in the field. For example, the Network Enforcement Act came into force in Germany at that time. This law not only caused debate about how to identify criminal content in the volatile interactive spaces of the Internet and about who should be responsible for regulating these spaces, but it has also been

1 The interdisciplinary research project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners" was funded by the German Federal Ministry of Education and Research [grant number: 01UG1735AX]. It brought together communication scholars from Freie Universität Berlin, computer scientists from the Berliner Hochschule für Technik, and computer linguists from VICO Research & Consulting.

used to justify the introduction of restrictive social media laws in autocratic states and flawed democracies. Thus, it renewed questions about contextual factors in our thinking about norms and boundaries in public debates.

Other examples that strongly affected our and others' research were Facebook's decision to restrict its API after the Cambridge Analytica scandal came to light in early 2018, and the General Data Protection Regulation (GDPR) of the European Union, which unsettled many blog operators, and eventually led to the closure of their comment sections.

To respond to these developments and the implications they had for our research, we invited a group of colleagues working on similar topics to a workshop at the Berlin Weizenbaum Institute in 2019 to share experiences with common theoretical, conceptual, and methodological issues in our field of research.[2] We discussed questions around data collection, protection and exchange, identification and classification of norm-transgressive user-generated content, as well as data analysis and automation. One important outcome of this workshop was the realization that considering perspectives from different political and cultural contexts, as well as from different academic disciplines, is crucial to better understand hate speech as a global and multifaceted phenomenon. Furthermore, the exchange confirmed how important debates around theoretical concepts and definitions are for the growing and transdisciplinary field of research on hate speech.

With the aim to further address these points together with a broader group of people, we planned an international and interdisciplinary conference on hate speech analysis for mid March 2020 in Berlin. In addition to a few invited presentations, our main idea for this conference was to provide space and opportunities for in-depth discussions and exchange among the participants. The large number of registrations we received from scholars from many different countries and disciplines showed that there is indeed a great interest in such discursive formats within the community of hate speech researchers. Unfortunately, the conference had to be canceled a few weeks before it was scheduled due to the onset of the COVID-19 pandemic.

2 The participants of the workshop were (in alphabetic order): Arndt Allhorn, Chris Biemann, Svenja Boberg, Ines Engelmann, Katharina Esau, Annett Heft, Dominique Heinbach, Jakob Jünger, Tim König, Constanze Kuechler, Sebastian Kuehn, Laura Laugwitz, Wiebke Loosen, Alexander Löser, Hanna Marzinkowski, Teresa Naab, Pablo Porten-Cheé, Cornelius Puschmann, Liane Reiners, Susanne Reinhardt, Diana Rieger, Julian Risch, Tim Schatto-Eckrodt, Anke Stoll, Betty van Aken, and Marc Ziegele.

Since we were determined that it was important to provide a forum for research-related discussions amongst hate speech scholars, we decided to organize this volume and reached out to a number of scholars, who had registered for our then-canceled conference, as well as colleagues from the closer environment of our research project to contribute. To do justice to all the discussions we have missed in the panels and coffee breaks of the conference, we asked these colleagues for short programmatic papers that question current research threads, point out new ways, and give impulses for future research. In addition, we invited texts that respond to these papers as well as discuss and contextualize them in relation to each other.

To our great pleasure, almost all colleagues accepted our invitation, and those who did liked the assignment, confirming that they too see a need for this kind of exchange. As a result, we could realize an even more diverse authorship and hopefully have a bigger outreach than the conference would have been able to. We are excited that, with a total of 26 chapters, we can now cover a wide range of topics that contribute to the field of hate speech research by (1) focusing on recent research and policy developments in countries that are less visible in literature, (2) discussing the multiplicity of theoretical concepts, definitions, and measurements, and (3) presenting new approaches of interdisciplinary research and machine learning that come with new questions, challenges, and implications.

2 Political perspectives: Current issues and developments

The first section of this volume opens with contributions dedicated to the foundations of hate speech research. One of these foundations is that the assessment of speech as hate speech is context-dependent, for example, with respect to the legal and political framework in which the public discourse takes place. This fact comes with issues of generalizability and comparability of findings and touches concerns of specific biases in international hate speech research. In particular, the issue of a Western bias of contemporary social research also manifests in the field of hate speech research (Matamoros-Fernández & Farkas, 2021). As in many other research areas, much more resources go into research on hate speech in the US, Europe or East Asian countries than in countries of the Global South. For this reason, we aimed to include perspectives from Non-Western researchers into this volume to have a better picture of global hate speech research.

However, context is not the only cause of blind spots in hate speech research. Insufficient definition, conceptualization and operationalization of the phenomenon in question also contribute to this issue. Hate speech legislation or automated text analysis software often simply work on the basis of a binary "hate / no hate" logic, which does not reflect the various shades on the continuum of problematic and disruptive speech. Thus, some authors in this section aim to advance our understanding of hate speech and its variants from different perspectives, providing theoretical conceptualizations or recommendations for more thorough methodological approaches.

As a start, *Afonso de Albuquerque* and *Marcelo Alves* analyze the specific conditions under which the Bolsonaro family in Brazil managed to build a social media-based ecosystem that combined strategies of disinformation, fake accounts and hate speech to support Jair Bolsonaros finally successful campaign for presidency. In their comprehensive account of the situation in Brazil, the authors highlight both national peculiarities and general tendencies of the evolution of hate speech in the context of political campaigns.

Zahera Harb adds the perspective of Lebanon, a country strongly impacted by severe confrontations of ethnically-defined political groups. Using the events around the explosions in the Beirut harbor, she widens the perspective to the role of journalists in the distribution of hate speech in society. In her study, she shows that in Lebanon many journalists do not have a differentiated understanding of hate speech and often spread hate messages amongst (legitimate) criticism of politicians. The difficult political situation of the country, which is mirrored in public discourse, requires a very thorough definition and understanding of hate speech and its consequences.

Using a feminist campaign in Poland as an example, *Dagmara Szczepańska* and *Marta Marchlewska* are exploring the boundaries between hate speech and offensive and vulgar language as means to attract attention and start a discourse in society. From their national background, they contribute to the debate about a context-sensitive definition of hate speech. It is not an expression or a term per se that constitutes hate speech, so their argument, but whether it is used—as in the example of the All-Poland Women's Strike—to point towards abuse and raise awareness for a societal problem or to attack an isolated group aiming at degrading their dignity and incite violence against them.

Anna Litvinenko connects to the preceding texts problematizing contextual factors by providing a theoretical categorization of different levels of context. Opening up a spectrum between situational and sociocultural contexts, she refers to the problem of a too simple black-and-white understanding of hate speech, which is not only part of many scientific approaches but also of current legislation. Such shortcomings can seriously harm anti-hate speech measures, for example by negatively affecting free speech, which is why she argues in favor of more context-sensitive approaches both in science and regulation.

Issues with regulatory interventions against hate speech are also in the focus of *Tomiwa Ilori*. In his example and from a legal perspective, the practical conflict between the prevention of hate speech and the violation of freedom of expression becomes apparent. Referring to the Nigerian context, but also including the wider approach of the African Commission on Human and People's Rights, he discusses alternative approaches to countering hate speech while preserving citizens' right to free speech.

A crucial field of fighting hate speech, both promising and potentially harmful, is the subject of *Sana Ahmad*, who takes a closer look at the internal content moderation policies of social media platforms. While many of us still hope that platforms sorting out negative content may be a solution for hate speech, disinformation and other sorts of content, her study on content moderation workers and sub-contractors in India puts the spotlight on moderation processes and working conditions as relevant contextual factors in the ecosystem of anti-hate speech actors and strategies. Connecting to the organizational layer of context outlined by Anna Litvinenko before, working conditions and power relations appear as important factors for the effectiveness of anti-hate speech measures.

The first section concludes with a text by *Christian Schemer* and *Liane Reiners*. Written as a response to the articles above, their contribution focuses on questions of comparability of hate speech studies from a basic, methodological perspective. The two authors discuss various aspects of concepts like the core term "hate speech", sampling and operationalization. As contexts of research are always quite different by nature, they argue that functional equivalence should be the goal in comparative hate speech research. However, they do not focus on comparative research alone but on hate speech research in general, which needs to produce findings that can be interpreted across studies to produce progress in our understanding of the phenomenon.

3 Theoretical perspectives: Terms, concepts and definitions

Taking up the question of which concepts we should work with in our research, the second part of this volume is devoted to the multiplicity of terms and definitions in the field of hate speech research and its neighboring strands. There are two main motivations behind this focus: First, we have a growing set of concepts competing in the broader field, but only little discussion of the implications and issues related to this inflation of terms and definitions (Sellars, 2016, p. 4). Accordingly, we see the need for a broader conversation about the theoretical and empirical contributions of each concept. How can we balance the demand for comparability of research with the need for specification and focus?

Second, we see not only a growing number of concepts but also a sort of camp formation in terms of who works with which of these concepts. For example, in a recent review paper on racism and hate speech in social media, Matamoros-Fernández and Farkas (2021) note "striking differences in the conceptual vocabularies used across quantitative and qualitative studies, with the former predominantly using the term 'hate speech' and the latter using 'racism'" (p. 216). Based on this finding, they detect a "terminological divide in the field" (p. 212). And indeed, our observation as editors of this volume is a similar one: Conceptual issues were discussed quite passionately between the authors in the course of the mutual reviews. There is clearly a need for more in-depth discussion here.

Our collection of texts on different concepts can hopefully be a start for this discussion, especially since it does not cover all of them. That said, our hope is that it initiates a more intense and informed conversation and helps building bridges in the process. We think academia has a special responsibility to address conceptual and definitional issues, given the fact that hate speech is also the subject of intense public debate.

We start with the "hate speech" concept as it is prominently included in this book's title, and also because we work with this concept in our own research as well. Nevertheless, we asked *Liriam Sponholz* to write a plea for this concept, since she is a renowned expert in this regard. In the first text of this section, she elaborates on the origins of the hate speech term in critical race theory, which has already embedded the consideration of social inequalities and power asymmetries in the definition of the term. According to this understanding, hate speech is defined as a symbolic attack against historically or systematically marginalized

groups and their (supposed) members. Against this background, she then discusses the issues of concept stretching, concept shrinking and conceptual inflation in the recent literature and their consequences for academia, politics, and society.

Lena Frischlich discusses the specific fallouts of hate speech from a social psychological perspective, similarly concluding that hate speech cannot be understood without taking into account pre-existing power structures and resource inequalities. In the second part of the text, she discusses the psychological research on perpetrators of hate speech and derives valuable insights for preventive measures.

With the concept of "dangerous speech," *Susan Benesch* contributes a perspective that also focuses on the harm of speech acts but draws on empirically observed patterns in public speech in the run-up to genocides and mass violence in different parts of the world and historical periods. Specific to the concept is the observation that speech acts have a cumulative effect on people through repetition and that different contextual factors play a role in assessing the (gradual) dangerousness of a speech act.

Marike Bormann and *Marc Ziegele* argue for the concept of (political) "incivility," which is rooted in social theory (e.g., deliberation theory and politeness theories) and has a long research tradition. The two authors discuss current challenges of the research strand related to the inconsistency of definitions and measures, the reliability of incivility measurement, and normative implications. Moreover, they offer a multidimensional model of political incivility that integrates different strands of incivility research and encompasses violations of five different norms of communication.

With the concept of "toxicity," *Julian Risch* presents a quite different perspective. The concept originated in computer science and application-oriented, industry-led research in the area of automated user comment classification (and hiding/removal). It focuses on the impact of user comments in online discussions and on the goal of ensuring that no users are pushed out of these discussions. Similar to incivility, toxicity is a comparatively broad concept that can encompass various subcategories and can be adapted to the specific needs of the potential users of a classification solution.

In her text on "extreme speech," *Sahana Udupa* introduces a critical perspective on digital practices that departs from established definitions of hate speech and mis-, dis- or malinformation but calls for a holistic, culturally and historically sensitive approach to these practices. Rather than replacing existing concepts, extreme

speech research aims to add new perspectives to hate speech research and considers ambivalences in the context of (political and economic) power relations, colonialism, and socio-technological transformations. Therefore, the framework emphasizes the need to balance the close contextualization of immediate contexts with a deep contextualization of underlying historical and colonial continuities.

The text by *Thorsten Quandt* and *Johanna Klapproth* revises the umbrella concept of "dark participation," introduced by the first author in 2018. This concept offers a systematization of various forms of negative or destructive user participation on the Internet along the main categories of actor, reasoning, object/target, audience, and process. However, the original article was also motivated as a commentary on the prevailing, one-sided focus of research (and, thus, also of this volume) on such negative aspects and as a call for more integrative theorizing and research. In their text for this volume, the two authors now reemphasize this motivation, discuss the resulting conceptual limitations of the dark participation model, and summarize the reactions and recommendations of the research community that followed the original publication.

Gina M. Masullo concludes the second part of this book with a text calling for a new approach to incivility research, which can also be read with regard to other concepts. In this text, Masullo pleads for addressing the specific forms of incivility, rather than continuing to treat it "as a monolith." In particular, she points to the need for multidimensional approaches that take into account the different theoretical underpinnings of incivility and allow for more specific research questions to be asked, for example, regarding the harmfulness of certain forms of incivility or contextual factors. She further identifies three research areas that need more research: the impact of online incivility on marginalized social groups and the protection of these groups, the role and power of social media platforms in regulating online incivility, and the dynamics between incivility and other forms of problematic online communication such as mis- and disinformation.

4 Methodological perspectives: Operationalization, automation and data

The third section of this volume focuses on methodological issues in the context of hate speech research. As in any other field, valid and reliable methods are key to scientific evidence on hate speech, especially because this field of re-

search brings together different disciplinary perspectives and methodological standpoints. As an object of academic research, hate speech in social media is not conventional media content but rather a form of applied language sitting in the ambivalent space between interpersonal and public communication, shaped by social interactions, algorithmic decision-making, business models and design decisions of platform companies. Given the fast evolving possibilities for the collection and analysis of (big) data, empirical hate speech research not only demands for new theoretical models of public spheres and social discourse but also has to solve challenges of accessing, archiving, sharing and analyzing data.

The section opens with a text by *Babak Bahador*, who presents an approach to monitoring hate speech that he and his team have used to analyze U.S. media. Starting from a critique of common hate speech definitions, he introduces an hate speech intensity scale that ranges from "disagreement" to "death." He justifies the necessity of such an early warning system, which also includes weaker forms of antagonistic criticism, by pointing out that "[o]nce more extreme hate speech takes hold, it could also be a sign that it is too late to implement more peaceful preventative actions."

Salla-Maaria Laaksonen provides valuable insights into lessons learnt in a use case for automated hate speech detection. She describes which compromises and simplifications are necessary to develop and apply a successful machine learning model for the identification of hate speech and emphasizes the importance of human training and monitoring. In her use case, contextual factors regarding the message, the author and the public impact of the postings increased the model quality and its lifetime.

Christian Baden discusses the numerous challenges of language for machine-assisted hate speech detection. For example, changes in language can be used metaphorically and ironically and thus mask insults and hate. In addition, the expansion of classification models through contextual data could lead to more ambiguity and evasive language use by those who use hate speech. It is a kind of arms race. The methods are refined but still cannot overcome the evolving social abysses behind animosity and hate.

Besides ambiguity and irony, implicity is another major challenge for identifying hate speech. Falling back on a corpus from their research project "Decoding Antisemitism," *Matthias J. Becker* and *Hagen Troschke* present examples of implicit statements that contain antisemitic stereotypes and prejudices but that are not

clear at first glance. They distinguish three areas of knowledge that help to extrapolate the implicit, and eventually identify those forms of antisemitism that are often disguised. In order to secure one's own interpretations in this context, the authors give concrete examples of "how implicitness can be realized at the different levels and how these levels can interact."

"Machines do not decide hate speech" is the title and claim of the text by *Jae Yeon Kim***.** The author understands the establishment of what counts as hate speech as a negotiation process between social groups based on norms. Transparency and debate on the applied definitions of hate speech must therefore precede the model-building process. Accordingly, he argues that persons and groups affected by hate speech need to be included into the process, which would make it both more accurate and democratic.

Anke Stoll critically comments machine learning as part of the artificial intelligence hype. In a kind of recipe, she shows how, in four simple steps, a phony classifier can be trained to deliver seemingly outstanding results that are nothing but artifacts. In this context, she discusses potential pitfalls and flaws of machine learning models and shows how not to proceed if we aim for meaningful results.

In the next text, *Laura Laugwitz* demonstrates how validity as a major quality criterion for empirical studies can be applied to automated content analyses. She explains various supervised text classification methods and shows that the functional descriptions of these models are not suitable for an assessment of validity in the empirical sense. Following an interdisciplinary approach, she pleads for closer cooperation between computer science and communication science to develop such criteria.

From a legal perspective, *Paddy Leerssen, Amélie Heldt* and *Matthias C. Kettemann* look at the accessibility of social media data for researchers in Europe. There are many laws that make access difficult and some regulations that should make it easier to get data from platforms. Privacy, freedom of information, data protection and copyright are rights and areas of law that partly overlap and can make scientific access to platform data difficult. Finally, the authors call for a clear and unambiguous framework for scientific data access.

Jakob Jünger takes a look at social media data from a hermeneutic perspective. Data collection here is an uncertain process that requires many interpretative decisions and therefore has a great influence on the later research results. The selection and availability of data, access restrictions, the systematics of the websites as well

as the archiving of the data show the tension between creativity and standardization that we as researchers face and that we have to dissolve thoughtfully.

Paula Fortuna, Juan Soler-Company and *Leo Wanner* discuss challenges for both building and comparing annotation datasets. Studies in the context of abusive language research have shown the importance of such data for machine learning models, a lack of common understandings in this context, and the presence of bias and artifacts in recognition and evaluation. Against this background, the authors provide guidelines to address the most pressing issues in a step-by-step guideline to improve the quality of annotated datasets.

In their response to the texts in this third section, *Jaime Lee Kirtz* and *Zeerak Talat* reflect on the various methodological challenges that each step of hate speech research faces, providing a broader orientation for each text of this section they discuss. In this context, they attach particular importance to social issues that need to be addressed in future research on hate speech detection.

Taken together, the third part of this book critically reflects the diversity and heterogeneity of methodological perspectives on machine-based models for the detection of linguistic constructs in social media. Against the background of these contributions, we think that the field of hate speech research is unlikely to succeed without true interdisciplinary exchange, discussions and collaboration. With this volume, we hope to contribute to such a project, and to stimulate first steps toward building bridges between disciplines, theoretical perspectives, and methods.

Last but not least, we would like to thank all authors of this volume for their excellent contributions, the rich discussions during the review process, and for their infinite patience with us editors. From our point of view, the experiment of a discursive collection of texts on the various challenges and future perspectives of hate speech research was more than successful. Perhaps it can even serve as a model for other research fields that are considering similar endeavors. To you, the reader, we wish an exciting and insightful read.

Sünje Paasch-Colberg is a Research Associate at the German Centre for Integration and Migration Research (DeZIM) in Berlin, Germany. https://orcid.org/0000-0002-0771-9646

Christian Strippel is research unit lead of the Weizenbaum Panel and the Methods Lab at the Weizenbaum Institute for the Networked Society in Berlin, Germany. https://orcid.org/0000-0002-7465-4918

Martin Emmer is Professor for Media and Communication Studies at Freie Universität Berlin and Principal Investigator at the Weizenbaum Institute for the Networked Society in Berlin, Germany. https://orcid.org/0000-0002-0722-132X

Joachim Trebbe is Professor for Media and Communication Studies at Freie Universität Berlin, Germany.

References

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–224. https://doi.org/10.1177/152747642098223

Sellars, A. F. (2016). Defining Hate Speech. Research Publication No. 2016–20. Berkman Klein Center. https://doi.org/10.2139/ssrn.2882244

I. POLITICAL PERSPECTIVES: CURRENT ISSUES AND DEVELOPMENTS

Digital
Communication
Research.de

Recommended citation: de Albuquerque, A., & Alves, M. (2023). Bolsonaro's hate network: From the fringes to the presidency. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 27–42). Digital Communication Research. https://doi.org/10.48541/dcr.v12.2

Abstract: This chapter discusses the origins, development, and characteristics the hate network built around the President of Brazil Jair Bolsonaro. How is it structured? How does it work? What factors have allowed it to exist? How does it affect the health of Brazilian democracy? It argues that Bolsonaro's followers took advantage from an institutional crisis taking place in the late 2010s. At that time, the legacy media provided a massive coverage associating PT (Workers' Party), which was ahead of the presidency, to corruption, thus fostering a hate campaign against it. In that scenario, Bolsonaro emerged as the leader of the far-right opposition to PT. His followers firstly built a network on Facebook, using fake profiles and mixing hate speech with humor. Coordinated unofficially by several cabinet staff members, this network articulates the official profiles of the Bolsonaro family to followers and sympathizers and a vast array of anonymous supporting pages. This structure allowed Bolsonaro's activists to blur the boundaries between official and spurious discourses, and powered a series of flaming wars against political adversaries and reporters perceived as hostile to Bolsonaro's interests.

DOI 10.48541/dcr.v12.2

Afonso de Albuquerque & Marcelo Alves

Bolsonaro's Hate Network

From the fringes to the presidency

1 Introduction

The first interview with Jair Bolsonaro as Brazil's elected president, on November 1, 2018, was different from any other before. Its scenario was quite exotic: a handful of microphones positioned on a surfboard. Traditional news media outlets, such as the newspapers *O Globo*, *O Estado de São Paulo*, and *Folha de S. Paulo*, were not allowed to take part in it. Still, the most notable part of the interview was its content. Bolsonaro minimized the importance of the news media and praised social media. According to him, "the people will decide which media vehicles will survive, and which won't" (Andrade & Maia, 2018). Ten days later, Bolsonaro presented, through his Twitter account, a list "of excellent information channels on Youtube." It included notorious hate speech disseminators, such as Olavo de Carvalho, Nando Moura, and Bernardo Küster. Of course, Bolsonaro is not the only far-right leader of the executive branch who uses hate speech as a part of his political arsenal. Yet, the open manner through which he does it is remarkable. The coordination of his hate speech and fake news strategy is in the hands of a group known as "Gabinete do Ódio" ("Office of Hate"). The president's son Carlos Bolsonaro is a notorious member of this group.

This chapter analyzes Bolsonaro's hate speech network's nature and political impact. How is it structured? How does it work? What factors have allowed it to exist? How does it affect the health of Brazilian democracy? The chapter is organized as follows. First, it presents the political context of Bolsonaro's presidency. We argue that as authoritarian as he can be, Bolsonaro was not the prime guilty party in terms of the crisis of democracy. Instead, he benefited from an already existing crisis, associated with the "Lava Jato" ("Car Wash") operation, a major anti-corruption initiative that had a strong impact on the Brazilian political system by fostering a climate of distrust with regard to the representative institutions. The second section explores the building of Bolsonaro's hate speech network, and the third section analyzes the structure of this network.

2 The institutional context of hate speech in Brazil

Hate speech is a pervasive phenomenon. It exists in most societies, if not all. Conventional wisdom associates hate speech with fringe groups, rather than with representative institutions. Famous exceptions include Nazi Germany and the Rwandan genocide. In the first case, the defeat in World War I and a huge economic crisis provided a fertile ground for hate politics. In the other, longstanding ethnic rivalry and civil war did the same. None of these factors was present in Brazil. Less than a decade before, the future looked promising for Brazilian democracy. Brazil experienced an economic boom. Poverty diminished. The country adopted progressive policies aiming to promote racial and gender equality. Accountability institutions, such as the judiciary, the prosecutors' office, and the press, became more active than before (Praça & Taylor, 2012). According to political scientists, this would provide a solid barrier against human rights abuses (Levitsky & Ziblatt, 2018). Against this backdrop, Bolsonaro's rise to the presidency has been described as an "illiberal backlash" (Albuquerque, 2021; Hunter & Power, 2019). How could this happen?

To understand what has gone wrong in Brazil, it is necessary to take a closer look at its political institutions. The web of accountability institutions, which includes, among others, the judiciary, the prosecutors' office, the federal police, and the press (Power & Taylor, 2011), deserves special attention in this respect. According to an influent view, the active role of these institutions is a key factor

in building a solid democracy, as they provide a barrier against the concentration of powers in the hands of the executive power (O'Donnell, 1998) and fight corruption (Power & Taylor, 2011). In this sense, they work as an immune system for democracy, preventing it from being infected by external, authoritarian agents (Albuquerque, 2021; Mounk, 2018). In line with this perspective, it is possible to suggest that the crisis of Brazilian democracy resulted from the passivity of these institutions. Still, this did not happen. In fact, these institutions have been very active in Brazil in the last decades. The problem is not that they refrained from acting but the manner in which they did it.

Autoimmune diseases provide a valuable metaphor for understanding this. In such diseases, the immune system mistakes parts of the body as foreign threats and attacks them. In extreme cases, this can lead to the death of the organism (Albuquerque, 2021). The "Lava Jato" ("Car Wash") operation provides a powerful example of how the autoimmune disease logic undermined Brazilian democracy. Conducted by Federal Judge Sergio Moro and a team of federal prosecutors led by Deltan Dallagnol, the operation originated as a judicial investigation aiming to tackle corruption in Petrobrás, the Brazilian state-owned oil company. However, its focus soon changed, and Lava Jato acquired a clearer political tone, as it became primarily oriented against the former President Luis Inácio Lula da Silva, and the Workers' Party (Partido dos Trabalhadores, hereafter PT; Meyer, 2018). As a consequence of Lava Jato, Lula was imprisoned in 2018 and forbidden to run for the presidency (Engelmann, 2020; Silva, 2020).

The historical significance of Lava Jato has been disputed in the academic milieu. For some authors committed to the "web of accountability" perspective, it was a turning point for corruption control in Brazil. According to Professor Ana Luiza Aranha (2020): "The success of the Lava Jato investigations resulted from a historic level of coordination among Brazilian institutions of accountability, suggesting that Lava Jato might represent a turning point in the effectiveness of Brazil's web of accountability institutions" (p. 94). Recent evidence suggests that Aranha is right but for the wrong reasons. In 2019, a hacker accessed the messages exchanged by the team of prosecutors ahead of the case and leaked them to the news site *Intercept Brasil*, which published them in the "Vaza Jato" news series. These messages suggest that Judge Moro, the prosecutors' team, and journalists colluded in convicting Lula, motivated by political reasons (Duarte & Intercept Brasil, 2000).

The massive media coverage of Lava Jato systematically associated representative politics with corruption (Albuquerque & Gagliardi, 2020; Damgard, 2018). By doing this, it fostered political polarization and suspicion regarding the democratic institutions. In special, a hate campaign against Lula and his partisans ("petistas" and leftists, in general) took place. Digital haters even commemorated the passing of Arthur, Lula's seven-year-old grandson, who died from sepsis (Arias, 2019). Lava Jato provided the context that allowed the impeachment of President Dilma Rousseff to take place in 2016. It also resulted in the condemnation of Lula in 2017, which put him in jail in 2018. Given that Lula was the clear favorite to win the 2018 presidential election, his removal from the dispute opened the way for Bolsonaro's victory. Bolsonaro invited Moro to serve as his minister of justice. Moro accepted and worked in his government for nine months. All in all, the institutions that were supposed to contain hate speech fostered it.

3 Building a hate network with public money

The election of Jair Bolsonaro in 2018 was a turning point for political communication strategies in Brazil. For the first time, a fringe politician won a national election without massive television electoral campaigning (Santos & Tanscheit, 2019). For almost three decades, Bolsonaro was known for violent statements and no significant legislation (Nascimento et al., 2018). How could a local deputy without any political power become the leader of a national far-right movement? In the aftermath of the Lava Jato institutional corrosion, Bolsonaro voiced feelings of distrust and anger. The digital communication strategy extended far beyond managing official social media profiles. In fact, the president spearheaded a vast network of operatives that was built many years before the 2018 election (Alves, 2019a). This section describes the early creation of Bolsonaro's digital communication network.

Once Jair Bolsonaro was elected as the Brazilian president, his confrontational style did not diminish. On the contrary, his fiery speech was directed toward anyone who criticized his actions. Bolsonaro's digital network ("Bolsonaristas") spread hate against journalists, scientists, artists, and politicians (Mello, 2020). Not even members of his own government or the allied branch of the parliament was safe. In several noteworthy events, bolsonaristas harassed ministers and party members that were perceived as enemies. A group named the "Office of Hate" ("Gabinete do Ódio") was

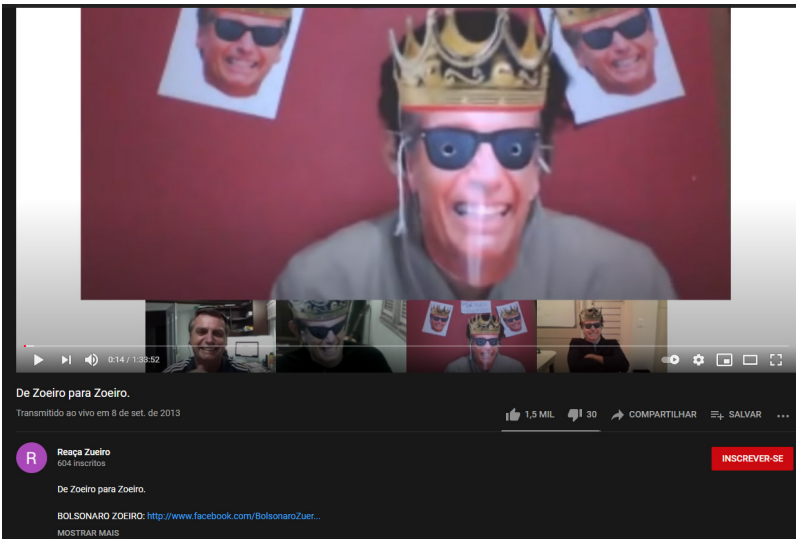
referenced as the coordinator of the attacks. The Office of Hate was built by a group of digital-savvy young conservatives employed by Bolsonaro's family to create memes and videos on social media. Bolsonaro's official channels did not publish this inflammatory content. In fact, the Office of Hate owns many sock puppets—fake profiles dedicated to amplifying hate speech (Alves, 2019b; Lerner, 2020). This anonymous network mobilizes followers to smear political targets usually as a joke.

The Office of Hate is a shady operation composed of young supporters hired to manage fake parody profiles to increase Bolsonaro's visibility. The model dates back to 2013. It was designed by Carlos Bolsonaro, who began to monitor pages and Facebook groups that supported his father and his ideals. Carlos thought that the mass media had a strong left bias and ignored conservative thinking (Gaspar, 2019). Back then, Jair was one of Brazil's most voted for federal deputies in Rio de Janeiro. He also appeared sporadically on television in auditorium and humor shows. Fast-growing social media platforms, such as Facebook, provided a way to spread conspiracy theories or hate speech online. This was the main driver of Bolsonaro's popularity in the years to come, along with a network of WhatsApp groups.

At the same time, Bolsonaro had a relatively popular Facebook presence, not only on his official fan page but also in fan-club style parodies that replicated faster than content moderation efforts. In his research, Carlos Bolsonaro found parodies, such as "Bolsonaro Zuero" (Joker Bolsonaro) and "Bolsonaro Opressor" (Opressive Bolsonaro), which were anonymous Facebook pages that celebrated Jair Bolsonaro. These parodies framed his father as an authentic myth that challenged political correctness and spoke from the heart (Ribeiro et al., 2016). They were also quite popular, followed by circa 30,000 people at the time. Carlos posted to his page in April 2013: "I'm having a bad laugh with the page 'Bolsonaro Zuero.'"

Carlos was responsible for contacting anonymous managers and arranging meetings with Bolsonaro. In September 2013, Jair Bolsonaro joined a live interview transmitted by the YouTube channel "Reaça Zuero" ("Joker Reactionary"). He talked with three young people who appeared on camera wearing masks depicting a photo of the far-right politician and a crown for the "king of lulz." In October 2013, another live appearance took place when the deputy received one of the anonymous creators in his office at the National Congress. Five years before Bolsonaro was elected, the fake network managers had direct access to Bolsonaro's personal cabinet as a federal deputy. Again, the interviewer disguised himself, covering his face with sunglasses and a wig.

Figure 1: Reaça Zuero's live transmission of Jair Bolsonaro interviewed by anonymous creators of far-right content and Bolsonaro in 2013



Source: Reaça Zueiro (2013)

So far, there is no evidence that Bolsonaro funded this video. However, it is worth noting that Bolsonaro's political district is São Paulo and Rio de Janeiro, which are Brazilian southeastern states. The first youngsters hired were actually from Ceará, a northeastern state. It is quite improbable that Bolsonaro initiated the strategy of parody accounts at the very beginning. All the available federal and journalistic evidence found suggests that he later hired the creators to work as members of his staff (Gaspar, 2019). Even so, there is no doubt that Bolsonaro's family was responsible for orchestrating and financing the managers, greatly enhancing their production and reach. The once playful humorous accounts became a central part of Bolsonaro's network of hate.

Bolsonaro Zuero stands for a model of recruiting young people in public offices to create unofficial content. Journalistic investigations found that phantom workers were employed by Bolsonaro's family as early as 2015 (Ghirotto et al., 2019). In 2019, José Mateus Sales Gomes (Bolsonaro Zuero) and Tércio

Arnaud Thomaz (Bolsonaro Opressor) held positions as the president's advisors. This strategy expanded beyond the two parodies, coordinating a vast network of fakes and copycats.

The core idea was to create several political personas as a humorous parody of Bolsonaro. Members of the communication team anonymously coordinated several false profiles. The strategy attracted the spotlights when Bolsonaro was elected president. In July 2020, Facebook deactivated an inauthentic network attributed to the Bolsonaro family. It had 35 accounts on Facebook and 38 on Instagram aimed at spreading misinformation and harassing opponents. Among the accounts deactivated was "Bolsofeios" ("Ugly Bolsonaros"), which had hundreds of thousands of followers on Instagram. Bolsofeios was an active collaborator of the group that defined targets and coordinated digital attacks on Bolsonaro's behalf. The account manager, Eduardo Guimarães, served as a direct advisor to Eduardo Bolsonaro (Rezende, 2020). He created the persona from the deputy's office using his personal email.

Among multiple hate speech cases driven by the network, Bolsofeios took part in the persecution of *Folha de São Paulo* reporter Patricia Campos Mello. During the 2018 election, her stories revealed business owners' illegal donations to finance services of message forwarding against the Workers' Party on WhatsApp. Other accounts, such as Bolsonewsss and Bolsonaro Opressor 2.0, administered by Tércio Thomaz, are examples of this pattern. Investigations also found false profiles managed by communication advisors from federal and state deputies who support Bolsonaro, such as Alana Passos and Anderson Moraes.

One of the most significant outcomes of this political communication model was Gil Diniz (Baptista Jr., 2020). Diniz was a poor postal service worker. His life changed when he created the Facebook page Carteiro Reaça (Reactionary Postman), which enthusiastically praised Bolsonaro. He introduced himself to Eduardo in 2014, who hired him for his parliamentary staff. Diniz was responsible for operating false profiles to share memes and positive news on social media. In the 2018 election, São Paulo elected him as state representative with over 214,000 votes. Diniz's support relied mainly on the far-right wave that elected several digital influencers. During his term, the public prosecuting office accused him of an illegal salary deduction. The investigation discovered that the deputy had created his fake news operation to harass political opponents (Dal Piva & Saconi, 2019).

4 Bolsonarista network structure on Facebook

The Brazilian news media coined the term "Office of Hate" when far-right activists flooded social media with attacks against democratic institutions. It represents the staff in charge of digital communications that occupy a room close to Bolsonaro's. The investigation held by the National Congress raised a large trove of evidence, such as payments to advisors and companies. However, this strategy is certainly not the only method used by Bolsonaro's communication network. In this section, we will present the empirical results of a social network analysis of the connections between dozens of false pages that amplify far-right hate speech in Brazil.

The operation of the Office of Hate represents the strategizing head of a vast fake news network. The main feature is hiring young people with public resources to operate fake profiles and spread hatred on social media. It is an ideological community articulated on digital channels by Bolsonaro's family and allies, who define political targets and schedule the messages. In general, these publicly funded fake news networks are part of the parliamentary quotas for hiring staffers. In this sense, these profiles are the most faithful and are closely supervised by the family.

However, it is unlikely that all the channels are sock puppets orchestrated by the president's family. Cesarino (2019) argues in favor of a layered organization of Bolsonaro WhatsApp groups. This idea is quite useful for understanding their general communication network. At the center is the family itself and its closest advisors. They control the official accounts and groups, as well as anonymous profiles, on a wide range of platforms. Digital activists and supporters are in the intermediate circles. They manage most of the parodies and amplify the attacks and frames initially created by the advisors. Finally, profiles of ordinary people contribute by sharing content with their friends. In this sense, the network is orchestrated by the family and its closest allies, advisors, and employees. Nevertheless, the network overlaps with independent activists who share the ideology and causes of the far-right.

To determine the pages of an extended Bolsonarista network, we sought to retrieve and organize digital traces. This study analyzed the following network between right fan pages on Facebook. The data collection procedure used was automated snowball sampling. From a list of 500 right fan pages discovered by previous studies (Alves, 2019a), a crawler extracted the following network (i.e., the pages followed by the seeds). This is a network mapping technique that adds

new nodes to the initial sample. Data cleansing procedures kept only the pages that mentioned the term "Bolso" in the title. We selected this word because most formerly known channels used some adaptation of the name "Bolsonaro" online, such as Bolsonaro Zuero.

Finally, Gephi's social network analysis software processed the connections to identify the network structure. Figure 2 shows the result of the application of the Force Atlas 2 layout on the network composed of 85 nodes and 1,255 edges. The size of the nodes represents the degree of input, that is, how many network pairs follow that page.

Figure 2: Bolsonaro's expanded communication network on Facebook

The result shows a very cohesive network structure organized around the Bolsonaro family members. Jair, Eduardo, Flavio, and Carlos had by far the highest indegrees. This pattern indicates their reference role with regard to the other channels. However, the network of fake news and hate speech operates far beyond official profiles. A large number of parodies, hyper-partisan media, and meme factories increases the numbers of flaming wars waged against the adversaries.

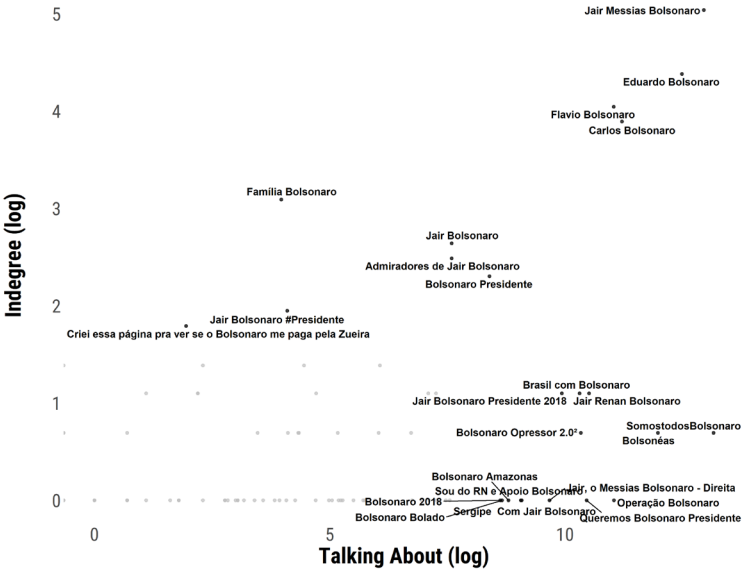
One of the main puzzles in terms of understanding Bolsonaro's digital communication is the dynamic relations between the official discourse and the apocryphal content. This does not mean that the president and his family are not themselves producers of hate speech. The aggressive behavior is constantly observed on social media and in interviews. One example is the smear campaign against journalists who revealed the Flavio Bolsonaro illegal scheme of improper salary deductions (Mello, 2020). One point to note is that there is an infrastructure carefully set up to guide favorable themes and frameworks in the public agenda.

In this network, specific roles are performed by different types of Facebook pages. The combination of social network and engagement metrics shows some hierarchies between official and anonymous support pages. The "talking about" metric counts external mentions on Facebook, and indegree is a connection metric that counts how many followers a node has in the studied sample.

Figure 3 shows how the members of the Bolsonaro family stand out both in terms of metrics of popularity and as a reference in the network. They are the only pages located in the upper right quadrant. However, SomosTodosBolsonaro, Bolsonéas, Operation Bolsonaro (Bolsonaro's Operation), and Bolsonaro Opressor 2.0 are in the lower right quadrant, which means they yield large popularity but are not followed to a great extent by other pages. They are the parodies and political personas that act as mobilizers and agitators of sympathizers. Bolsonaro's communication staff operates anonymously through some of the pages, investigative reporters have shown. Bolsonéas, for example, was run by the team of state representative Alana Passos (Brandt et al., 2020). Many of these accounts are financed by entrepreneurs (Toledo, 2018) and benefit from the monetization of content on YouTube, and some receive advertising resources from the federal government.

Some nodes have few public mentions but several connections in the network: Familia Bolsonaro, Jair Bolsonaro #President, and Criei essa página pra ver se o Bolsonaro me paga pela Zueira (I created this page to see whether Bolsonaro would pay me for the lulz). They are not financed or contracted by the Bolsonaro

Figure 3: Scatterplot of Bolsonaro's network



family or its political supporters. However, these smaller pages follow the general trends and frames advanced by the far-right network. This facet of the problem portrays acts of the co-creation of meaning by the sympathizers themselves.

There is an assemblage of strategies created by the supporters themselves. Before the 2018 election, Carlos Bolsonaro promoted a meeting with several participants to articulate this coordination strategy. Leaders of the movement for the impeachment of Dilma Rousseff, digital influencers, and content creators attended the meeting. Often, activists cite the hashtag #MarketeirosdoJair or refer to themselves ironically as robots as an internal community joke to articulate actions.

This broad network is very active and engaged in content production, usually using memes and conducting flaming wars. One example of such an orchestration was the harassment and hate speech against the investigative journalist and award-winner Patricia Campos de Mello. The reporter published a series of news

pieces revealing that Bolsonaro's 2018 campaign benefited from illegal services of bulk messages sent to the WhatsApps of Brazilian citizens in the second turn. In response to the publications, Bolsonaro's support network resorted to a massive public shaming campaign, spreading multiple sexist and misogynous attacks, including saying that the reporter had exchanged the scoop for sexual favors (Neder, 2021). This type of smear campaign against female journalists is a pattern emerging from Bolsonaro's family that has been repeated very often by the president himself. Reporters Without Borders (RSF) has tallied 580 attacks against media in Brazil during 2020, with most of them directed toward female journalists (2021).

5 Concluding remarks

Jair Bolsonaro revolutionized Brazilian political communication, both in terms of its formal aspects and content. For the first time, a candidate won a presidential dispute without having the mass media as his campaigning backbone. Instead, he used social media as his main communication resource. Even during his term ahead of the presidency, he has privileged social media over mass media. His communication style is extremist: Bolsonaro often expresses racist and misogynistic views. Hate speech and fake news are important elements of his rhetorical toolkit. How could this happen? To start with, Bolsonaro took advantage of a previously existing institutional crisis. The same institutions that were supposed to defend citizens' rights attacked them. This provided hate speech groups with tremendous opportunities. Social media offered them the means to exploit this situation. Bolsonaro used these groups for his own benefit. In exchange for their support, he hired them as part of his government's communication team.

In this paper, we demonstrated how Bolsonaro's ferocious hate speech campaigns are orchestrated on social media by multiple anonymous accounts that enhance the visibility of sexist and misogynous attacks by the president and his family. This network is funded and supported by public resources since some of its administrators were hired to work in the deputy cabinet of Bolsonaro's family elected members. Upon his election as president, Bolsonaro increased this modus operandi to harass, persecute, and publicly shame anyone considered an enemy or a traitor. Social media platforms should enforce policies against hate speech to moderate and deplatformize such speech coming from elected officials in Brazil.

Further research is necessary to understand how Bolsonaro's family coordinates the attacks and how alt-tech platforms are used to determine targets and spread hate speech in Brazil.

Afonso de Albuquerque is Professor at the Graduate Program in Communication at the Fluminense Federal University, Brazil. https://orcid.org/0000-0002-2608-7605

Marcelo Alves is Professor at the Graduate Program in Communication at the Pontifical Catholic University of Rio de Janeiro, Brazil (PPGCOM/PUC-Rio). https://orcid.org/0000-0003-4995-6612

References

Albuquerque, A. (2021). The two sources of the illiberal turn in Brazil. *Brown Journal of World Affairs*, *27*(2), 127–144.

Albuquerque, A., & Gagliardi, J. (2020). Democracy as corruption: The news media and the debunking of democracy. In X. Orchard, S. Garcia Santamaria, J. Brambila, & J. Lugo-Ocando (Eds.), *Media & governance in Latin America: Towards a plurality of voices* (pp. 77–96). Peter Lang.

Alves, M. (2019a). *VaipraCuba: A gênese das redes de direita no Facebook* [#GotoCuba: The genesis of rightist networks on Facebook]. Appris.

Alves, M. (2019b). Desarranjo da visibilidade, desordem informacional e polarização no Brasil entre 2013 e 2018 [Visibility distress, informational disorder and polarizarion in Brazil between 2013 and 2018] [PhD thesis]. Communication Program, Fluminense Federal University.

Andrade, H., & Maia, G. (2018, November 1). *Bolsonaro sobre o papel da imprensa: Cheguei ao poder graças às mídias sociais* [Bolsonaro on the role of the press: I came to power thanks to social media]. UOL. https://noticias.uol.com.br/politica/ultimas-noticias/2018/11/01/bolsonaro-relacao-com-imprensa-midias-sociais.htm

Aranha, A. L. (2020). Lava Jato and Brazil's web of accountability: A turning point for corruption control? In P. Lagunes & J. Svejnar (Eds.), *Corruption and the Lava Jato scandal in Latin America* (pp. 94–112). Routledge.

Arias, J. (2019, March 2). *A morte do inocente neto de Lula soltou os monstros do ódio* [The death of Lula's innocent grandson liberated the hate monsters]. El País. https://brasil.elpais.com/brasil/2019/03/02/opinion/1551487708_675741.html

Baptista Jr., J. (2020, July 30). *Por que o deputado Gil Diniz é um elo importante com a família Bolsonaro* [Why representative Gil Diniz is not an important link with the Bolsonaro family]. Veja. https://veja.abril.com.br/blog/veja-gente/por-que-o-deputado-gil-diniz-e-um-elo-importante-com-a-familia-bolsonaro/

Brandt, R., Macedo, F., Moura, R., Ortega, P., & Netto, P. (2020, May 27). *Veja as conexões de bolsonaristas e os perfis de internet do 'gabinete do ódio'* [See the connections between Bolsonaristas and the internet profiles of the 'hatred cabinet']. Estadão. https://politica.estadao.com.br/blogs/fausto-macedo/veja-as-conexoes-de-bolsonaristas-e-os-perfis-de-internet-do-gabinete-do-odio/

Cesarino, L. (2019). Identidade e representação no bolsonarismo: Corpo digital do rei, bivalência conservadorismo-neoliberalismo e pessoa fractal [Identity and representation in Bolsonarismo: The king's digital body, bivalence conservatism-neoliberalism, and the fractal person]. *Revista de Antropologia*, *62*(3), 530–557. https://doi.org/10.11606/2179-0892.ra.2019.165232

Dal Piva, J., & Saconi, J. P. (2019, November 1). *Central de boatos de Gil Diniz no Whatsapp tem memes contra João Doria e outros adversários* [Gil Diniz's rumor center on Whatsapp has memes against João Doria and other opponents]. Época. https://oglobo.globo.com/epoca/brasil/central-de-boatos-de-gil-diniz-no-whatsapp-tem-memes-contra-joao-doria-outros-adversarios-24054119

Damgaard, M. (2018). Cascading corruption news: Explaining the bias of media attention in Brazil's political scandals. *Opinião Pública, 24*(1), 114–143. https://doi.org/10.1590/1807-01912018241114

Duarte, L., & Intercept Brasil. (2020). *Vaza Jato: Os bastidores das reportagens que sacudiram o Brasil* [Vaza Jato: The background of the reports that shook Brazil]. Mórula Editorial.

Engelmann, F. (2020). The 'fight against corruption' in Brazil from the 2000s: A political crusade through judicial activism. *Journal of Law and Society*, *47*(s1), 74–89. https://doi.org/10.1111/jols.12249

Gaspar, M. (2019). *O pit bull do papai* [Daddy's pitbull]. Revista Piaui. https://piaui.folha.uol.com.br/materia/o-pit-bull-do-papai/

Ghirotto, E., Vieira, M. C., & Lopes, A. (2019, July 12). *As milícias digitais* [The digital militias]. Veja. https://veja.abril.com.br/brasil/as-milicias-digitais/

Hunter, W., & Power, T. J. (2019). Bolsonaro and Brazil's illiberal backlash. *Journal of Democracy*, *30*(1), 68–82. https://doi.org/10.1353/jod.2019.0005

Lerner, C. (2020). A direita unida em torno de Bolsonaro: Uma análise da rede conservadora no Facebook [The right united around Bolsonaro: An analysis of the Conservative network on Facebook]. In F. G. Faria & M. L. B. Marques (Eds.), *Giros à direita: Análises e perspectivas sobre o campo líbero-conservador* [Turns to the right: Analyses and perspectives on the Liberal-Conservative field] (pp. 90–121). Editora SertãoCult.

Levitsky, S., & Ziblatt, D. (2018). *How democracies die*. Crown.

Maiwaring, S., & Welna, C. (2003). *Democratic accountability in Latin America*. Oxford University Press.

Mello, P. C. (2020). *A máquina do ódio: Notas de uma repórter sobre fake news e violência digital* [The hate machine: notes of a reporter about fake news and digital violence]. Companhia das Letras.

Meyer, E. P. N. (2018). Judges and courts destabilizing constitutionalism: The Brazilian judiciary branch's political and authoritarian character. *German Law Journal*, *19*(4), 727–768. https://doi.org/10.1017/S2071832200022860

Mounk, Y. (2018). *The people vs. democracy: Why our freedom is in danger and how to save it*. Harvard University Press.

Nascimento, L., Alecrim, M., Oliveira, J., Oliveira, M., & Costa, S. (2018). "Não falo o que o povo quer, sou o que o povo quer": 30 anos (1987–2017) de pautas políticas de Jair Bolsonaro nos jornais brasileiros [I don't speak what the people want, I am what the people want: 30 years (1987–2017) of Jair Bolsonaro's political agenda in the Brazilian newspapers]. *Plural*, *25*(1), 135–171. https://doi.org/10.11606/issn.2176-8099.pcso.2018.149019

Neder, R. (2021). *Brazilian journalist Patrícia Campos Mello sued President Bolsonaro's son for moral damages – and won*. Committee to Protect Journalists. https://cpj.org/2021/03/brazilian-journalist-patricia-campos-mello-sued-president-bolsonaros-son-for-moral-damages-and-won/

O'Donnell, G. (1998). Horizontal accountability in new polyarchies. *Journal of Democracy*, *9*(3), 112–126.

Power, T. J., & Taylor, M. M. (2011). *Corruption and democracy in Brazil: The struggle for accountability*. University of Notre Dame.

Praça, S., & Taylor, M. M. (2012). Inching toward accountability: The evolution of Brazil's anticorruption institutions, 1985–2010. *Latin American Politics & Society*, *56*(2), 27–48. https://doi.org/10.1111/j.1548-2456.2014.00230.x

Reaça Zueiro. (2013, September 9). *De Zoeiro para Zoeiro* [Video]. YouTube. https://www.youtube.com/watch?v=k1q4hTFOXR4

Ribeiro, L. G. M., Lasaitis, C., & Gurgel, L. (2016). Bolsonaro Zuero 3.0: Um estudo sobre as novas articulações do discurso da direita brasileira através das redes sociais [A study on the new articulations of the Brazilian right's discourse through the social networks]. *Anagrama*, *10*(2), 1-16.

RSF. (2021). *RSF tallied 580 attacks against media in Brazil in 2020*. Reporters Without Borders. https://rsf.org/en/reports/rsf-tallied-580-attacks-against-media-brazil-2020

Santos, F., & Tanschett, T. (2019). When old actors leave: The rise of the new political right in Brazil. *Colombia Internacional*, (99), 151–186. https://doi.org/10.7440/colombiaint99.2019.06

Silva, F. S. (2020). From Car Wash to Bolsonaro: Law and lawyers in Brazil's illiberal turn (2014–2018). *Journal of Law and Society*, *47*(s1), 90–110. https://doi.org/10.1111/jols.12250

Toledo, L. F. (2018, October 12). *Rede pró-Bolsonaro engaja mais do que Madonna e Neymar* [Bolsonaro's network engages more people than Madonna's and Neymar's]. Estadão. https://politica.estadao.com.br/noticias/eleicoes,rede-pro-bolsonaro-engaja-mais-do-que-madonna-e-neymar,70002544629

Digital
Communication
Research.de

Recommended citation: Harb, Z. (2023). Journalists as messengers of hate speech: The case of Lebanon. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 45–53). Digital Communication Research. https://doi.org/10.48541/dcr.v12.3

Abstract: The term "crisis" has become synonymous with describing the political, social, and economic state of many Arab countries including Lebanon. These continuous crises, including a global pandemic, have manifested in Lebanese news and current affairs through messages of hate disseminating via the media and journalists. Hate speech circulated via airwaves and the Internet has been shown to cause more harm than having hate shared in private conversations. The global pandemic, followed by the Beirut Port explosion in August 2020, has raised the level of hate speech in public, and Lebanese journalists have been used directly or indirectly as tools for propagating hate speech. This reflective account engages Lebanese journalists with the aim of producing a set of guidelines for tackling hate speech in news coverage and current affairs programs. Two workshops were conducted with Lebanese journalists in Lebanon in an attempt to understand the level of awareness of hate speech and its consequences among Lebanese journalists, assess how they understand hate speech, and determine the importance of guidelines and tools in helping journalists identify and tackle hate speech.

DOI 10.48541/dcr.v12.3

Zahera Harb

Journalists as Messengers of Hate Speech[1]

The case of Lebanon

1 Explosive hate

On August 4, 2020, a few hours after finishing my first workshop with journalists on "hate speech in the Lebanese media in times of crisis" in the offices of the pan-Arab magazine *180 Post* in Beirut, a huge explosion shattered the city, and I was one of its victims. I sustained a few face wounds from broken glass that required sutures, but those visible wounds had little impact compared to the invisible scars the Beirut port explosion left inside every one of us Lebanese people. A sense of despair, anger, and sorrow swept us all. Feelings of helplessness and hopelessness overwhelmed us. Personally, that sorrow and anger grew bigger a few days after the explosion when the Lebanese political factions and sect leaders started a war over airwaves and social media as to whom was to blame for the explosion. A war of hate messages erupted that took Lebanon back

1 This chapter is part of a larger project investigating closely media and journalism practices in Lebanon and Egypt, and the relationship between hate speech and journalism in times of crisis. Passages of this chapter was published in an article the author has written for the Ethical Journalism Network website, of which she is a board member and trustee.

to civil war divisions: Christians versus Muslims and Shia versus Sunni, with the blame mainly falling on the Shia community at large in their generalized and assumed affiliation to Hezbollah. Several journalists took those hate messages to their hearts and became their driving force. Hate speech demonstrated over social media was soon passed onto TV screens and vice versa. The harm caused by the highly divisive rhetoric was transmitted to the homes of millions through journalists, either on purpose or out of ignorance. Hate messages fueled the insecurities among different sectarian communities toward each other.

In the aftermath of the explosion, two scenes dominated the country—one of solidarity demonstrated by the "army of brooms" of volunteers pouring from all over the country to help those affected by the explosion, and another one of hate, gaslit by journalists and media personalities. The despair and anger caused by the explosion seemed to be channeled into hate among many Lebanese against "the other," rather than against those ruling politicians, who at different stages of the six years the ammonium nitrate was stored in the port knew about the explosive material and its devastating impact if exploded. The extreme hate demonstrated by members of the public on social media disturbed me, but not as much as observing journalists share, write, broadcast, and post hate images and texts while declaring their informed support for those messages. Why are many Lebanese journalists keen on jumping on board of sectarian hate with little attention to what that might cause, including reigniting the Lebanese civil war (1975–1990) that took thousands of lives, left thousands with injuries or disabilities, and internally displaced hundreds of thousands? An answer might be related to the fact that Lebanon, as a nation, is still struggling to come to terms with its traumatic past. However, another interpretation might lie within one revelation that came out while discussing the term 'hate speech' with Lebanese journalists, which is that many of them were unfamiliar with the term (or at least the Arabic translation of it). In the next section, I will highlight the main findings of the two workshops conducted in Beirut with 15 mid-careers to senior journalists. These workshops raised more questions than answers regarding the definition of hate speech and its implications for journalism and journalists. This chapter ends by introducing some suggestions for tackling hate speech in Lebanese media.

2 Beirut hate speech workshops

The lack of a relative understanding of what hate speech is, what it means, and what consequences it entails surfaced during the workshops. There is not one accepted international definition of hate speech, and, according to the Ethical Journalism Network (2015), "the tolerance levels of speech vary dramatically from country to country." However, the common understanding is whether speech aims to harm others' harm, "particularly at moments when there is the threat of immediate violence."

The United Nations included a definition in its "Strategy and Plan of Action on Hate Speech" guidance, published in 2019, understanding hate speech as:

> [A]ny kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor. This is often rooted in, and generates intolerance and hatred and, in certain contexts, can be demeaning and divisive (United Nations, 2019).

However, not realizing the framework and the meaning of hate speech, some participants of the workshop raised questions that pointed to very different directions: Is exposing a corrupt politician or civil servant in the absence of a fair and just judiciary system hate speech? Where do journalists draw the line? Should they ignore different international and European definitions of hate speech that speak of hate based on discrimination along race, ethnic background, gender, and sexual orientation among others[2], and define one specifically for Lebanon that would focus more on community cohesion and avoiding sectarian divisions? Should we try to add the need to avoid hate based on class but not include political figures or the ruling ranks? This has led me to question the existence of a link between advocacy journalism, adopted by many journalists in Lebanon, and hate speech, and how it is widely defined. Should we make a clear distinction here between hate speech and advocacy journalism in any hate

2 For more on hate speech and hate crime evaluation in the EU, see this study: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf

speech definition or just mainly in countries that have a similar political context as Lebanon? These seem to be valid questions that we need to consider while promoting media spaces free from hate throughout the globe, especially under the conditions of non-functioning judiciary systems.

Fifteen journalists from various media outlets (print, broadcast, and online) participated in the workshop discussions. Many of them hinted that they rarely had to think of checking for hate speech in what they produced or wrote. Some shared their frustration with other colleagues who, while covering clashes between different communities within neighboring areas, were not aware of the political history and sectarian nuances of these areas, hence reporting without responsibility and inciting hostility among communities. Social responsibility came across in the workshops as a major need for journalists in Lebanon to consider while reporting. To achieve this, journalists need to stay away from sensational reporting. They need to avoid rushing to publish or broadcast. Some TV journalists in Lebanon believe that serving their political or sectarian sponsor or media organization owners with their writing and news production is, as a matter of fact, a responsible act. This is extended by the tendency of many journalists to be melodramatic, posting extreme and hateful content on social media to enhance their celebrity profiles and get more clicks and followers ("clickbait syndrome," as identified by the workshop participants).

How do we cut the umbilical cord between journalists and their political and sectarian leaders? How do we convince them that their loyalty as journalists should be to the public and not to their political and sectarian leaders? How do we remind them that being a journalist requires us to be skeptical, especially about our own political and sectarian affiliation? Being skeptical is crucial to detecting hate speech in the Lebanese context. Politicians in Lebanon have been known for using sectarian fears of "the other sect" to ensure their power continuity and preserve their political and economic interests. Accordingly, journalism in Lebanon is being instrumentalized by political institutions (see Harb, 2013; 2019).

Fact-checking is another pertinent need in the Lebanese context to detect and avoid hate speech. The amount of "fake news" that has dominated the media scene (including digital and social media) has flourished, particularly following the Beirut explosion. The Beirut case is a clear example of how fake news generate hate speech. Journalists, as a whole, might not be the source of hate speech, but ignorance of the historical context of the internal conflicts in Lebanon interprets

itself in coverage that incites hate and violence among neighboring sectarian communities. The "five-point test for hate speech," published by the Ethical Journalism Network (2015), is a good and helpful tool for journalists to use. One of the tool's points is to test the speakers' status before quoting them, sharing, or posting their speech.

Knowing the speech reach is also crucial in the Lebanese context, especially with journalists who rush to use tweets or Facebook posts as sources. Identifying the speech reach will help Lebanese journalists realize and detect hate speech in their journalism. Journalists who took part in the workshop discussions in Beirut consistently referred to other journalists' loyalty to their political and sectarian sponsors as the main obstacle to achieving hate-free reporting. It is true that Lebanon's media have always operated within proximity to the political sphere editorially and financially (see Dajani, 2019; Harb, 2013; 2019; Richani, 2016), but as in any other nation prone to conflict settings, in the absence of representative and independent journalists' unions in Lebanon, it is the obligation of journalists to attempt redeeming some of the good journalism Lebanese have demonstrated through tougher times (see Harb, 2011). To achieve this, solidarity among journalists is crucial.

What came out clearly from the workshops with journalists is that there is no clear distinction between polarization, bullying, libel, offensive language, and hate speech. There was little realization that not all polarization, libel, and offensive language can be classified as hate speech, but hate speech very likely involves all of these acts. Many journalists seem to struggle between issues regarding their margin of control over what they write and broadcast, including hate speech. Tools to help them tackle hate speech are important, but for many journalists, the priority lies in not being forced by their bosses to sensationalize their stories or rush into publishing and broadcasting before verifying the authenticity of their story. They are more concerned with their ability to do proper journalism and not become tools in the hands of media bosses who serve their own political and financial agendas. However, journalists are weary of "naming and shaming" politicians for their negligence and corruption, being labeled "hate speech."

The picture is not completely gloomy, as there are still journalists in Lebanon who stick to good journalism and its role in seeking truth and holding those in power accountable. One such journalist is Edmond Sassine of the Lebanese Broadcast Corporation (LBC). Sassine, in live coverage from a protest spot outside

Beirut following the port explosion, refused to open the airwaves to angry protestors. He was clearly heard on TV instructing his cameraperson not to move close to the protesters while live on air, as he did not want the protesters to use the live broadcast as a tool to channel more hate against other protesters from the opposite political affiliation, who were standing only a few meters away, which might have resulted in clashes erupting again between the two groups after being brought to a still by the Lebanese Army.

3 Hate speech in the Lebanese media—an ongoing challenge

Hate speech has been floating across the Lebanese media for some time, with journalists engaging in calls to physically silence those opposing their political views and affiliation. A very flagrant example is the article written by the chief editor of the Lebanese daily *Al Akhbar* newspaper Ibrahim Al Amin, which included direct threats to anti-Hezbollah activists, threatening to wring their necks (Annahar, 2021).[3] The threats that came out in 2012 resurfaced and were linked to the killing of anti-Hezbollah activist Lokman Slim in February 2021 in South Lebanon. Many of Slim's colleagues and friends saw a direct correlation between the newspaper's incitement to harm and Slim's assassination. The danger was linked to a list of names of those labeled as "traitors and collaborators" which was published alongside the threatening article.

Hate speech in the Lebanese media has many facets and is not necessarily bound to inciting political violence. In a fragile state where sectarian tension is high, spreading false news about mischiefs by one sect will generate hate against the accused, which might result in harm not necessarily on the individual level but on the collective level as well. However, we need to emphasize that advocacy journalism, led by investigative journalists in the country, should not be equated with hate speech. Investigations into the corrupt ruling class and their agencies are not a facet of hate speech, as those in power claim in an attempt to clamp down on media freedom. As one workshop participant put it, "In Lebanon, even if you decided not to broadcast or publish one politician's speech and not the other's, it will be seen as an act of hate speech." The scene is so complicated,

3 There is no direct translation in English of the phrase used in Arabic تحسسوا رقابكم

but the fear is that hate has become the dominant discourse. Journalists are increasingly becoming transmitters of the accumulated political and sectarian rivalry, translated in hate narrative. What is alarming in the Lebanese scene is that journalists may not be aware that they are being used as tools in a war of hate messages between different factions.

This alarming state of affairs of Lebanese journalism reminded me of the two Rwandan journalists sentenced for life in jail "for their roles in fueling the 1994 genocide in which 800,000 Tutsis and Hutus were murdered." How would a threat to slaughter rival political activists ("wring their necks"), aimed at opposition figures in Lebanon, differ from the Rwandan message that "the graves are not yet full"? What and who would stop these journalists in Lebanon who have willingly or unwillingly become messengers of hate speech?

The given example of calling for murder is not unique in the Lebanese media scene or exists only on one side of the political spectrum. Marcel Ghanem, the host of the Murr Television (MTV) talk show "It is About Time," has facilitated the spread of many false news about who is responsible and what caused the Beirut port explosion. In one of his episodes following the port tragedy, he built a theory based on a WhatsApp message he received from an anonymous viewer who claimed to have "confirmed insider information." He does not seem to hesitate to spread any news, even when those stories have not been verified to help implicate Hezbollah. His incitement against the Shia' political party has evolved to become incitement against the Shia' sect collectively in Lebanon.

Hate speech in Lebanon is not restricted to political and sectarian rivalry. The Syrian and Palestinian refugees have been the target of hate campaigns led by media organizations, fueled by journalists and demonstrated themselves, for example, by curfews imposed on Syrian refugees in many Lebanese villages or by equating Palestinians in Lebanon with the deadly Coronavirus (Khalil, 2020).

Many attempts have taken place over the years, mainly in the 21st century, to bring the Lebanese media to recognize hate speech as a contrast to their social responsibility role as journalists, including those initiated by the Maharat Foundation in Lebanon[4] and the Ethical Journalism Network in 2014 and 2016.

4 Maharat is a Lebanese NGO, established by a group of Lebanese journalists; it advocates the values of freedom of expression and respect for human rights in Lebanon. http://www.maharatfoundation.org/en

Further, the "Media Ethical Code for Promoting Civil Peace," facilitated by the Maharat and launched on June 25, 2013, by the "UNDP Peace-Building Project in Lebanon," was signed by 13 different Lebanese media organizations.[5] The initiative succeeded in raising awareness, but its impact had washed out at first signs of political tension in the country. Hence, the focus needs to be shifted to raising awareness among journalists themselves on the individual level in hopes that it might bring change on the collective level.

4 Suggestions for tackling hate speech

As mentioned earlier, not all journalists are forces of hate speech in Lebanon. Those who took part in the Beirut workshops in August believed in the need to avoid and tackle hate speech, and in the need for ethical reporting free of hate. Nevertheless, to achieve this, an assessment and redefinition of the core principles of journalism in Lebanon are required. Combating hate speech should be at the top of those role redefinitions. Meanwhile, Lebanese journalists need to realize that negative speech is not hate speech. Hate language and discourse can incite harm. Disinformation generates hate that incites harm. To fact-check, to not rush to publish or broadcast, to be sensitive to sectarian vulnerabilities, and to educate oneself of Lebanon's civil war history and geography to avoid triggering new hostilities between different communities have become necessary steps Lebanese journalists need to consider to avoid disseminating hate speech during times of crisis. Hate speech in the Lebanese context is speech that leads to harm, speech that is based on unverified and fabricated information, speech that uses sensational inflammatory language, and speech that feeds enmities among different publics.

The situation in Lebanon proves that there is a need to establish and implement a definition of hate speech that would take into account the socio-cultural and political context more strongly. Other countries in a similar situation of polarization could benefit from such definition. The Lebanese context is not unique, and journalists in similar political settings need to set some time aside to reflect on their profession and to be clear on defining their role as journalists in the society. This becomes even more crucial in times of crisis.

5 For more on this initiative, visit http://www.maharatfoundation.org/en/talkshows

Zahera Harb is the director of the MA International Journalism (MAIJ) and MA Media and Globalization (Erasmus Mundus) programs (International Journalism Studies Cluster leader) at the City, University of London, UK. https://orcid.org/0000-0002-7630-1171

References

Annahar. (2021, March 20). Assassination of Lukman Sleem. https://www.annahar.com/arabic/section/76-04022021110126010/سياسة

Dajani, N. (2019). *The media in Lebanon: Fragmentation and conflict in the Middle East*. IB Tauris.

Ethical Journalism Network. (2015). *Five points test of hate speech*. https://ethicaljournalismnetwork.org/5-point-test-for-hate-speech-english

Harb, Z. (2019). Journalism cultures in Egypt and Lebanon: Role perception, professionalism and ethical considerations. In M. Iqani & F. Resende (Eds.), *Media and the Global South: Narrative territorialities, cross-cultural currents* (pp. 119–139). Routledge.

Harb, Z. (2013). Mediating internal conflict in Lebanon and its ethical boundaries. In D. Matar & Z. Harb (Eds.), *Narrating 'conflict' in the Middle East: Communication practices in Palestine and Lebanon* (pp. 38–57). IB Tauris.

Harb, Z. (2011). *Channels of resistance: Liberation propaganda, Hezbollah and the media.* IB Tauris.

Khalil, B. (2020). *Al Joumhoriya and their moral fall.* Al Modon. https://www.almodon.com/media/2020/4/14/ةعيضولا-امتطقسو-ةيروموجلا-قديرج

Richani, S. (2016). *The Lebanese media: Anatomy of a system in perpetual crisis.* Palgrave.

United Nations. (2019). Strategy and plan of action on hate speech. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

Digital
Communication
Research.de

Recommended citation: Szczepańska, D., & Marchlewska, M. (2023). Unfree to speak and forced to hate? The phenomenon of the All-Poland Women's Strike. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 55–71). Digital Communication Research. https://doi.org/10.48541/dcr.v12.4

Abstract: This chapter explores abusive language's role when employed by a group with a lower social status whose rights are threatened by political authorities. We focus on the language of protest that emerged during the 2020 All-Poland Women's Strike, following a court ruling that almost totally banned legal abortions in Poland. Since some slogans used by these protesters could be interpreted as expressions of abusive language, we decided to analyze their meaning in a wider socio-political context. We show that women's use of vulgarisms and offensive language can serve as a tool of social and political change and that it may lead to empowerment. Moreover, given the cultural underpinnings of Polish society's gender-based social norms, we show that the use of abusive language may symbolize the process of redefining the traditional gender contract in Poland.

DOI 10.48541/dcr.v12.4

Dagmara Szczepańska & Marta Marchlewska

Unfree to Speak and Forced to Hate?

The phenomenon of the All-Poland Women's Strike

1 Introduction

On October 22, 2020, Poland's Constitutional Tribunal ruled that abortions were unconstitutional in cases where a fetus is diagnosed with a severe and irreversible birth defect. The judgment ended the so-called "abortion compromise," a law allowing voluntary pregnancy termination under certain circumstances, which had been in force since 1993 (Gliszczyńska-Grabias & Sadurski, 2021). The tribunal revised this piece of legislation after an official request had been filed by a group of 119 members of Parliament, all from socially conservative, right-wing parties (i.e., Law and Justice, Confederation, and the Polish People's Party—Kukiz'15). As a result, in Poland, abortion is now only allowed in cases of rape, of incest, or where a pregnant woman's life is at risk (Wigura & Kuisz, 2020). This change in legislation has translated into an almost total ban on voluntary pregnancy termination since, according to data gathered by the BBC, 96% of all legal abortions in Poland in 2018 were performed because of fetal defects (BBC, 2020, October 23b). Many mass street protests, organized by the All-Poland Women's Strike, flooded large and small Polish cities after the ruling was published, gathering up to 400,000 people in around 400 locations

across the country at its peak (Dziennik Gazeta Prawna, 2020). Despite the government's restrictions on public gatherings due to the ongoing COVID-19 pandemic (Garda World, 2020), these protests continued throughout the rest of 2020 and into 2021, since the new law only entered into force on January 27, 2021.

This legislative proposal was not the first attempt to change Poland's abortion law in recent years. The most memorable such efforts date back to May 5, 2016, when—by the right of a legislative initiative—a project called "Stop Abortion" was submitted for consideration to Parliament. However, it was ultimately rejected by Parliament, similar to previous attempts. Yet this project fomented remarkable social unrest (BBC, 2016). In September 2016, the Black Protest movement[1] emerged, and on October 3, 2016, the first All-Poland Women's Strike was organized, uniting approximately 100,000 participants across the country and creating a new political actor in the shape of a women's social movement (Gwiazda, 2016). This mass manifestation of opposition to the proposal to restrict the abortion law was unprecedented in Poland. Neither of the previous attempts had met with such outright displays of disapproval or united so many people. According to Korolczuk (2016), these former projects were less successful because their initiators lacked the support of the ruling political parties while, in 2016, the Law and Justice MPs regarded the "Stop Abortion" project favorably. In any case, the proposal failed, though the exact reason for this failure is subject to interpretation. Did it fail because of the citizen protests, or because the political opportunity structure at the time was unfavorable? Given that the ruling party did not submit the project itself and that the proposal included some controversial changes regarding the penalization of abortion, both reasons seem plausible (Korolczuk, 2016).

Nevertheless, the All-Poland Women's Strike clearly seems to have shed new light on civic participation among Poles. It has shown that many people—especially young women—are willing to fight for their rights and are ready to do so even less conventionally, choosing non-normative (i.e., street protests) rather than normative (e.g., electoral voting, petition signing) forms of political participation. Throughout the protests, language proved an important tool for expressing emotions that the situation evoked. The protesters often used swear words and taboo

1 The series of protests against the legislative proposal became known as the "Black Protest" because its participants wore black clothes as a symbol of mourning (see Korolczuk, 2016).

language to address political authorities and other individuals deemed responsible for the ruling, some of which we can classify as examples of abusive language or even hate speech. Given some authors' emphasis that hate speech is directed primarily at minority groups (see Sponholz in this volume), we decided to analyze an opposite situation, where expressions of abuse were employed by a systemically discriminated group toward a political majority.

In this chapter, we investigate whether the slogans used by the All-Poland Women's Strike protesters could have an emancipatory function in this particular socio-political context despite being vulgar or potentially offensive. We also analyze their role in boosting gender-related identities and protestors' willingness to act on behalf of their in-group. We believe this perspective is relevant because it explores abusive language and vulgarisms' role as instruments of political and social progress, leading to empowerment. Moreover, political forces change quickly, and a group that is now in power may soon fall out of grace and become a minority itself.

2 The discourse of protests: Abusive language or hate speech?

The concept of *discourse* is much broader than the mere use of language. Drawing on the reflections of Laclau and Mouffe (2014), discourse theory relies on the assumption that all objects and actions have a meaning, which depends on the historically constituted systems of rules. Therefore, *discourse* consists of all the social practices and systems of symbolic meanings that shape, and are shaped by, a given group of social and political actors in a given context (Laclau & Mouffe, 1987). As Gee (2015) explained, this range of semiotic practices, associated with the "social construction of knowledge," includes postures, ways of thinking, attires, and other artifacts that define people and shape our identity.

An inherent characteristic of protest movements throughout the entire world is the use of visual signs and banners (Linke, 1988). Over the past decade, this phenomenon has also spread from the streets into the world of social media: hashtags, profile picture frames, memes, and a variety of other resources to exhibit support for certain causes are becoming increasingly popular (see, e.g., Li et al., 2020). This manner of expressing one's opinion and clarifying demands reflects movements' performative character and is vital for constructing their wider meaning.

After all, these elements can be regarded as a movement's language or discourse, and they are constantly analyzed by sociolinguists and political scientists (Blackwood et al., 2016). Accordingly, an extensive body of related literature already exists—for example, on the languages of protest (Frekko, 2009; Kumar, 2001; Sonntag, 2003), labor (Wood, 2000; Woolfson, 2006), the environment (Linke, 1988), and women's movements (Mathonsi & Gumede, 2006; Ukeje, 2004).

Unsurprisingly, visual signs were also important for the 2016 anti-abortion-restriction protests in Poland. Their symbols were a coat hanger (commonly associated with self-induced, unsafe abortion), the color black (representing mourning and despair), and the slogan, "My body, my business" (referring to an individual's fundamental right to decide about their own health). Since then, the repertoire of signs used both in public spaces and on the internet has evolved significantly, including allusions to popular culture, historical events and figures, works of literature, and—ironically—the people supporting the opposed changes in abortion law. In 2020, the Polish protests' main focus gradually shifted from criticizing the tribunal's ruling as such to encompassing an overarching discontent with the government and the ruling Law and Justice Party. Beside banners directly referring to women's situation in Poland, other banners aimed to offend Law and Justice Party members and supporters. Some of these banners were explicitly offensive and used vulgarisms, while others were more subtle and used cultural references to imply politicians' lack of intelligence and complete misunderstanding of the contemporary world (Agence France-Presse, 2020). Therefore, considering this language's abusive character, we intended to verify whether they could be categorized as *hate speech*.

The difficulties in defining and operationalizing *hate speech* are extensively discussed in Liriam Sponholz's chapter in this volume; nevertheless, we aim to briefly explain how this concept is understood in this chapter and how it differs from the concept of *abusive language*. Since the *hate speech* concept is fluid and heavily depends on country-specific legal regulations, the international scientific community lacks a universal interpretation of the term *hate speech* that could apply to all historical and cultural contexts. Warner and Hirschberg (2012), for example, define it rather broadly as "abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation" (p. 19; see also Bilewicz et al., 2017). Given that this approach focuses on speech acts directed against minority groups, recent research has also drawn attention

to other practices that, at times, may intersect with hate speech while differing in fact. These practices include abusive language (Waseem et al., 2017), incivility (see Bormann & Ziegele and Masullo in this volume), offensive language (Davidson et al., 2017), and dangerous speech (see Benesch in this volume). Since we are analyzing a wide variety of linguistic acts in this chapter and they are not aimed at a systemically marginalized group, we believe using the broader concept of *abusive language* is more adequate for this purpose.

While considering the different examples of abusive language used during the 2020 women's protests, we consider some questions about the motivations that underpinned them, for example: *Which set of criteria should be used to evaluate the origin of an expression (e.g., hate or prejudice) and its consequences (e.g., invoking hatred or violence)? How important are historical and cultural antecedents in categorizing examples of abusive language?* In doing so, we were inspired by two studies addressing these issues. The first study focuses on the identification of a class of linguistic acts intersecting with hate speech, including the role of context. By bringing together legal and machine learning approaches focusing on the linguistic content of speech, Kennedy et al. (2018) propose the use of the term "hate-based rhetoric," which they define as language "that intends to—through rhetorical devices and contextual references—attack the dignity of a group of people, either through an incitement to violence, encouragement of the incitement to violence, or the incitement to hatred" (p. 8). The key concept in this approach—which emphasizes both the speaker's intention and the wider historical, cultural, ideological, and political context—is a person's dignity.

The second study seeks to create a typology synthesizing the variety of subcategories of *hate speech* presented in previous studies, and it demonstrates how these subcategories interrelate. Waseem et al. (2017) draw attention to two crucial factors while categorizing examples of *abusive language*—the target (i.e., whether it is "directed at a specific individual or entity or is directed toward a generalized group") and whether "the abusive content is explicit or implicit" (p. 78). When guiding researchers studying the topic, they emphasize that, by nature, abusive language is entirely subjective and that human annotators are influenced by existing social biases, which may lead them to disregard certain types of abuses. Such is the case of racism, for example, which is generally coded higher on hate speech scales than sexism (Waseem et al., 2017).

3 The All-Poland Women's Strike

Let us now briefly cite and analyze a few examples of speech acts used during the 2020 All-Poland Women's Strike. We will consider both their contextual allusions and the distinctions suggested by Waseem et al. (2017).

Among the most frequently used vulgar slogans were "Wypierdalać" (Get the fuck out) and "Jebać PiS" (Fuck PiS). The former expressed a desire for a change of government and had a deeper meaning, especially considering the wider socio-political context and the string of reforms carried out since Law and Justice first won a majority in the 2015 parliamentary elections. Changing the abortion law was, indeed, only the beginning of judicial reforms initiated soon after the party came to power in 2015, and one of its first tangible consequences (Ziółkowski, 2020). "Jebać PiS" was quickly adapted into various forms of expression: people, for example, chanted it at the top of their voices, replacing the chorus to the famous Eric Prydz song "Call on Me," which was played not only from special DJ trucks present at the protests but also from the open windows of passing cars and neighboring flats. After the non-protesting public's cry of outrage against protesters' use of vulgarisms, it was also turned into a visual symbol of ***** ***, sparking the rise of an informal group initiative later called the "Eight Star Movement." Both slogans are explicitly abusive in their expression and directed at a specific group, therefore exemplifying *abusive language*, according to the definition by Waseem et al. (2017). Moreover, they may be considered offensive by people who identify with the slandered party and exhibit a conservative worldview. Indeed, a prior analysis showed that political conservatism correlates with the perception of these two slogans as offensive (Szczepańska et al., 2021).

Furthermore, a selection of implicitly offensive expressions (i.e., that do not imply abuse through the use of vulgarisms but through more subtle, context-dependent allusions) is worth a closer look, too. These texts include, "Law and Justice cheated at the pregnancy test," "Law and Justice thinks *In Vitro* is a pizzeria," "Law and Justice believes Dąbrowski's Mazurka is a cake,"[2] "Law and Justice likes its own posts," "Every country has its own Voldemort," and "Even mephedrone has

2 "Dąbrowski's Mazurka" is the title of the Polish national anthem. In English, it is officially known by its incipit, "Poland Is Not Yet Lost," while a mazurka is a popular cake eaten during Easter in Poland (see Wikipedia, 2021, for further information).

better composition than Law and Justice." These texts are not explicitly offensive, do not use vulgarisms, and do not show direct aggression, but they are meant to be funny, and their message is clear: the ruling party is detached from what is happening in society and lacks the cognitive ability to comprehend these events.

Interestingly, the inclusion of pop-cultural references also points to the generational gap between the protesters and the authorities, since behaviors such as liking one own's posts on social media or misunderstanding foreign-sounding words are considered highly laughable by internet-savvy youths. Moreover, choosing in-vitro fertilization (IVF) for these texts is no coincidence, adding another layer of irony since the Law and Justice government recently removed infertility treatment from the National Health Program for 2021 to 2025 (Szczepańska & Klinger, 2021). As a prior analysis shows, even these rather implicit slogans were perceived as offensive by people who adhered to a conservative worldview (Szczepańska et al., 2021).

The Law and Justice members and supporters were not the only groups targeted by offensive slogans during the abovementioned protests, which is worth mentioning. One other such group was the Polish Catholic Church, whose hierarchs have long pressured political authorities to limit access to legal abortion in Poland (BBC, 2020, October 25). Among the most characteristic slogans against the church was "Kuria mać" (a wordplay using one of the most common Polish swear words, *kurwa mać*, and the word *kuria*, which designates the body of congregations, courts, and offices through which the Pope governs the Roman Catholic Church). Another example, hinting at an increasing number of pedophilia-related scandals, is: "If altar boys could get pregnant, abortion would be a sacrament."

Finally, the names of specific individuals—including political figures and activists—also appeared on protesters' banners. One such example is Jarosław Kaczyński, the leader of the Law and Justice party and deputy prime minister in charge of defense, widely considered the country's main powerbroker (AFP, 2019). The slogan "Moja pusia, nie Jarusia," which protesters frequently used, breaks several social norms. First, it employs the word pussy, an anglicism used by the younger generation and an explicit name for the vagina. Second, it uses a diminutive of Kaczyński's first name, which seems condescending, may be perceived as imputing a lack of professional competence, and can be considered disrespectful toward the addressee, especially given his age and public function. Moreover, not only was he addressed directly in such slogans, but his house on Mickiewicza

Street became the final destination of many street marches and swiftly became one of the most police-protected buildings in Poland. According to unofficial information published by the news portal Onet.pl, 82 police vans were stationed along that street on December 13, 2020—the anniversary of martial law's introduction by the Polish communist government in 1981 (Associated Press, 2020). The fact that so many police officers protected Kaczyński's house showed that the authorities took these direct threats during the protests very seriously.

Over time, more and more banners started appearing in what could be perceived as a race or competition to come up with even funnier, yet still thought-provoking, allusions to Poland's current political situation. Kaja Godek, the face of the anti-choice Life and Family Foundation (BBC, 2020, October 23a), also became an object of the protesters' mockery, best embodied by the slogan: "If this were *The Sims*, I'd remove Godek's pool ladder" (In *The Sims*, a popular life simulation video game, removing a ladder from a swimming pool is a means to kill a character). Though such slogans seem innocent and may be considered only jokes by some people, we would categorize these expressions as *hate speech* since they directly incite violence. Godek herself felt unsafe after the outbursts of protests in 2020, especially when some All-Poland Women's Strike activists published her private address and phone number on social media, together with the data of a few other well-recognized anti-choice figures. Godek started receiving hateful messages and phone calls, while offensive graffiti began appearing in her neighborhood. As a result, she filed an offense notification with the police and requested protection (World Today News, 2020). Given that the repertoire of slogans and direct actions during the 2020 protests was very wide, they also included some examples of hate speech. However, these examples were the exception, rather than the rule, and they highlighted the fact that access to voluntary pregnancy termination proved significant for many people.

4 The myth of the "Polish Mother"

Let us now reflect on the implications of employing this particular discourse during the protests against the Constitutional Tribunal's ruling on abortion law in Poland. As we have seen, these examples can undoubtedly be categorized as *abusive language* since they are not only offensive and directed at specific

people or groups but also perpetuate stereotypes and incite violence. However, since they form part of the specific discourse of a social movement and are, in themselves, a form of protest, they must be viewed from the perspective of their function within that phenomenon as well. The movement in question revolves around the issue of reproductive rights; therefore it extends to women's rights. Although a distinction between *human rights* and *women's rights* is still debated among scholars and politicians, many violations of women's human rights do indeed differ from the violations of men's rights and are intrinsically linked to their gender (Joachim, 2010). Therefore, we would like to point out other possible interpretations of this particular discourse's role, taking into account that it emerged from a women's rights movement.

Soon after the movement's outbursts, voices of outrage at the use of vulgarisms could be heard in the media, especially from more conservative politicians and public figures alike (Associated Press, 2020). Perhaps this outrage should be unsurprising since swearing has always been condemned and proscribed due to its association with subversion and its potential to undermine the status quo (Montagu, 2001). However, intensifying this sense of offense—besides the fact that these expressions were directed at one's own in-group—was that the words were uttered by women. After all, women of all ages around the world have been socialized to believe "swearing does not suit a lady" (Eagly, 1987; O'Neil, 2001). Politeness, compliance, emotionality, and care are but a few example characteristics of the stereotypical model of women's social role, and their gender-based ascription has been hotly debated since the appearance of Bem's sex role inventory—a tool used to self-measure one's perception of their own masculinity and femininity—almost half a century ago (see Bem, 1974).

The particular matter of Polish women's identity has been largely studied by scholars, such as Fuszara (2011), Graff (2008), Siemieńska (2008), and Titkow (2007). All of them have shown that a very traditional understanding of women's roles in society is prevalent among the Polish population. Although the percentage of women who choose motherhood and marriage as their primary life goals has been gradually diminishing since the first measurement in 1979, the percentage of women embodying the characteristics traditionally ascribed to the female gender remained at about 45% in successive studies (Titkow, 2007). Based on the results of various studies, Titkow (2007) argues that the tendency to reject traditional gender roles is actually stronger among Polish men than women, because women are still influenced by

the myth of the "Polish Mother," which is present in the national consciousness. This figure embodies what it means to be a woman in Polish society—a heroic, selfless individual capable of sacrifice for the sake of the family and the country (Imbierowicz, 2012). Moreover, while, due to history, Polish men can be said to have lost the distinctive features of their social identity after 1945, this very same political system can be said to have actually reinforced women's traditional role (Siemieńska, 2008). Therefore, some people view juxtaposing this figure of a devoted mother with a swearing, aggressive individual as offensive.

Research on linguistic impoliteness has already demonstrated that individuals learn to judge which content is offensive based on the cultural norms they are embedded in, and this evaluation is an ongoing process (Jay & Janschewitz, 2008). Therefore, if one's perception of a woman's social role complies with the traditional stereotype, that individual can be hypothesized to feel more offended by swear words uttered by women than more progressive people. A previous analysis confirmed this influence by showing that both political conservatism and support for the ruling positively predicted the conviction that the use of vulgarisms during the protests did not suit women (Szczepańska et al., 2021).

These results also align with previous findings on the positive relationship between support for hate speech prohibition and right-wing authoritarianism (RWA, see Bilewicz et al., 2017), a factor manifesting right-wing political views (Altemeyer, 1981), consisting of a willingness to submit to authorities, aggressiveness toward people who do not respect authoritarian values, and attachment to traditions decreed by authorities (Altemeyer, 1981). In their research, Bilewicz et al. (2017) found that people with high (versus low) RWA do not tolerate hate speech, probably because they perceive it as an extreme case of norm violation. This relationship should be even stronger when hateful expressions are used by women, who are traditionally perceived as not allowed to swear (Vingerhoets, 2013).

5 The language of a revolution

After the protesting women were condemned for using vulgarisms and told to refrain from using them in the public sphere, a new set of banners started appearing. They offered variations on the statements, "I am extremely aggravated," or, "I've been polite before." The vulgar language employed by the protesting

women in Poland can be argued to not only serve the purpose of expressing emotions—such as anger or outrage—but also manifesting change. This shift marks the rejection of the so-called gender contract regulating the social relationships and roles ascribed to men and women in Polish society (Fuszara, 2021). On one hand, "I've been polite before" may be interpreted as a direct reference to the protest itself and to the fact that resorting to standard, socially, and systemically acceptable measures of influencing abortion law (such as petitions or legislative initiatives) has been ineffectual, so protest must be taken to the streets. On the other hand, the statement may allude to the traditional perception of women as polite, symbolizing a repudiation of that image. Women no longer wish to be labeled as courteous and passive; they express a readiness to take control of their own fates and decide for themselves what type of social role they wish to fulfill. In fact, previous research found that swearing influences the swearer's perceived credibility, intensity, and persuasiveness. It can also help boost gender-related identity, promoting group bonding and solidarity (for a review, see Vingerhoets, 2013).

Using vulgarisms can be perceived as a form of empowerment, similar to the process of reclaiming the meaning of certain words traditionally meant to offend women. One such example is the word *kurwa*, which not only translates as *fuck* but also means *bitch* or *prostitute* and has been appropriated by the Coeducational Revolutionary Liberation Anarchist Union, whose Polish abbreviation reads KURWA. Related to the phenomenon of reclaiming certain words' meaning is also the term *witch*, a word with a pejorative connotation, denoting a disobedient woman with magical powers who is capable of influencing the world around her. Portraying powerful women as witches is still common in Western societies. Not only was Hilary Clinton labeled a witch during her 2016 US presidential election campaign but so was the United Kingdom's prime minister Theresa May (Miller, 2018). Classicist Mary Beard argues that stories of powerful women, such as the *Tale of Medusa*, are parables in which women are disempowered (Beard, 2017). Recently, the word *witch* is being reclaimed, though, both in various pop cultural productions, as well as historical studies (Buckley, 2017). During the women's protests in Poland, the word appeared in the slogan, "We are the granddaughters of the witches you couldn't burn"—drawing a connection between contemporary protesters and historical figures killed for their daring and independence. Protesting women intended to be perceived as rebellious and strong, unlike the stereotypically passive female figure present in the Polish national consciousness.

6 Context is everything

Thus, the more concepts such as *abusive language* or *hate speech* are researched, the more complex they become. Undeniably, hate speech can be harmful and even dangerous to the groups or individuals it addresses (Bilewicz & Soral, 2020; Soral et al., 2018). After all, linguistic acts are examples of discursive practices that shape social reality, and the normalization of certain expressions of anger may lead to the normalization of other forms of violence. However, its function must always be evaluated through the lens of *context*, a task requiring deeper analysis that brings together different scientific approaches (see Litvinenko in this volume). Moreover, hate speech heavily depends on the social status of the group it is employed by. In the case of the All-Poland Women's Strike, abusive language performed several roles, resulting directly from the movement's context and the fact that women are a group that has been historically marginalized and silenced in the public sphere (Houston & Kramarae, 1991). As we have illustrated, in this movement, abusive language served an emancipatory function and became a tool for social change. It drew attention to the issue of reproductive rights among not only the movement's supporters but also its opponents. Finally, it allowed for the creation of a sense of group identity among the protesters, who understood the allusions included in the slogans and laughed at the different jokes they included. Certainly, further research on the specific motivations of the individuals who participated in the protests could be revelatory, since some banners were also carried by men. We, therefore, encourage the use of interdisciplinary methods to better understand of the role that specific manifestations of abusive language play in the contemporary social reality.

Funding details

This work was supported by the Polish National Science Center (grant number 2019/35/B/HS6/00123) and the Polish Ministry of Science and Higher Education Grant (DIALOG Grant No. 0013/2019; financing period: 2019–2021).

Dagmara Szczepańska is a researcher at the Institute of Psychology, Polish Academy of Sciences, and a lecturer at Maria Grzegorzewska University, Poland. https://orcid.org/0000-0001-6216-5773

Marta Marchlewska is the head of the Political Cognition Lab at the Institute of Psychology, Polish Academy of Sciences, and a lecturer at the University of Social Sciences and Humanities, Poland. https://orcid.org/0000-0003-2807-5189

References

AFP. (2019, October 13). *Jaroslaw Kaczynski: Poland's polarising powerbroker.* France24. https://www.france24.com/en/20191013-jaroslaw-kaczynski-poland-s-polarising-powerbroker

Agence France-Presse. (2020, October 28). *Abortion protests attempting to "destroy" Poland: ruling party.* ABS-CBN. https://news.abs-cbn.com/overseas/10/28/20/abortion-protests-attempting-to-destroy-poland-ruling-party

Altemeyer, B. (1981). *Right-wing authoritarianism*. University of Manitoba Press.

Associated Press. (2020, October 27). Poland's PM defends abortion ruling, condemns protests. *Courthouse News Service*. https://www.courthousenews.com/polands-pm-defends-abortion-ruling-condemns-protests/

Associated Press. (2020, December 14). *Poland: Protesters march to home of PiS party leader on anniversary of communist crackdown.* Euronews. https://www.euronews.com/2020/12/13/anti-government-protesters-march-to-home-of-pis-party-leader-on-anniversary-of-communist-c

BBC. (2016, October 6). *Poland abortion: Parliament rejects near-total ban.* https://www.bbc.com/news/world-europe-37573938

BBC. (2020, October 23a). *Poland abortion ruling: Police use pepper spray against protesters.* https://www.bbc.com/news/world-europe-54657021

BBC. (2020, October 23b). *Poland abortion: Top court bans almost all terminations*. https://www.bbc.com/news/world-europe-54642108

BBC. (2020, October 25). *Poland abortion ruling: Protesters disrupt church services*. https://www.bbc.com/news/world-europe-54683735

Beard, M. (2017). *Women & power: A manifesto*. Profile Books.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2), 155–162. https://doi.org/10.1037/h0036215

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41(S1)*, 3–33. https://doi.org/10.1111/pops.12670

Bilewicz, M., Soral, W., Marchlewska, M., & Winiewski, M. (2017). When authoritarians confront prejudice. Differential effects of SDO and RWA on support for hate-speech prohibition. *Political Psychology*, *38*(1), 87–99. https://doi.org/10.1111/pops.12313

Blackwood, R., Lanza, E., & Woldemariam, H. (Eds.). (2016). *Negotiating and contesting identities in linguistic landscapes*. Bloomsbury Publishing.

Buckley, C. G. (2017, May 19). *Hag, temptress or feminist icon? The witch in popular culture*. The Conversation. https://theconversation.com/hag-temptress-or-feminist-icon-the-witch-in-popular-culture-77374

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515.

Dziennik Gazeta Prawna. (2020, October 29). *Komendant Główny Policji o protestach: Zatrzymano blisko 80 osób; prowadzonych jest ponad 100 postępowań ws. dewastacji* [Chief of Police about the protests: About 80 people were detained; over 100 cases of devastation are being investigated]. https://www.gazetaprawna.pl/wiadomosci/artykuly/1494857,komendant-glowny-policji-o-protestach-zatrzymano-blisko-80-osob-prowadzonych-jest-ponad-100-postepowan-ws-dewastacji.html

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Psychology Press.

Frekko, S. E. (2009). Signs of respect: Neighborhood, public, and language in Barcelona. *Journal of Linguistic Anthropology*, *19*(2), 227–245. https://doi.org/10.1111/j.1548-1395.2009.01032.x

Fuszara, M. (2011). Divorce in Poland. *Societas/Communitas*, *12*(2), 211–228.

Fuszara, M. (2021, February 1). *Feminizm stosowany, czyli bunt kobiet AD 2020* [Applied feminism, i.e. the rebellion of women] [Video]. YouTube. https://www.youtube.com/watch?v=l2FlUHBnjnk

Garda World. (2020, October 23). *Poland: Restrictive measures increased amid record rise in COVID-19 cases October 23 / update 18*. https://www.garda.com/crisis24/news-alerts/392331/poland-restrictive-measures-increased-amid-record-rise-in-covid-19-cases-october-23-update-18

Gee, J. (2015). *Social linguistics and literacies: Ideology in discourses*. Routledge.

Gliszczyńska-Grabias, A., & Sadurski, W. (2021). The judgment that wasn't (but which nearly brought Poland to a standstill): 'Judgment' of the Polish Constitutional Tribunal of 22 October 2020, K1/20. *European Constitutional Law Review, 17*(1), 130–153. https://doi.org/10.1017/S1574019621000067

Graff, A. (2008). The land of real men and real women: Gender and EU accession in three Polish weeklies. In C. Elliott (Ed.), *Global empowerment of women: Responses to globalization, politicized religions and gender violence* (pp. 191–212). Routledge.

Gwiazda, M. (2016). *Polacy o prawach kobiet, „czarnych protestach" i prawie aborcyjnym* [Poles on women's rights, "black protests" on abortion law]. CBOS. https://www.cbos.pl/SPISKOM.POL/2016/K_165_16.PDF

Houston, M., & Kramarae, C. (1991). Speaking from silence: Methods of silencing and of resistance. *Discourse & Society*, *2*(4), 387–399. https://doi.org/10.1177/0957926591002004001

Imbierowicz, A. (2012). The Polish mother on the defensive? The transformation of the myth and its impact on the motherhood of Polish women. *The Journal of Education Culture and Society, 3*(1), 140–153. https://doi.org/10.15503/jecs20121.140.153

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research*, *4*(2), 267–288. https://doi.org/10.1515/JPLR.2008.013

Joachim, J. (2010). Women's rights as human rights. In *Oxford research encyclopedia of international studies*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190846626.013.430

Kennedy, B., Kogon, D., Coombs, K., Hoover, J., Park, C., Portillo-Wightman, G., Mostafazadeh, A., Atari, M., & Dehghani, M. (2018). *A typology and coding manual for the study of hate-based rhetoric.* PsyArXiv.

Korolczuk, E. (2016). Explaining mass protests against abortion ban in Poland: The power of connective action. *Zoon Politikon*, *7*, 91–113. https://civitas.edu.pl/wp-content/uploads/2015/03/Zoon_Politikon_07_2016_091_113.pdf

Kumar, A. (2001). *Rewriting the language of politics: Kisans in colonial Bihar*. Manohar Publishers.

Laclau, E., & Mouffe, C. (1987). Post-Marxism without apologies. *New Left Review*, *166*(11–12), 79–106.

Laclau, E., & Mouffe, C. (2014). *Hegemony and socialist strategy: Towards a radical democratic politics*. Verso Trade.

Li, M., Turki, N., Izaguirre, C. R., DeMahy, C., Thibodeaux, B. L., & Gage, T. (2020). Twitter as a tool for social movement: An analysis of feminist activism on social media communities. *Journal of Community Psychology, 49*(3), 854–868. https://doi.org/10.1002/jcop.22324

Linke, U. (1988). The language of resistance: Rhetorical tactics and symbols of protest in Germany. *City & Society, 2*(2), 127–133. https://doi.org/10.1525/city.1988.2.2.127

Mathonsi, N., & Gumede, M. (2006). Communicating through performance: Izigiyo zawomame as gendered protest texts. *Southern African Linguistics and Applied Language Studies, 24*(4), 483–494. https://doi.org/10.2989/16073610609486436

Miller, M. (2018, April 7). *From Circe to Clinton: Why powerful women are cast as witches*. The Guardian. https://www.theguardian.com/books/2018/apr/07/cursed-from-circe-to-clinton-why-women-are-cast-as-witches

Montagu, A. (2001). *The anatomy of swearing*. University of Pennsylvania Press.

O'Neil, R. P. (2001). *Sexual profanity and interpersonal judgement*. LSU Historical Dissertations and Theses.

Siemieńska, R. (2008). Gender, family, and work: The case of Poland in cross-national perspective. *International Journal of Sociology*, *38*(4), 57–75. https://doi.org/10.2753/IJS0020-7659380403

Sonntag, S. K. (2003). *The local politics of global English: Case studies in linguistic globalization*. Lexington Books.

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Szczepańska, A., & Klinger, K. (2021). *Program walki z niepłodnością nie wypalił. Rząd nie zamierza dłużej wspierać prokreacji* [Infertility program didn't work out. The government does not borrow a place of procreation]. Dziennik Gazeta Prawna. https://serwisy.gazetaprawna.pl/zdrowie/artykuly/8061850,program-walki-z-nieplodnoscia-nie-wypalil-rzad-nie-zamierza-dluzej-wspierac-prokreacji.html

Szczepańska, D., Marchlewska, M., & Karakula, A. (2021). [Unpublished raw data on predictors of Poland's aborion ban support]. Institute of Psychology, Polish Academy of Sciences.

Titkow, A. (2007). *Tożsamość polskich kobiet: Ciągłość, zmiana, konteksty* [The identity of Polish women: Continuity, change, contexts]. Wydawnictwo Instytutu Filozofii i Socjologii PAN.

Ukeje, C. (2004). From Aba to Ugborodo: Gender identity and alternative discourse of social protest among women in the oil delta of Nigeria. *Oxford Development Studies*, *32*(4), 605–617. https://doi.org/10.1080/1360081042000293362

Vingerhoets, A. (2013). *Why only humans weep: Unravelling the mysteries of tears*. Oxford University Press.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (pp. 19–26). Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 78–84). Association for Computational Linguistics. https://doi.org/10.18653/v1/w17-3012

Wigura, K., & Kuisz, J. (2020, October 28). *Poland's abortion ban is a cynical attempt to exploit religion by a failing leader*. The Guardian. https://www.theguardian.com/commentisfree/2020/oct/28/poland-abortion-ban-kaczynski-catholic-church-protests

Wikipedia. (2021, March 3). *Poland is not yet lost.* https://en.wikipedia.org/wiki/Poland_Is_Not_Yet_Lost

Wood, A. G. (2000). Urban protest and the discourse of popular nationalism in postrevolutionary Mexico: The case of the Veracruz rent strike. *National Identities*, *2*(3), 265–276. https://doi.org/10.1080/713687698

Woolfson, C. (2006). Discourses of labor protest. *Atlantic Journal of Communication, 14*(1–2), 70–96. https://doi.org/10.1080/15456870.2006.9644770

World Today News (2020, October 28). *Kaja Godek: Due to the wave of aggression, I am not staying in my apartment*. https://www.world-today-news.com/kaja-godek-due-to-the-wave-of-aggression-i-am-not-staying-in-my-apartment/

Ziółkowski, M. (2020). Two faces of the Polish Supreme Court after "reforms" of the judiciary system in Poland: The question of judicial independence and appointments. *European Papers*, *5*(1), 347–362. https://doi.org/10.15166/2499-8249/362

Digital
Communication
Research.de

Recommended citation: Litvinenko, A. (2023). The role of context in incivility research. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 73–85). Digital Communication Research. https://doi.org/10.48541/dcr.v12.5

Abstract: *Incivility* is a concept with a wide scope of interpretations, ranging from impoliteness to aggressive and extremist speech. The definition of *uncivil speech* is highly context-sensitive, and this contextual sensitivity should be considered in future research. In this chapter, I argue that choosing to omit context from incivility research may result in the diffusion of authoritarian norms in online content regulation and negatively influence freedom of speech in different sociopolitical settings. I suggest considering four layers of context in incivility research: (1) sociocultural context (the macro level), (2) sociopolitical context (the macro level), (3) organizational context (the meso level), and (4) situational context (the micro level). I elaborate on each level's role in defining and regulating *uncivil speech*, and I conclude by suggesting paths for future research.

DOI 10.48541/dcr.v12.5

Anna Litvinenko

The Role of Context in Incivility Research

1 Defining incivility: Why context matters

Certain forms of incivility are widely considered to negatively influence deliberation and, ideally, to be eliminated from online discussions (Ng & Detenber, 2005; Chen, 2017). This black-and-white attitude toward uncivil speech has been increasingly implemented in national internet legislation (Mchangama & Fiss, 2019). In many countries, global platforms—such as Facebook or Twitter—are legally required to delete or quarantine uncivil speech. Dealing with immense amounts of data and using both manual and automated content moderation, such platforms tend to overregulate online communication (Gostomzyk, 2020).

In colloquial discussions of content moderation, the terms *incivility*, *harmful language*, and *hate speech* are often used interchangeably despite scholarly attempts to distinguish between harmful hate speech and other types of uncivil content (Paasch-Colberg et al., 2021). Some scholars (Chen, 2017; Sydnor, 2018) conceptualize *incivility* as a broad spectrum of speech phenomena ranging from impoliteness, profanity, and offensive language to hate or harmful speech (see Sponholz and Frischlich in this volume) and dangerous speech (see Benesch in this volume) – that is, language that can provoke violence, on the other hand. In this chapter, I will use the term *uncivil speech* bearing in mind the wide amplitude of possible interpretations of the concept in both colloquial use and scholarly works.

Since incivility, in a broad sense, is a type of communication that violates societal norms (see Bormann & Ziegele in this volume), whether a certain kind of speech actually violates a society's norms, as well as the extent to which it might harm participants in such communication, is subject to interpretation (van Mill, 2021). Many scholars have emphasized this concept's context sensitivity (Coe et al., 2014; Chen et al., 2019). However, studies of uncivil speech often neglect contexts' role, an unfortunate tendency since it might, for instance, lead to regulatory decisions with negative consequences for certain contexts. In this chapter, I explain the importance of considering different levels of speech context in both research on uncivil communication and internet regulation debates. A closer look into the sociocultural, sociopolitical, and situational circumstances of uncivil online communication can help explain not only uncivil speech's potential harm to participants but also the potential harm of banning this type of speech from a particular context.

In the past decade, scholars have observed some negative outcomes of the generalized approach to speech regulation on global media platforms. For instance, according to a study by Mchangama and Fiss (2019), Germany's *Network Enforcement Act* (NetzDG)—the so-called Facebook law—has triggered a wave of restrictive social media laws in flawed democracies and autocracies, which the authors assess as a "global cross-fertilization of censorship norms" (p. 6). They explore cases from 13 countries, including Russia, the Philippines, and Venezuela, where NetzDG has been cited as a justification to tighten online speech regulation. Depending on sociopolitical contexts, the restriction of incivility on global social media platforms can become a new censorship tool for authoritarian regimes, helping them curb remaining free speech enclaves in their countries, as in Russia (Litvinenko, 2020). Moreover, even in rule-of-law states, such as Germany, Facebook tends to overreact when assessing harmful speech and prefers to delete ambiguous content in order not to avoid legal liability (Gostomzyk, 2020). Platform content moderators are likely to censor borderline cases, such as satire or subcultural communication, as was the case with the German satire magazine *Titanik* (Martin, 2018). Due to social media platforms' global nature, content moderation decisions for one specific context are often applied to users in other contexts, and this global echo of national policies should be considered by both researchers and decision-makers.

2 Four layers of speech context

According to Teun van Dijk (2015), *speech context* is "how language users dynamically define the communicative situation" (p. 4). It comprises the following communication aspects: setting, participants, goals, and communicative interaction (p. 5). This definition emphasizes both the subjectivity of context assessments, which are conducted by participants themselves, and context's dynamic nature. Both of these aspects seem important when evaluating the harm of uncivil communication.

In addition to the micro-level context of speech—that is, the text itself—linguists also identify a macro context, "the broader social, political, and cultural conditions of discourse" (van Dijk, 2007, p. 6). All of these layers are obviously important in assessing uncivil speech's role in a particular situation. When discussing online speech regulation, distinguishing between context levels would obviously be appropriate since these context levels correspond to various levels of information regulation: political actors, communities, media, intermediary organizations, and personal users.

Figure 1: Four layers of speech context

I, therefore, suggest considering the following context levels in uncivil speech research (see Figure 1): (1) sociocultural context: the sociocultural roles of uncivil speech in a society (the macro level); (2) sociopolitical context: the political roles of uncivil speech in a specific context (the macro level); (3) organizational context: the norms of conduct on a media platform or in a community (the meso level); and (4) situational context: the role of uncivil speech in a particular communicative situation (the micro level). I elaborate on each of these context levels to highlight their importance in research and internet regulation.

3 Sociocultural context

In their comparative study of hate speech practices in India and Ethiopia, Pohjonen and Udupa (2017) emphasize the need to bring context into this debate "with an attention to user practices and particular histories of speech cultures" (p. 1173). In their case studies, they give compelling examples from speech cultures, such as the so-called "wax and gold" tradition in Ethiopia, which implies the importance of "complex double meanings, wordplay, and the use of metaphor" (p. 1185). This tradition is used alongside other elements to express offensive content in a disguised way. Prosecuting this kind of content in online discussions under the premise of incivility could alter the speech culture itself.

The example of swear language and its perception in different cultures is particularly suitable for illustrating the importance of the cultural context. Swearing is part of incivility, and it can be perceived as undesirable by different social groups. However, sensitivity to certain swear words differs from culture to culture. For instance, the "f-word" in English is largely tolerated in English-speaking media productions while, in some other languages, the corresponding word is considered unacceptable for professional media use. Thus, English swear words in movies are, in many languages, translated using euphemisms.

Subcultures that differ in some ways from the mainstream culture often develop a certain type of vocabulary that becomes a part of their identity and that might strike outsiders as uncivil. Consider, for instance, rap music and hip-hop culture, which are often accused of being sexist, racist, and violent (Rebollo-Gil & Moras, 2012). At the same time, researchers acknowledge that this subculture has

contributed to the emancipation of Black women and men worldwide (Rebollo-Gil & Moras, 2012; Loots, 2003).

Moreover, some minority subcultures tend to reclaim offensive language, reframing it within their community and then using it in a positive sense (Davidson et al., 2017; Allan, 2017). Van Aken et al. (2018) found out that several widely used automatic hate speech detection models show racial bias since they identify some dialectic words in African American English as offensive language. This finding shows that the use of automatic content moderation without considering sociocultural speech peculiarities leads to the discrimination of groups that already face discrimination and could endanger their speech culture by "flattening" it and forcing users to avoid wordplay, undertones, or hidden meanings.

Contexts' sociocultural and sociopolitical layers usually intertwine and influence one another. Political and legal context obviously plays a particularly noticeable role in determining the norms of incivility at a particular moment in a given country. At the same time, it is more flexible and subject to changes than the sociocultural layer of context, which concerns historically developed speech cultures.

4 Sociopolitical context

Political talk is central to the uncivil and hate speech debate since, usually, the most heated discussions arise around controversial political topics (Boberg et al., 2018). While some undesirable types of speech are universally accepted and defined as *hate speech* in international treaties (e.g., Council of Europe, 1997), exact interpretations of—for instance—racial discrimination or appeals to violence vary, depending on political and legal contexts. As Brown (2017) notes, the *hate speech* concept in colloquial use is often stretched to indicate any kind of offensive language and is used "in ways that merely serve political or ideological ends" (p. 453).

In authoritarian contexts, any type of incivility—especially from political opposition—can be treated as dangerous speech. For example, in the West, current debates about harmful or hate speech in online discussions are dominated by the threat of right-wing populist discourse, which challenges democratic principles of civility (Ebitsch & Kruse, 2021; Council of Europe, 2019). In (semi-)authoritarian contexts, which—according to The Economist's Democracy Index—constitute 55% of the world's polities (The Economist Intelligence Unit, 2020), liberal discourse is

often under attack under similar debates about harmful or hate speech in online communication. Facebook posts that would be considered moderately uncivil in an established democracy can be perceived as extremist speech in more restrictive political settings, which might lead to severe sanctions. Moreover, conservative political regimes often accuse liberal actors in their countries of violating so-called "traditional values" or offending an older generation with their online behavior. For instance, in Russia, the obscene sublanguage called "*mat*" has been banned from use in registered media since 2014 (Pilkington, 2014). This ban made the use of this type of language a gesture of political disobedience in certain cases. Consequently, *mat* has been widely used in alternative formats of news journalism produced by independent media professionals on YouTube (Bodrunova et al., 2021). This demonstrative loosening of language rules challenged conservative discourse of pro-state television. Our study of political talk on Russian YouTube has shown that politically motivated uncivil language plays an important role in not only fueling political discussions but also consolidating oppositional counter-publics (Bodrunova et al., 2021). In February 2021, a new law obliged social media platforms to filter *mat*. In such cases, under the threat of fines, global social media platforms might censor speech even more rigorously than state institutions, which are known in Russia to apply such laws rather selectively (Vendil Pallin, 2017). Social media platforms automatically detect undesirable word stems and can easily ban accounts or deny monetization of their content in cases where users decide to use swear language. In Russia's case, this law against swearing in social media can be considered a new tool to curb political dissent.

Another example of uncivil language's emancipatory political role is protests against conservative anti-abortion laws in Poland during 2020 and 2021 (see also Szczepańska & Marchlewska in this volume). Protesters' slogans and hashtags on social media often contained uncivil language that was clearly offensive to government officials—for instance, "Wypierdalać" [Fuck off] (Ciobanu, 2020). In this case, the use of uncivil language served as a tool for a women's rights movement to challenge conservative political discourse.

Suppressed groups' political emancipation usually accompanies the use of aggressive speech in the process of challenging hegemonic discourse since antagonism is an intrinsic part of political struggle (Mouffe, 2002). Democratic institutions can "diffuse the potential for hostility that exists in human societies by providing the possibility for antagonism to be transformed into 'agonism'"

(Mouffe, 2002, p. 58). However, in the cases of flawed democracies or authoritarian regimes, Mouffe's model of "agonistic pluralism" seems unfeasible, and suppressed communities are forced to voice their discontent antagonistically.

The examples noted in this section show that the global practice of banning online incivility might tighten authoritarian censorship, which is now reinforced by global social media platforms' algorithms. Ignoring differences in how a particular political setting affects incivility's role in a message might help globally diffuse authoritarian norms.

5 Organizational context

The organizational context level comprises formal and informal organizations, which provide rules of speech behavior for their participants—such as social media platforms' "discourse architecture" (Freelon, 2015) for users' communication, or companies and communities that specify their own rules of conduct on their websites and social media accounts.

Tech companies that own social media platforms play a significant role in regulating online speech, as well as creating online communities, making this context layer particularly important for research of online discussions. Several studies have shown that platform architectures and content curation mechanisms influence openness (Stockmann et al., 2020), as well as the civility of speech (Rösner & Krämer, 2016) on a platform. Sydnor (2018) explored perceptions of incivility across various media channels, concluding that "certain characteristics of media platforms can shape a message's perception as civil or uncivil" (p. 97).

Although social media platforms are usually global, their affordances can be used differently, depending on sociocultural and sociopolitical contexts; thus, the organizational level often intertwines with the macro level of context. Nagy and Neff (2015) introduced the term "imagined affordances," which highlights the importance of users' perception and employment of a tech platform's functional features. Thus, Telegram—which is known for its liberal approach to content filtering—has hosted very different types of alternative communities around the globe. While it has been celebrated as a tool of liberal protest movements in, for example, Belarus or Hong Kong (Litvinova, 2020), it is known in Germany as a meeting place for right-wing populists who circulate much hate

speech in their chats and channels (Ebitsch & Kruse, 2021). In any case, tech platforms' content moderation policies and affordances obviously play a role in shaping online discussions' tone.

By using a platform, users are expected to agree to its terms of use. On social media platforms' globalized level, the so-called informed consent to terms of use is often criticized as a mere formality since users barely read terms before agreeing (Dogruel, 2019). However, on online communities' level, depending on moderation styles (Strippel & Paasch-Colberg, 2020), users might have opportunities to negotiate the rules of conduct and their interpretation. The organizational level of context is, thus, more sensitive to interactions with users and their feedback, and it has the potential per se to be a truly democratic mechanism of speech regulation.

Chen et al. (2019) argue that allowing communities to formulate the norms of communication for themselves and define *incivility* could present an effective option for regulating uncivil communication. They give an example from the Civil Comments project, which existed from 2015 to 2017—a commenting plugin for news sites based on crowd-sourced moderation. It was designed to make each group of users who adopted the plugin define the standards of communication for themselves. Of course, communities can misuse users' freedom to create their own rules, as has been the case for the imageboard website 4chan. This website, where users can anonymously create message boards, is known for its lack of content moderation, which has resulted in racist and other aggressive content flourishing on the website (Arthur, 2020). This example shows that a liberal approach to content moderation should still be balanced by some basic rules of conduct and control mechanisms provided by a platform to avoid the spread of violent rhetoric.

Considering the meso level of context in incivility research will shed light on organizational actors' role in setting norms and shaping definitions of incivility in specific environments. It can also help create effective mechanisms of democratic regulation in online discussions.

6 Situational context

Situational context is linked to a specific communicative situation and, alongside other elements, accounts for participants' shared knowledge and personal communication styles. In other words, a group of friends might use a coded

language that would seem offensive to outsiders but would not be perceived as such by these friends.

In legal cases, participants' individual perceptions and sensibilities usually play a role in assessing the harm of hate speech (van Mill, 2021). As Sellars (2016) notes, "an epithet devoid of context may lead a scholar to see hate speech where the speaker, recipient, and subject of discussion may not" (p. 14). In his experimental study of individuals' perception of uncivil interactions among politicians, Muddiman (2017) demonstrated that political actors from the party with which a person associates are perceived as more civil than others. This finding proves that group identity influences perceived incivility in communication.

Overregulating uncivil speech could lead to increased self-censorship by users, forcing them to avoid ambiguity and playfulness in their communication. The micro level of context is closely connected with the sociocultural layer of context since knowledge of cultural codes is often required to assess immediate communicative situations.

7 Paths for future research

Considering different layers of context and their interplay certainly further complicates the analysis of uncivil speech. However, the examples presented in this chapter show that omitting context aspects from uncivil speech debates could seriously damage free speech worldwide. In the age of a "platform society" (van Dijck et al., 2018), we should recognize the effects that norms and concepts introduced in one context could have on other localities, as well as on transnational communities.

Introducing different layers of context to studies of uncivil speech opens new paths for future research. Scholars could, for example, compare the perceptions and roles of uncivil content in different political and sociocultural settings or examine online speech regulation's effects on different user groups' online behavior, including their willingness to participate in discussions and their levels of self-censorship. Comparative studies of uncivil speech across different contexts can, further, help reveal this content's effects on online discussions, depending on communicative situations. This revelation, in turn, would help explain the potential harm of various types of incivility, as well as the consequences of its banning in different settings.

Anna Litvinenko is a postdoctoral researcher at the Institute for Media and Communication Studies at Freie Universität Berlin, Germany. https://orcid.org/0000-0002-4029-0829

References

Allan, R. (2017). *Hard questions: Who should decide what is hate speech in an online global community?* Facebook. https://about.fb.com/news/2017/06/hard-questions-hate-speech/

Arthur, R. (2020). *The man who helped turn 4chan into the internet's racist engine*. Vice. https://www.vice.com/en/article/m7aap8/the-man-who-helped-turn-4chan-into-the-internets-racist-engine

Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication*, *6*(4), 58–69. https://doi.org/10.17645/mac.v6i4.1493

Bodrunova, S. S., Litvinenko, A., Blekanov, I., & Nepiyushchikh, D. (2021). Constructive aggression? Multiple roles of aggressive content in political discourse on Russian YouTube. *Media and Communication*, *9*(1), 181–194. https://doi.org/10.17645/mac.v9i1.3469

Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, *36*(4), 419–468. https://doi.org/10.1007/s10982-017-9297-1

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Palgrave Macmillan.

Chen, G., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, *5*(3). https://doi.org/10.1177/2056305119862641

Ciobanu, C. (2020). *Protests over abortion ruling widen, radicalise and threaten Polish government*. BalkanInsight. https://balkaninsight.com/2020/10/26/protests-over-abortion-ruling-widen-radicalise-and-threaten-polish-government/

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Council of Europe (1997). *Recommendation of the Committee of Ministers to member states on "hate speech" (No. R [97] 20).* https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505d5b

Council of Europe (2019). *News of the European Commission against Racism and Intolerance (ECRI).* https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/-/hate-speech-and-xenophobic-populism-remained-major-concerns-in-europe-in-2018

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In O. Varol, E. Ferrara, C. A., Davis, F. Menczer, & A. Flammini (Eds.), *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* – ICWSM 2017 (pp. 512–515). AAAI.

Dogruel, L. (2019). Privacy nudges as policy interventions: Comparing US and German media users' evaluation of information privacy nudges. *Information, Communication & Society*, *22*(8), 1080–1095. https://doi.org/10.1080/1369118X.2017.1403642

Ebitsch, S., & Kruse, B. (2021). *Wer den Hass verbreitet* [They who spread hate]. https://www.sueddeutsche.de/digital/steckbriefe-akteure-telegram-1.5278290

Freelon, D. (2015). Discourse architecture, ideology, and democratic norms in online political discussion. *New Media & Society*, *17*(5), 772–791. https://doi.org/10.1177/1461444813513259

Gostomzyk, T. (2020). Mehr oder weniger Meinungsfreiheit durch soziale Netzwerke? Technologischer Wandel und seine Herausforderungen für die freie Rede [More or less freedom of speech through social networks? Technological change and its challenges for free speech]. In T. Schultz (Ed.), *Was darf man sagen? Meinungsfreiheit im Zeitalter des Populismus* [What is allowed to say? Freedom of expression in the age of populism] (pp. 55–74). Kohlhammer.

Litvinenko, A. (2020). Social media in Russia: Between state and society. *Russian Analytical Digest*, *258*. https://css.ethz.ch/en/publications/rad/details.html?id=/n/o/2/5/no_258_media_capture

Litvinova, D. (2020). *"Telegram revolution": App helps drive Belarus protests.* AP News. https://apnews.com/article/international-news-technology-business-ap-top-news-europe-823180da2b402f6a1dc9fbd76a6f476b

Loots, L. (2003). Being a "bitch": Some questions on the gendered globalisation and consumption of American hip-hop urban culture in post-Apartheid South Africa. *Agenda: Empowering Women for Gender Equity*, *57*, 65–73. https://www.jstor.org/stable/4066391

Martin, D. (2018). Satire magazine back on Twitter after ban. *Deutsche Welle*. https://www.dw.com/en/german-satire-magazine-titanic-back-on-twitter-following-hate-speech-ban/a-42046485

Mchangama, J., & Fiss, J. (2019). *The digital Berlin Wall: How Germany (accidentally) created a prototype for global online censorship*. https://globalfreedomofexpression.columbia.edu/publications/the-digital-berlin-wall-how-germany-accidentally-created-a-prototype-for-global-online-censorship/

Mouffe, C. (2002). Which public sphere for a democratic society? *Theoria: A Journal of Social and Political Theory*, *99*, 55–65. https://www.jstor.org/stable/41802189

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*, 3182–3202.

Nagy, P., & Neff, G. (2015). Imagined affordance: Reconstructing a keyword for communication theory. *Social Media + Society*, *1*(2). https://doi.org/10.1177/2056305115603385

Ng, E. W., & Detenber, B. H. (2005). The impact of synchronicity and civility in online political discussions on perceptions and intentions to participate. *Journal of Computer-Mediated Communication*, *10*(3), JCMC1033.

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180. https://doi.org/10.17645/mac.v9i1.3399

Pilkington, A. (2014). *The rich, sweary sub-language that will protect Russia from Putin's latest crackdown*. The Conversation. https://theconversation.com/the-rich-sweary-sub-language-that-will-protect-russia-from-putins-latest-crackdown-26362

Pohjonen, A., & Udupa, S. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, *11*, 1173–1191.

Rebollo-Gil, G., & Moras, M. (2012). Black women and Black men in hip hop music: Misogyny, violence and the negotiation of (White-owned) space. *The Journal of Popular Culture*, *45*(1), 118–132. https://doi.org/10.1111/j.1540-5931.2011.00898.x

Rösner, L., & Krämer, N. C. (2016). Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media + Society*, *2*(3). https://doi.org/10.1177/2056305116664220

Sellars, A. F. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, *20*. https://papers.ssrn.com/sol3/papers.cfm?abstractid=2882244

Stockmann, D., Luo T., & Shen M. (2020). Designing Authoritarian Deliberation: How Social Media Platforms Influence Political Talk in China. *Democratization* 27 (2), 243–64. https://doi.org/10.1080/13510347.2019.1679771

Strippel, C., & Paasch-Colberg, S. (2020). Diskursarchitekturen deutscher Nachrichtenseiten [Discourse architectures of German news websites]. In V. Gehrau, A. Waldherr, & A. Scholl (Eds.), *Integration durch Kommunikation: Jahrbuch der Publizistik- und Kommunikationswissenschaft 2019* (pp. 153–165). https://doi.org/10.21241/SSOAR.68129

Sydnor, E. (2018). Platforms for incivility: Examining perceptions across different media formats. *Political Communication*, *35*(1), 97–116. https://doi.org/10.1080/10584609.2017.1355857

The Economist Intelligence Unit (2020). *Democracy Index 2020. In sickness and in health?* https://www.eiu.com/n/campaigns/democracy-index-2020/

Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). *Challenges for toxic comment classification: An in-depth error analysis.* arXiv. https://arxiv.org/abs/1809.07572

van Dijck, J., Poell, T., & de Waal, M. (Eds.) (2018). *The Platform Society*. Oxford University Press.

Van Dijk, T. A. (2007). Macro contexts. In U. Dagmar Scheu Lottgen & J. Saura Sánchez (Eds.), *Discourse and international relations* (pp. 3–26). Lang.

Van Dijk, T. A. (2015). Context. In K. Tracy, C. Ilie, & T. Sandel (Eds.), *International Encyclopedia of language and social interaction.* Wiley-Blackwell. https://doi.org/10.1002/9781118611463.wbielsi056

van Mill, D. (2021). Freedom of speech. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2021/entries/freedom-speech/

Vendil Pallin, C. (2017). Internet control through ownership: The case of Russia. *Post-Soviet Affairs*, *33*(1), 16–33. https://doi.org/10.1080/1060586X.2015.1121712

Digital
Communication
Research.de

Recommended citation: Ilori, T. (2023). Beyond the law: Toward alternative methods of hate speech interventions in Nigeria. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 87–109). Digital Communication Research.
https://doi.org/10.48541/dcr.v12.6

Abstract: Effective and rights-respecting hate speech interventions should encourage cohesion more than hate and inclusion more than division. Importantly, they should also guarantee the right to be and to express. Unfortunately, most countries, including Nigeria, lack effective hate speech interventions. This chapter considers specific ways to make hate speech interventions in Nigeria more effective while guaranteeing the protection of the right to freedom of expression, especially in the digital age. Thus, this chapter considers international human rights law and ideas on hate speech interventions, various hate speech interventions in Nigeria, and Nigeria's obligation to comply with international human rights instruments to achieve better results. It concludes that one of the best ways to ensure these interventions' effectiveness is combining alternative methods, such as strategic training, education, public awareness, and a multistakeholder approach.

DOI 10.48541/dcr.v12.6

Tomiwa Ilori

Beyond the Law

Towards alternative methods of hate speech interventions in Nigeria

1 Introduction

The law ought to not only define societal rules but also use these rules to solve societal problems. First, it ought to define the formal and golden rules a society abides by. Second, it ought to actively solve that society's problems to justify its relevance (Barret & Gaus, 2020; Biccheri, 2016). International human rights law, the system of rules that most sovereign states subscribe to, has established a consensus on prohibiting hate speech. Whether through emotional or physical violence, racial slurs or discrimination online or offline, hate speech is clearly forbidden by international human rights law through its various interpretations, justifying international human rights law's relevance (ICCPR, 1966, Art. 19–20; United Nations, 2019; Brown, 2015; Fino, 2020). However, despite this position, the use of violence through hate speech has risen (Tontodimamma et al., 2021; Futtner & Brusco, 2021; Deutsche Welle, 2020). While the definitions of *hate* or *prohibited speech* may vary in a language or context, they often share a purpose: to deter the use of any means of communication that may incite violence or discriminate against a set of protected characteristics (Mendel, 2012). Hence,

beyond a formal system of hate speech interventions, the law must devise effective ways to combat hate speech. However, in national contexts, the laws on hate speech—unlike international law—are ineffective and as such do not justify their relevance (Bakken, 2002; Fino, 2020).

Considering the law and its limited use yet central role in regulating hate speech, this chapter examines the viability of hate speech interventions in Nigeria. It considers the Nigerian context, the country's approach to hate speech regulation through laws, and how this approach has fared so far. It finds that major laws on hate speech interventions in Nigeria are ineffective due to their vague and excessive provisions that do not consider alternative intervention measures and, consequently, violate international human rights law.

In arriving at these findings, this chapter is divided into six broad parts. Part I introduces the chapter, while Part II considers various hate-speech intervention positions, including normative and theoretical approaches. Part III focuses on the *Declaration of Principles on Freedom of Expression and Access to Information* (ACHPR, 2019) in Africa as an opportunity to combat hate speech more effectively. It analyses the common principles of the various positions under Part II and how the *Declaration* offers a promising perspective on ensuring effective hate speech interventions. Part IV then applies these principles to the Nigerian context. As a result, Part V proffers possible solutions as rights-respecting and democratically viable hate speech interventions in digital-age Nigeria. Part VI concludes that for Nigeria to combat hate speech, its interventions must not be limited to mere criminalization of hate speech but must also include other alternative measures such as *strategic training*, *education*, *public awareness*, and *a multistakeholder approach*.

2 Major approaches to hate speech regulation in Africa

Primarily, hate speech is prohibited by international law (Scheffler, 2015). Various international human rights law instruments exemplify this prohibition through provisions for states to prohibit hate speech through law (United Nations, 1948a, Art. 3(c); ICCPR, 1966, Art. 19–20; ICERD, 1969, Art. 4; United Nations, 1948b, Art. 19; African Union, 1986, Art. 9). In addition to the law, there have been various explanations that analyzed hate speech and its regulations (Dworkin, 2009; Baker, 1989, 1997; Mill, 1859; Rawls, 1993). Both legal and scholarly approaches to hate

speech regulation, especially within the African human rights system, offer perspectives on how hate speech can be regulated (ACHPR, 2019). Practically, these perspectives should effectively use the law to actually prohibit hate speech.

For a working definition, Parekh's (2012) description of *hate speech* and its most obvious challenge—regulation—offer some clarity for this chapter. He states:

> Hate speech expresses, encourages, stirs up, or incites hatred against a group of individuals distinguished by a particular feature or set of features such as race, ethnicity, gender, religion, nationality, and sexual orientation. Hatred is not the same as lack of respect or even positive disrespect, dislike, disapproval, or a demeaning view of others... The difficult and much-debated question is whether it should be not merely discouraged by moral and social pressure but prohibited by law. Although law must be our last resort, its intervention cannot be ruled out for several important reasons (p. 55).

Parekh's view suggests that moral and social pressure are "alternative methods" of regulating hate speech and that legal intervention should only be the last resort. This position further suggests that, while the law plays its own roles, moral and social pressure are equally pertinent (Workneh, 2020; Benesch, 2014; Esimokha et al., 2019; Nkrumah, 2018; Breen & Nel, 2011; Asogwa & Ezeibe, 2020; Cassim, 2015). Consequently, a strong connection between the law's rhetoric and other alternative methods as forms of interventions on hate speech seems apparent. Therefore, considering the various perspectives on hate speech interventions in Africa is important.

2.1 Key standards of the normative approach to hate speech interventions

Various international human rights and humanitarian law instruments proscribe hate speech. Using different words yet a common purpose of prohibiting hate speech, and all their various mechanisms prohibit hate speech. Though all of these instruments prohibit hate speech, only the ICCPR and the ICERD explicitly mandate the traditional approach: the use of law to prohibit hate speech.

The most pressing concern of hate speech interventions is how not to violate the right to freedom of expression (Elbahtimy, 2014). This question is one of the greatest challenges facing governments and other stakeholders, including social media companies, in combating hate speech since social media companies have

been said to have a horizontal obligation to protect the right to freedom of expression (Nowak, 2005; Callamard, 2019; United Nations, 2018; Kaye, 2019).

A closer look at articles 19 and 20 of the ICCPR offers a perspective on balancing the contending needs for freedom of speech and freedom from hate speech. Article 20 of the ICCPR provides for three instances when the right to freedom of expression provided for under article 19 may be limited: (1) advocacy for discrimination, (2) hostility and violence based on protected characteristics, and (3) the incitement of imminent violence and propaganda for war. Combined with article 19(3), which allows for restrictions to free speech in order to protect others' rights, both articles form the fulcrum of international human rights law on limiting and regulating hate speech offline and online (Mendel, 2012, p. 420).

The relationship between these two articles can be understood in two major ways. First, the cumulative and conjunctive three-part test under article 19(3) (legality, proportionality, and necessity) provides a framework for the application of the limitations under Article 20. For example, a law on hate speech must not only be formulated with sufficient, precise meaning but it must also not provide a government with unfettered discretion, and it must be directed toward combating hate speech specifically as defined under international law (to protect the rights of others and public interests and use the least restrictive means for a specific aim) (United Nations, 2019). Article 19(3) presents the direct formula for solving the provisions of Article 20 or any claim for restricting the right to freedom of expression. The second relationship between these two articles is that, when they are combined, they ensure a high threshold of regulating the right to freedom of expression based on hate speech (United Nations, 2013).

One major challenge for hate speech jurisprudence under international human rights law is how to balance articles 19 and 20 of the ICCPR. This tension is obvious, especially when Article 20 suggests that "any advocacy"—which can include the right to freedom of expression as provided for under article 19(2)—may be restricted as prohibited speech, reading as a direct limitation of the right as provided for under article 19(2). However, the tension is more obvious even when applied narrowly to the prohibition of hate speech. What do human rights advocates mean when they demand that hate speech interventions must comply with international human rights law? While specific principles govern what qualifies as *hate speech*, these principles require a contextually sensitive application to be effective.

Considering the various international law texts above, what are the possibilities for effective and rights-respecting hate speech interventions in Africa?

2.2 Theories of hate speech interventions

Two major theoretical approaches address how best to regulate hate speech with respect to the right to freedom of expression: *absolutism* and *pragmatism*. Absolutism, which is popular in the United States' legal system, primarily argues against limitations of the right to freedom of expression (Dworkin, 2006, 2009; Baker, 1989, 1997). Its core argument is anchored on the claim that the freer the speech, the more open the society. Absolute hate speech intervention is further divided into two categories. First, *self-ordering absolutism* argues that a society will always "self-order" or "self-correct" in the course of debates and exchanges of ideas, whether popular or unpopular and through a free press (Mill, 1859). Second, *institutional absolutism* contends that, so far strong institutions are in place—such as the courts, law enforcement, and public service—higher guarantees protecting free speech are available by not using only the law (Rawls, 1993; Nickel, 1994).

Traditionally, pragmatism centers the use of hate speech interventions—laws and other measures that prohibit hate speech. Such interventions may include *traditional* or *non-traditional interventions*. Traditional interventions are the use of laws to combat hate speech, while non-traditional interventions are the use of other social methods, such as education, training, and public awareness (Workneh, 2020; Nkrumah, 2018; Cassim, 2015). Non-traditional interventions may also be called *alternative methods* or *alternative measures* of hate speech interventions.

Oftentimes, on one hand, most states adopt traditional interventions as they seek to combat hate speech through laws; on the other hand, most international law instruments use non-traditional interventions by referring to the use of other social methods in hate speech interventions. What distinguishes non-traditional interventions from other approaches is that it considers hate speech as socio-pathological and for this reason, requires more than criminalization and the legislative impulse to combat hate speech (Cassim, 2015).

Absolutist arguments against limiting speech through hate speech interventions are unsubstantiated since examples show that hate speech precipitates violence (Viljoen, 2005). Additionally, many societies are unable to "self-order" as

a result of weak democratic institutions that are meant to effectively lead such "self-ordering."

Traditional interventions equally pose a problem for the regulation of hate speech. Often, when the law or provisions that criminalize hate speech are not far-reaching in criminalization and punishments and are used to restrict the right to freedom of expression, they focus on corrective measures, rather than preventive methods (Scheffler, 2015, p. 82). However, in understanding hate speech as a social problem, the non-traditional intervention requires the law as a necessary tool to be combined with other social and alternative methods. Thus, hate speech interventions can be adjusted to various contexts while also protecting free speech and guarding against prohibited speech.

The normative and theoretical approaches are similar in providing the basis for assessing hate speech interventions in various contexts. The normative approach provides the prescriptive basis for balance between hate speech and the right to freedom of expression, while the theoretical approaches provide a more context-based and practical application of these laws. The normative framework convergently aims to prohibit hate speech, and the theoretical approach provides divergent perspectives on applying legal goals. A fine blend of both approaches is usefully exemplified in the African human rights system's reviewed *Declaration*.

3 The Declaration of Principles on Freedom of Expression and Access to Information in Africa and hate speech interventions

The African Commission on Human and Peoples' Rights (ACHPR) adopted the *Declaration* under its promotional mandate. The *Declaration* was made pursuant to Article 45(1) of the *African Charter on Human Rights* (African Union, 1986), which requires the African Commission to "promote human and peoples' rights, among others, by formulating and laying down principles and rules to solve legal problems relating to human and peoples' rights and fundamental freedoms upon which African States may base their legislation" (ACHPR, 2019).

In fulfilling this obligation, the *Declaration* was adopted to provide policy guidance for states' protecting the right to freedom of expression and access to information in the digital age under Article 9 of the *African Charter*.

Within the African human rights system, the *Declaration* benefited in its drafting from extensive consultations between April 2018 and October 2019, including perspectives from both the normative and theoretical approaches (ACHPR, 2019). As a result, it provides a prime example of a non-traditional intervention on hate speech in Africa. It is the only regional instrument that combines both forms of pragmatism described above. Principle 23 provides for the nature and extent of enforcing a human rights-focused hate speech intervention:

1. States shall *prohibit* any speech that advocates for national, racial, religious or other forms of discriminatory hatred which constitutes incitement to discrimination, hostility or violence.

2. States shall *criminalise prohibited speech as a last resort* and only for the most severe cases. In determining the threshold of severity that may warrant criminal sanctions, States shall take into account the:

    - prevailing social and political context;
    - status of the speaker in relation to the audience;
    - existence of a clear intent to incite;
    - content and form of the speech;
    - extent of the speech, including its public nature, size of audience and means of dissemination;
    - real likelihood and imminence of harm.

3. *States shall not prohibit speech that merely lacks civility or which offends or disturbs.*[1]

In demonstrating an example of non-traditional intervention, the principle addresses the specific nature of speech that is prohibited and considers at what point criminalization of this speech should occur—thus, criminalization is not the first step of intervening against hate speech. For example, 23(1) provides that states "shall prohibit" various categories of speech but does not refer to any specific method of regulation. This provision is presented before 23(3), on the criminalization of speech as a "last resort" because laws are not the only means of prohibiting speech and where such means arise, they would be suitable for only the *most severe cases*. The use of criminalization as a "last resort" readily suggests other

1 The italics here are added for emphasis by the current chapter's author.

methods than criminalization, including non-traditional methods should exist. Moreover, criminalization should be proportional since the narrow limitations of the right to freedom of expression matter. Thus, the *Declaration* applies both non-traditional and traditional approaches to hate speech interventions.

Further considering what a "last resort" criminalization of prohibited speech might look like, the *Declaration* considers six factors to assess whether certain speech is prohibited under 23(2): social and political context, the speaker-audience relationship, intention or motive, speech content, reach and likelihood, and proximity of harm. In concluding that prohibited speech has been used and should be criminalized, stakeholders should consider all six factors in enforcement (United Nations, 2013, p. 11). Hence, in regulating prohibited speech in African countries, non-traditional means must be considered before criminalization, which should only be used as a last resort, and such a last resort should be reserved for the "most severe cases."

To limit the right to freedom of expression based on hate speech interventions, such interventions require a high threshold of compliance due to the right's importance. Therefore, traditional and non-traditional approaches to hate speech prohibition should be combined. For example, a law on hate speech—even if it complies with the strict provisions of international human rights law—may be ineffective since hatred is reduced not only by imprisonment terms and fines but also through carefully chosen alternative methods that focus more on social dynamics than criminal elements. So, while a specific alternative method or a combination of alternative methods may genuinely teach about and prevent the dangers of hate speech, the law as a form of hate speech intervention should reinforce such alternative methods. Therefore, hate speech interventions in most severe cases should not be limited to imposing criminal sanctions but, also be used as a tool to mainstream alternative methods of hate speech interventions (Scheffler, 2015, pp. 96–98).

Perhaps closely related to traditional pragmatism on hate speech interventions is the proposition for a narrower application of hate speech, called *dangerous speech* (Benesch et al., 2018). In considering effective interventions for dangerous speech, Benesch (2014) notes:

> Most policies to counter inflammatory speech are punitive or censorious such as prosecuting, imprisoning, or even killing inflammatory speakers . . . these methods may curb freedom of expression, which must be protected, not only as a fundamental

> human right but also because denying it can increase the risk of mass violence, by closing off non-violent avenues for the resolution of grievances (p. 5).

This argument implies that mere the criminalization of dangerous speech, like all other forms of hate speech, is not only ineffective as an intervention but also often violates the right to freedom of expression and prevents opportunities to address hate speech through other measures.

The relationship between the international human rights instruments referred to above and the *Declaration* can be considered in two major ways. First, Article 9 of the *African Charter* can be used to strengthen international human rights law prescriptions on hate speech interventions, and to ensure this, the *Declaration* provided for specific obligations for African states on how to carry out such interventions under Principle 23 (United Nations, 2013, p. 11). Second, as a regional human rights instrument, the *Declaration* complements other international human rights systems. This second point is further reinforced by the window of complementarity permanently opened by virtue of Article 60 of the *African Charter*, which allows the African Commission to "draw inspiration from international law on human and peoples' rights."

These do not only tie the *Declaration* to the international human rights system, reinforcing its authoritative nature of issues with respect to the right to freedom of expression, but also grounds the *Declaration*'s provisions on prohibited speech in international human rights norms. This tie shows that any member state to the *African Charter*, including Nigeria, is free to consider either or both the international human rights system and the *Declaration* and still comply with international human rights law on hate speech interventions. This compliance is necessary because the "state bears the burden of demonstrating the consistency of such restrictions with international law with such restriction including those on the right to freedom of expression like hate speech" (ECOWAS, 2018, para 65).

4 Effectiveness of hate speech interventions in Nigeria

Since Nigeria gained independence in 1960, various interventions on hate speech have been implemented, mainly laws and rarely alternative methods. Recent interventions have arisen directly or indirectly through the 1999 constitution

(as amended), electoral laws, broadcasting laws, and proposed laws as hate speech interventions.

4.1 The 1999 Constitution (as amended)

Chapter 4 of the 1999 *Constitution of the Federal Republic of Nigeria* (as amended) provides for fundamental human rights. Section 39(2) limits the rights to freedom of expression, opinion, and the dissemination of ideas in its proviso. The proviso vests the power to limit the rights provided for under this section in the government. It empowers the government to carry out such limitations through an Act of the National Assembly in order to determine the ownership, establishment, and operation of any broadcasting station. Under Subsection 3, it provides that the basis for restricting the right to freedom of expression through laws must be "reasonably justifiable in a democratic society," requiring that government's powers to restrict the right be limited by reason and justifiability in a democratic system.

Section 45(1) provides for two other bases that apply to some rights contained under the chapter, including Section 39:

> (1) Nothing in sections 37, 38, 39, 40 and 41 of this Constitution shall invalidate any law that is reasonably justifiable in a democratic society
>
> (a) in the interest of defence, public safety, public order, public morality or public health; or
>
> (b) for the purpose of protecting the rights and freedom or other persons

Sections 39 and 45(1) suggest two categories of limitations with respect to the protection of the right to freedom of expression under the 1999 constitution. The first category is internal, contained in the provisions of sections 39(2) and (3), with (3) requiring that the limitation under (2) be reasonably justifiable. The second category is external, as contained in the provisions of Section 45.

Given the effects of both provisions' possible limitations to the right to freedom of expression through hate speech, such limitations must be provided for by law, be reasonably justifiable in a democratic society, and be proportionate toward protecting specific forms of public interests and the rights of others. These requirements demonstrate that, for example, in using law to limit the right to

freedom of expression through a hate speech law, under the Nigerian constitution, it must not only be specific toward such an aim but also be used reasonably in a democratic society. The ideals of a democratic society are respect for the rule of law, including finer practices such as respect for fundamental rights, limited government, periodic free and fair elections, the independence of the judiciary, and other crucial aspects of political power relations (Ihonvbere, 2000, p. 343).

4.2 Electoral Act

Section 95(1) of the *Electoral Act* of 2010 provides for the offenses of "abusive language directly or indirectly likely to injure religious, ethnic, tribal or sectional feelings." Subsection (2) further criminalizes "abusive, intemperate, slanderous or base or insinuations or innuendoes designed to likely provoke violent reactions or emotions." Subsection (7) further provides for various punishments, including imprisonment terms and fines.

Additionally, paragraph 7 of the *Code of Conduct for Political Parties* of 2013 provides that political parties and candidates shall refrain from "the use of inflammatory language, provocative actions, images or manifestation that incite violence, hatred, contempt or intimidation against another party or candidate or any person or group of persons on grounds of ethnicity or gender or for any other reason" (INEC, 2018).

The above provisions and language of the *Electoral Act* do not fall under the express limitations of *hate speech* under international human rights law. "Abusive language" may not be considered hate speech. It may be classified as a form of harm, but not hate speech, which does not include offensive or unpopular speech. The *Code of Conduct* provision may be further streamlined to cover the incitement of violence and advocacy for war and discrimination, based on the above-mentioned characteristics, while applying the various factors to be considered in determining whether hate speech has occurred.

4.3 Cybercrime Act

Section 26 of the *Cybercrimes (Prohibition, Prevention, Etc.) Act* of 2015 provides for racist, xenophobic, and genocidal offenses online. Section 26(1)(a–b)

criminalizes the production and sharing of racist and xenophobic material to the public. Additionally, the offense includes threatening anyone based on their race, color, descent, nationality, ethnicity, or religion. Section 26(1)(c), however, provides for the offense of insults based on these characteristics, while (d) criminalizes genocide or crimes against humanity. Each of these offenses carries various fines and imprisonment terms as punishments.

The provision of Section 26(1)(c) of the *Cybercrime Act* does not comply with international law in that "insults" are not covered under hate speech. For a speech to fall under the intendment of Section 26 as labeled, it must fall under the strict prescription of international human rights law, as explained immediately above.

4.4 Nigerian Broadcasting Code

Under the current *Nigerian Broadcasting Code*, paragraphs 3.0.2.1 and 3.0.2.2 provide that broadcasting incitement and hate speech is prohibited. It first paragraph states:

> No broadcast shall encourage or incite to crime, lead to public disorder or hate, be repugnant to public feelings or contain offensive reference to any person or organization, alive or dead or generally be disrespectful to human dignity (National Broadcasting Commission, 2016).

To the contrary, the code does not provide what constitutes *hate speech*. Words such as *public feelings*, *offensive reference*, and *disrespectful* do not convey a sufficient or precise meaning. For example, public feelings cannot be determined or contextualized, public feelings are not grounds for limiting the right to freedom of expression, and no international law instrument includes public feelings as bases for prohibiting hate speech.

4.5 National Commission for the Prohibition of Hate Speeches (2019)

The objective of the *National Commission for the Prohibition of Hate Speeches Bill* is to "promote national cohesion and integration by outlawing unfair discrimination, and hate speech." It seeks to establish a national commission for the prohibition of

hate speeches. The bill provides for various categories of offenses, including ethnic discrimination, hate speech, harassment on the basis of ethnicity, offense of ethnic or racial contempt, and discrimination through victimization and offense by companies and firms. It describes the offense of *hate speech* as the act of anyone who

> publishes, presents, produces, plays, provides, distributes and/or directs the performance of, any material, written and/or visual which is threatening, abusive or insulting or involves the use of threatening, abusive or insulting words or behaviour commits an offence if such person intends thereby to stir up ethnic hatred, or having regard to all the circumstances, ethnic hatred is likely to be stirred up against any person or person from such an ethnic group in Nigeria (National Commission for the Prohibition of Hate Speeches, 2019).

Of all the offenses provided for under the proposed law, only hate speech carries the punishment of life imprisonment, and where such speech results in death, it becomes punishable by death by hanging. Other offenses such as harassment on the basis of ethnicity and ethnic or racial contempt carry punishments of a five-year jail sentence or a fine of 10,000,000 nairas (26,000 US dollars) or both punishments if the accused is found guilty. Offenses by companies or firms carry the punishment of a one-year jail sentence or 2,000,000 nairas (5,000 US dollars) or both punishments if the accused is found guilty.

The bill, as an intervention, presents obvious irony since its hate speech provisions are not only excessive, non-compliant with international standards, and censorious (IPI, 2019; Media Rights Agenda, 2020; Tijani, 2019; Adibe, 2018) but also directly contravene its objectives to "promote national cohesion and integration" with its excessive punishments, including life imprisonment and death by hanging. Despite the provisions of the bill's Section 19 which considers other less intrusive means of combating hate speech, it fails to provide adequate clarity as a law, it is disproportionate and it does not demonstrate the necessity of its form of interventions.

Importantly, the Nigerian government is responsible for demonstrating its compliance with international law requirements limiting the right to freedom of expression (Land, 2020). This responsibility is that, aside from the use of such vague words as *insulting* or *abusive*, the bill did not provide for the contextual analysis of hate speech as under Principle 23(3). Additionally, it prescribes the outright criminalization of speech not as a last resort, while it also recommends death by hanging as

punishment when hate speech results in death. Therefore, its framing of *hate speech* and the necessary interventions do not demonstrate the consideration of other less intrusive means as one of the major tests for compliance with international law.

Currently, Nigeria lacks any elaborate provision in its Criminal Code Act or Penal Code Act—both laws that provide for criminal offenses of hate speech in Southern and Northern Nigeria, especially as prescribed under international law. Related to the prohibition of hate speech are the provisions of Section 417 of the Penal Code (Northern States) Federal Provisions Act. It provides for an offense of endangering public peace by exciting hatred among classes. Moreover, currently, no policies offer guidance on online hate speech in Nigeria. Therefore, the Nigerian government faces at least three urgent needs to review its hate speech interventions.

First, it must review all laws and existing policies to align them with human rights principles because for Nigeria to thrive, given its current constitution, it must allow for more speech and not less. This process involves aligning various laws with international human rights provisions, such as those provided for in the *Declaration* to ensure more debates and a tolerant system.

Second, since the laws have been in use for the most time and have not effectively reduced hate speech, more alternative methods should be considered (Scheffler, 2015; Bakken, 2002), such as the use of the law to achieve evidence-based policy-making on hate speech interventions. Rather than using the law simply as a criminalization tool, it could be used to devise normatively creative ways to combat hate speech.

Third, all forms of intervention must be truly transparent and inclusive to accommodate the realities of combating hate speech, especially in the digital age. This goal can be accomplished by considering the various recommendations in the subsequent parts of this chapter. They would assist in solving the twin challenge of ensuring more speech while protecting against harmful speech.

As this chapter has explained above, especially under the international human rights law, any form of hate speech intervention should aim to stop the spread and impact of hate speech. Any other aim could endanger human rights protections and democratic development. Hate speech interventions should not focus on using vague words to criminalize hate speech (United Nations, 2012). Rather, they should adopt creative means beyond the law, accommodating diverse perspectives to form various systems of rules that can both prevent hate speech and, simultaneously, protect free speech.

5 Beyond the law: Hate speech and alternative methods of intervention in Nigeria

Without the right policy to justify the use of criminalization for serious hate speech offenses, governments lack legitimacy and legality in their use of most hate speech legislation (Egbunike, 2019; Nkanga, 2016; Busari, 2020; Nyathi, 2018). One requirement of the three-part test in limiting speech is that such laws must be formulated with sufficient precision so that everyone affected by those laws can understand them. The rise in hate speech in Africa offline and online does not necessarily suggest that perpetrators of hate speech fully understand its impacts.

In suggesting various alternative approaches to curbing hate speech, using Rwanda and Kenya as case studies, Scheffler proposed five ways to divide responsibilities across stakeholders (Scheffler, 2015, pp. 89–94). These stakeholders include government and state officials, the public, media, monitoring institutions, and the international community. This chapter takes a slightly different approach but includes some of these methods as other means of conducting hate speech interventions in Nigeria.

Using various ways to resolve the three issues highlighted above, after policy review, more stakeholders should be included to devise alternative methods for hate speech interventions in Nigeria (Ibrahim, 2021, p. 200). These methods will not only allow for the legitimacy of such interventions but also practically combat hate speech and increase the prospects of tolerance. Some such alternatives include *strategic training*, *education*, *public awareness*, and *a multistakeholder approach*.

5.1 Strategic training

Various stakeholders should be prioritized for training, especially in the public sector, to advance an incisive public-facing understanding of hate speech in Nigeria. While ensuring this understanding is primarily the responsibility of governments and their institutions, other stakeholders such as social media platforms, academia, and civil society should be willing to collaborate in this regard. Considering the strategic role played by some sub-sectors, such as the administration of justice, education, and internal affairs, designing targeted training programs fit for the purpose of each of these sub-sectors is salient.

These programs can be accomplished by identifying and provisioning specific resources that focus on hate speech's social dynamics. Since hate speech seeds violence, these proximate stakeholders who are most likely to make decisions on the public's behalf should have their training manuals updated occasionally, and ensure mandatory, continuous education including understanding the various dynamics of hate speech and its interventions in Nigeria. For example, various judicial institutions involved in the continuous education of magistrates, judges, and other judicial officers should incorporate the various dynamics of how hate speech plays out in today's society like contexts where such speech was used, the spread and impacts of such speech and others.

5.2 Education

States should mainstream academic modules—such as literary studies, civic studies, and history into academic curricula, making them stand-alone, compulsory subjects at the primary, secondary, and post-secondary levels. This measure would afford a fuller understanding of the various contexts that might influence hate speech in society through more objective formal education. It should focus on how the humanities preserve the society through social methods and correct hate speech through carefully planned educational systems that encourage thinking, beyond merely remembering. Additionally, as a public policy, governments should consider various promotional materials that can assist in contextualizing hate speech in various communities.

5.3 Public awareness

To stem hate speech through alternative methods, stakeholders such as the government, social media platforms, academia, and civil society in Nigeria should consider raising more awareness about the dangers of hate speech and the various contexts in which it might occur. Such awareness should be informed by comparative and contextual examples of hate speech. Clarifying the legal and social impacts of hate speech is vital, especially with respect to international law. In communicating these impacts, various institutions such as government

ministries and institutions should collaborate with other stakeholders. For example, the National Orientation Agency (NOA) and the National Human Rights Commission (NHRC) could coordinate stakeholders' activities to develop and implement a nationwide campaign on hate speech, according to its mandate (National Orientation Agency, 1993). This campaign may serve as a precursor to designing a hate speech policy for Nigeria. This campaign should draw on various stakeholders to design communicative, community-friendly, and easy-to-read facts about hate speech. For example, providing public educational materials in more minority languages that are designed for such a campaign in Nigeria would greatly complement a focus on the country's major languages.

5.4 A multistakeholder approach

A democratic, inclusive, and participatory system is central to balancing harmful speech and free expression in Nigeria. Such a system should accommodate as many stakeholders as possible to update policies on hate speech in Nigeria. Government officials, government institutions, the private sector (including telecommunication companies and social media platforms), civil society, academia, linguists, journalists, and traditional rulers at various levels should together determine the course of a nationwide policy on hate speech.

6 Conclusion

In order to lead with more effective interventions, key stakeholders including the Nigerian government, social media platforms, academia, and civil society should pay more attention to a combination of the methods presented above. For example, in regulating hate speech in Nigeria and other African countries, social media platforms can adopt internationally set human rights standards while also paying close attention to varying contexts. Such approaches would include encouraging education over permanent sanctions. Social media companies seeking to comply with national laws and their community guidelines is insufficient, given the overarching need to apply international human rights laws and social methods to regulate hate speech (Global Network Initiative, 2020). This application provides

a more objective basis for social media companies to push back against censorious practices and effectively help combat hate speech. Now, with increased reliance on technologies, social media companies must actively mainstream international human rights law into their algorithms while setting finer policy mandates through strategic training, education, public awareness, and a multistakeholder approach to collaborate on social methods that systematically combat hate speech.

The essence of clear and narrow restrictions on the right to freedom of expression—especially under international human rights law not only protects against harms such as hate speech, but also ensures that such protections do not render the right nugatory. In striking a careful balance between these two seemingly contrasting needs, the requirement to determine whether speech must be restricted must consider the least intrusive means. As Mendel (2010) states, "measures to protect the right must be rationally connected to the objective of protecting the interest, in the sense that they are carefully designed so as to be the least intrusive measures which would effectively protect it" (p. 18).

This agrees with the provisions of Principle 23 of the *Declaration*, which not only regards criminalization as necessary in serious cases but also considers its use only as a last resort. These provisions emphasize alternative approaches to criminalization or the law in combating hate speech in Nigeria.

Hate speech is a multifaceted social phenomenon, and it has been studied as a socio-pathological trait. Therefore, it has become even more amplified, given the rise of technologies. As a result, more normatively creative interventions on hate speech that do not only prevent it fundamentally but also arrest its harm to the society are needed. This chapter has shown that preventing and arresting such harm is possible but stakeholders must seek solutions beyond the law. "Beyond the law" here does is not mean *outside the law* but rather, *a creative use of the law as a tool to protect speech while combating its harmful aspects*. More definitive ideas on such creative normative interventions that combat hate speech effectively will be needed, so this chapter looks to spark more conversations about these ideas.

Thus, to ensure effective interventions against hate speech, Nigeria should consider alternative measures. These approaches consider not just education but the type that informs about the histories and dangers of hate speech, not just training but also a focus on proximate stakeholders in hate speech interventions and their implementation, and not just the obvious and easy approach of criminalization but also approaches treating hate speech as a social phenomenon.

Tomiwa Ilori is a researcher at the Centre for Human Rights, University of Pretoria, South Africa, and at the Expression, Information and Digital Rights Unit of the Centre. https://orcid.org/0000-0002-2765-3103

References

Adibe, J. (2018, October 2). *Should the law be used to curb hate speech in Nigeria?* Brookings. https://www.brookings.edu/blog/africa-in-focus/2018/10/02/should-the-law-be-used-to-curb-hate-speech-in-nigeria/

African Union (1986, October 21). African Charter on Human and Peoples' Rights. https://au.int/en/treaties/african-charter-human-and-peoples-rights

ACHPR – African Commission on Human and Peoples' Rights. (2019). *Declaration of Principles on Freedom of Expression and Access to Information in Africa 2019.* https://www.achpr.org/legalinstruments/detail?id=69

Asogwa, N., & Ezeibe, C. (2020). The state, hate speech regulation and sustainable democracy in Africa: A study of Nigeria and Kenya. *African Identities*, Advance online publication. https://doi.org/10.1080/14725843.2020.1813548

Baker, E. C. (1989). *Human liberty and freedom of speech*. Oxford University Press.

Baker, E. C. (1997). Harm, liberty, and free speech. *Southern California Law Review*, *70*, 979–1020.

Bakken, T. (2002). The effects of hate crime legislation: Unproven benefits and unintended consequences. *International Journal of Discrimination and the Law*, *5*(4), 231–246. https://doi.org/10.1177/135822910200500404

Barret, J., & Gaus, G. (2020). Laws, norms and public justification: The limits of law as an instrument of reform. In S. A. Langvatn, M. Kumm, & W. Sadurski (Eds.), *Public reason and courts* (pp. 17–19). Cambridge University Press. https://doi.org/10.1017/9781108766579.009

Benesch, S. (2014). *Countering dangerous speech: New ideas for genocide prevention*. Harvard Berkman Klein Centre for Internet and Society. https://doi.org/10.2139/ssrn.3686876

Benesch, S., Glavinic, T., Manion, S., & Buerger, C. (2018). Dangerous speech: A practical guide. *Dangerous Speech Project*. https://dangerousspeech.org/guide/

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Breen, D., & Nel, J. A. A. (2011). South Africa – a home for all? The need for hate crime legislation. *South African Crime Quarterly*, *38,* 33–43. https://doi.org/10.17159/2413-3108/2011/v0i38a851

Brown, A. (2015). *Hate speech law: A philosophical examination.* Routledge.

Busari, K. A. (2020). *A tsunami of legal limitations of press freedom? An analysis of Nigerian laws that impact journalists' freedom*. Unpublished Master's dissertation, Vrije Universiteit Brussel.

Callamard, A. (2019). The human rights obligations of non-state actors. In R. F. Jørgensen (Ed.), *Human rights in the age of platforms* (pp. 191–225). MIT Press.

Cassim, F. (2015). Regulating hate speech and freedom of expression on the Internet: Promoting tolerance and diversity. *South African Journal of Criminal Justice*, *28*(3), 303–336. https://hdl.handle.net/10520/EJC-60bf00235

ECOWAS. (2018). *Suit no: ECW/CCJ/APP/10/I5 Judgment No: ECW/CCJ/JUD/31/18*, paragraph 65. https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2019/09/ECW_CCJ_JUD_31_18.pdf

Deutsche Welle. (2020, October 27). *As hate speech online increases, so has the resistance*. https://www.dw.com/en/as-hate-speech-online-increases-so-has-the-resistance/a-55411555

Dworkin, R. (2006). A new map of censorship. *Index on Censorship*, *35*, 130–133. https://doi.org/10.1080/03064220500532412

Dworkin R. (2009). Foreword. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. v–ix). https://doi.org/10.1093/acprof:oso/9780199548781.001.0001

Egbunike, N. (2019, July 17). *How Nigeria uses the law to repress free speech: The case of journalist Jones Abiri.* Global Voices. https://globalvoices.org/2019/07/17/how-nigeria-uses-the-law-to-repress-free-speech-the-case-of-journalist-jones-abiri/

Elbahtimy, M. (2014). The right to be free from the harm of hate speech in international human rights law [Working Paper]. https://www.repository.cam.ac.uk/handle/1810/245215

Esimokha, G., Bobmanuel, K. B., & Asaolu, O. (2019). Perception of Nigerians on hate speech bill (a study of Akungba-Okoko residents, Ondo State). *IOSR Journal of Humanities and Social Science*, *24*(11), 59–67. https://www.iosrjournals.org/iosr-jhss/papers/Vol.%2024%20Issue11/Series-3/J2411035967.pdf

Fino, A. (2020). Defining hate speech. *Journal of International Criminal Justice*, *18*(1), 31–57. https://doi.org/10.1093/jicj/mqaa023

Futtner, N., & Brusco, N. (2021). *Hate speech is on the rise*. Geneva International Centre for Justice. https://www.gicj.org/images/2021/Hate_Speech_is_On_the__Rise-FINAL_1.pdf

Global Network Initiative. (2020). Content moderation and human rights: Analysis and recommendations. https://globalnetworkinitiative.org/wp-content/uploads/2020/10/GNI-Content-Regulation-HR-Policy-Brief.pdf

Ibrahim, H. (2021). The efforts at countering hate and dangerous speech. In J. Ibrahim & Y. Z. Yaú (Eds.), *Context and content in hate speech discourse in Nigeria* (p. 200). Centre for Information Technology and Development.

Ihonvbere, J. O. (2000). How to make an undemocratic constitution: the Nigerian example. *Third World Quarterly*, *21*(2), 343–366.

INEC – Independent National Electoral Commission. (2018). *Code of Conduct for Political Parties*. https://www.inecnigeria.org/wp-content/uploads/2018/10/Code_of_Conduct_For_Political_Parties_Preamble.pdf

ICERD – International Convention on the Elimination of All Forms of Racial Discrimination. (1969, January 4). https://www.ohchr.org/en/professionalinterest/pages/cerd.aspx

ICCPR – International Covenant on Civil and Political Rights. (1966, December 16). https://treaties.un.org/doc/publication/unts/volume%20999/volume-999-i-14668-english.pdf

IPI – International Press Institute. (2019, December 11). *Nigeria must amend vague "hate speech" bill*. https://ipi.media/nigeria-must-amend-vague-hate-speech-bill/

Kaye, D. (2019). *Speech police: The global struggle to govern the Internet*. Columbia Global Reports.

Land, M. (2020). Against privatized censorship: Proposals for responsible delegation. *Virginia Journal of International Law*, *60*(2), 409–410. http://dx.doi.org/10.2139/ssrn.3442184

Media Rights Agenda. (2020, April 24). *MRA sues National Assembly over hate speech bill.* IFEX. https://ifex.org/mra-sues-national-assembly-over-hate-speech-bill/

Mendel, T. (2010). *Restricting freedom of expression: Standards and principles.* Centre for Law and Democracy. http://www.law-democracy.org/wp-content/uploads/2010/07/10.03.Paper-on-Restrictions-on-FOE.pdf

Mendel, T. (2012). Does international law provide for consistent rules on hate speech? In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: Rethinking regulation and responses* (pp. 417–429). Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.029

Mill, J. S. (1859). *On liberty*. J. W. Parker & Son.

National Broadcasting Commission. (2016). *NBC Code* (6th ed.). https://www.nta.ng/wp-content/uploads/2019/09/1494416213-NBC-Code-6TH-EDITION.pdf

National Commission for the Prohibition of Hate Speeches (2019). A bill for an act to provide for the prohibition of hate speeches and for other related matters. https://placbillstrack.org/upload/Hate-Speech-Bill.pdf

National Orientation Agency. (1993). National Orientation Agency Act. http://www.noa.gov.ng/wp-content/uploads/2018/03/NATIONAL-ORIENTATION-AGENCY-ACT.pdf

Nickel, J. W. (1994). Rethinking Rawls's theory of liberty and rights. *Chicago-Kent Law Review*, *69*(3), 763–885.

Nkanga, P. (2016, September 21). *How Nigeria's cybercrime law is being used to try to muzzle the press.* CPJ. https://cpj.org/2016/09/how-nigerias-cybercrime-law-is-being-used-to-try-t/

Nkrumah, B. (2018). Words that wound: Rethinking online hate speech in South Africa. *Alternation Journal*, *23*, 108–133. https://journals.ukzn.ac.za/index.php/soa/article/view/1240

Nowak, M. (2005). *UN Covenant on Civil and Political Rights: CCPR Commentary*. N.P. Engel Verlag.

Nyathi, N. M. (2018). *The poverty of law: A critical analysis of hate speech jurisprudence in South Africa* [Master's dissertation, University of Pretoria]. UP Repository.

Parekh. B. (2012). Is there a case for banning hate speech? In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: Rethinking regulations and responses* (pp. 37–56). Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.006

Rawls, J. (1993). *Political liberalism*. Columbia University Press.

Scheffler, A. (2015). *The inherent danger of hate speech legislation: A case study from Rwanda and Kenya on the failure of preventive measures*. Friedrich-Ebert-Stiftung. https://library.fes.de/pdf-files/bueros/africa-media/12462.pdf

Tijani, M. (2019, November 25). *Death penalty for criticising Nigeria's government? What we know about the hate speech bill.* AFP Fact Check. https://factcheck.afp.com/death-penalty-criticising-nigerias-government-what-we-know-about-hate-speech-bill

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*, 157–179. https://doi.org/10.1007/s11192-020-03737-6

United Nations. (1948a, December 9). Convention on the Prevention and Punishment of the Crime of Genocide. https://www.un.org/en/genocideprevention/documents/atrocity-crimes/Doc.1_Convention%20on%20the%20Prevention%20and%20Punishment%20of%20the%20Crime%20of%20Genocide.pdf

United Nations. (1948b, December 10). Universal Declaration of Human Rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights

United Nations. (2012, September 7). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on hate speech and incitement to hatred*. https://digitallibrary.un.org/record/735838

United Nations. (2013). Rabat Plan of Action. In *Annual report of the United Nations High Commissioner for Human Rights* (p. 6–15). https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf

United Nations. (2018, April 6). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on online content regulation.* https://digitallibrary.un.org/record/1631686

United Nations. (2019, October 9). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf

Viljoen, F. (2005). Hate speech in Rwanda as a test case for international human rights law. *Comparative and International Law Journal of Southern Africa*, (38)1, 1–14. https://hdl.handle.net/10520/AJA00104051_68

Workneh, T. W. (2020). Ethiopia's hate speech predicament: Seeking antidotes beyond a legislative response. *African Journalism Studies*, *40*(3), 123–139. https://doi.org/10.1080/23743670.2020.1729832

Digital
Communication
Research.de

Recommended citation: Ahmad, S. (2023). Who moderates my social media? Locating Indian workers in the global content moderation practices. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 111–125). Digital Communication Research. https://doi.org/10.48541/dcr.v12.7

Abstract: Building on the growing concerns around hate speech and harmful content on social media, this chapter analyzes the processes by which content is moderated on leading social media platforms. The outsourcing practices of platform operators or social media companies to acquire content moderation services from third-party companies have been acknowledged in the public discourse. Details regarding these outsourcing relationships and power mechanisms remain obfuscated, however. Using empirical data from India, this chapter presents a global value chain perspective on the mechanisms by which US-based social media monopolies source content moderation services from Indian information technology business process outsourcing (IT BPO) supplier companies. The agreements established between the two parties direct the content moderation labor process through which Indian workers' labor power is transformed into productive labor.

DOI 10.48541/dcr.v12.7

Sana Ahmad

Who Moderates My Social Media?

Locating Indian workers in the global content moderation practices

1 Fixing "our broken" social media

In a recent media article in *The Guardian*, technology reporter Julia C. Wong put together a list of proposals by North American researchers and activists to fix "our broken" social media (Wong, 2021). These proposals rather being speculative underscore concrete actions to regulate social media platforms. "We cannot fix what we do not understand," notes one of the experts, Alex Abdo, litigation director at the *Knight First Amendment Institute* (Wong, 2021). Abdo advocates for enabling independent inquiry by researchers and journalists to explain how social media companies have managed to prioritize user retention at the cost of allowing hate speech and fake news to circulate on their platforms.

The calls for social media companies to have transparent content moderation policies and practices on their platforms have gradually increased over the last several years. Several nation-state governments today, especially with evidence of the use of social media to influence their election results, are using legal routes to prohibit the presence of hate speech, fake news, and other propaganda on social media. Furthermore, countries such as the United Kingdom (UK), India, and

others are in the process of delegating the responsibility of monitoring and controlling these spaces to social media companies. However, putting the onus of social media management on companies and privatizing law enforcement can have repercussions for users' freedom of speech, as has been pointed out by many, including Brigitte Zypries, in her former role as Minister of Justice in Germany (Agence France-Presse, 2017).

The positions offered in this chapter do not argue for or against regulation of speech on social media platforms. Instead, I take Abdo's proposal seriously on understanding social media and platform operations before trying to fix it. The material conditions underlying the functioning of social media platforms and the built-in power asymmetries are the focal points of this chapter. Drawing heavily from the labor process debate, the commercialized practice of content moderation is examined here, with specific attention placed on the working conditions of content moderators employed at third-party contracting companies in India. Noting the contemporary public discussions and significance assigned to the function of content moderation for social media, this chapter aims to motivate the reader to consider the ongoing treatment of content moderation practices as *industrial secrets* by social media companies.

However, this chapter does not chart a distinct relationship between content moderation production processes and the proliferation of hate speech on social media. Increased public attention to harmful content on these platforms has elicited, on the one hand, techno-solution-oriented responses and, on the other, the assurance of contracting additional human reviewers by social media companies. While transparency reports from global social media monopolies show that the prevalence of hate speech has reduced on their platforms, it cannot be confirmed whether this has been made possible using content moderators' labor power or through the exclusive application of automated filters and technologies[1].

1 The fourth quarterly reports from 2020 can be accessed on the official websites of Google and Facebook. The 'YouTube Community Guidelines Enforcement' report is accessible at https://transparencyreport.google.com/youtube-policy/removals, and Facebook's 'Community Standards Enforcement Report' is accessible at https://transparency.facebook.com/community-standards-enforcement. The biannual 'Twitter Transparency Report' is accessible at https://transparency.twitter.com/en/reports.html.

The analysis presented here is derived from research fieldwork in India, which I undertook as part of my doctoral inquiry. In total, 35 guided interviews were conducted with target participants in India. The chapter consists of the following sections. I start by defining content moderation and the commercialization of this practice. Following this, I attend to the question of why it is obfuscated from the public view. I then present an overview of the content moderation labor process and possibilities for resistance, if any. Finally, I conclude by underlining the importance of further research and the relevance of policies in regulating these outsourced practices.

2 Content moderation: Why it matters

The practice of content moderation follows a pattern of evolution similar to that of the internet-based services. With an increase in the commercial application of the Internet in the 1990s, and an expansion of Internet-based services, the need to screen and monitor these services grew as well. Commercial services based on the World Wide Web, such as email services (Hotmail.com, Yahoo, AOL, etc.), classified advertisement services (Craigslist), dating services (Match.com), and peer-to-peer file sharing services, were monitored and controlled according to local regulations and company standards. Information scientists and inter-personal communication researchers were quick to identify the growth of social media as a "computer-mediated communication" in the form of emails, forums, and Bulletin Board Systems (Rice, 1980; Kerr & Hiltz, 1982 in Burgess et al., 2018). Many of these text-based social communities, followed by an increasing number of social technologies in the 1990s and mid-2000s (MySpace, Wikipedia, Reddit, etc.), placed emphasis on online community management through open and voluntary moderation (Roberts, 2017).

The shift of focus on social media from *social network sites* (boyd & Ellison, 2007) to *social media platforms* (Gillespie, 2018b) accompanied a surge across several disciplines, including media and communication studies, to examine the ethics of data culture, especially the collection, monitoring, and monetizing of user-generated data by social media companies (Herman, 2014; Helmond, 2015). It was Roberts (2019), however, who, through her empirical investigation, was able to link large-scale social media platforms in the United States of America (USA)

with the commercialized practice of content moderation. According to her definition, "commercial content moderation is the organized practice of screening user-generated content posted to Internet sites, social media and other online outlets, to determine the appropriateness of the content for a given site, locality, or jurisdiction" (2017, p. 1).

In a similar vein, Gillespie (2018a) identified synchronously occurring processes of content moderation on social media platforms and public exchange. Similar to Roberts (2019), Gillespie considered content moderation as the core process for maintaining these platforms, and he goes on to equate it as an "essential, constitutional and definitional" function of social media platforms (Gillespie, 2018a, p. 21). However, much before content moderation as a commercial practice could receive scholarly attention, investigative articles in the media exposed its outsourcing to peripheral states in the USA and later its offshoring to geographically dispersed locations across the world (Stone, 2010; Chen, 2012; Chaudhuri et al., 2014). India, along with the Philippines, has been observed as crucial locations for content moderation outsourcing.

3 A closer look at the hidden practices of content moderation

In his seminal work on providing a historical materialist understanding of *digital materialism*, Gottlieb examined the "mystification or metaphysical obfuscation" of processes associated with digital technologies (Gottlieb & Karatzogianni, 2018, p. 2). Gottlieb's focus on digital materiality allows us to acknowledge the often-times hidden labor that goes into producing and maintaining the technologies of today. Further, it prompts us to investigate the underlying social relations that constitute the technological processes. Examining the political economy of digital media certainly opens new opportunities for studying the unpaid activities of social media users and their commodification by social media companies (Fuchs, 2014, 2010; Dyer-Witheford, 2010). However, Gandini argued that such a broad analysis of labor and digital technology could risk understudying the hidden dimensions of *digital labor* (2020).

Gandini proposed considering platforms, including social media, as organizational actors and examining the "manifold ways in which the capital-labor relationship is enforced through them" (2020, p. 9). Correspondingly, the focus of this

essay is on explaining the production model of content moderation, which remains, as we have determined before, an essential feature of social media platforms. Much of this content moderation work, which involves screening large amounts of user-generated information within a very short amount of time, is carried out by outsourced workers located in the world peripheries.

Content moderation practices and the outsourcing of labor are hidden from the public view, thus making it difficult for independent researchers and journalists to assess these processes. Some of the self-described motivations of social media companies to maintain this secrecy are as follows: to protect the identities of workers (Gillespie, 2018b), to prevent the users who post illicit content on social media platforms to "game the rules" (Roberts, 2016, p. 7), and to "guard the proprietary tech property and gaining cover from liability" (Buni & Chemaly, 2016, p. 12).

In the wake of leaks in media articles as well as lawsuits filed by content moderators against social media companies, the industry's secrets are spilling out. Yet, public and legal focus on the hidden labor of content moderation has remained rather limited. As mentioned above, national legislation in different countries is taking shape in trying to shift the liability on social media companies for hosting illegal online content, disputing the protection of these companies under the *safe-harbor* legislation in the USA.[2] While these developments have certainly allowed us to challenge what some have called a "marketplace orientation" of Section 230 (Medeiros, 2017, p. 2), they have yet to take into concern the production process of content moderation and labor, which goes into sustaining this essential practice.[3]

The outsourcing of content moderation work by social media companies has created global content moderation value chains. While the content moderation policies and software are designed within social media companies, the actual labor of content moderation, which is often low-paid and "rote, repetitive, quota-driven, queue based" (Roberts, 2019, p. 92), is outsourced to contracted content moderators who are placed at great distances from these companies. In public discourse, content moderation has often been understood as an automated task, and the reality of human content moderation has only been explored in the

2 Section 230 of the Communications Decency Act in the United States provides the Silicon Valley-based social media giants, along with other websites, a safe harbor from liability for user-generated content or third-party content posted on their platforms.

3 Medeiros notes that for these companies, "suppression of speech can be anathema to the marketplace theory" (2017, p. 2).

recent years. Studying the offshore practices of content moderation on social media platforms is challenging. Most notably, the term "content moderation" is not a standard business terminology. Instead, several other job titles, such as "system analyst," "website administrator," "process executive," and others, are assigned to moderators by supplier companies in India (Ahmad & Krzywdzinski, 2022).[4] Roberts (2019, p. 40) noted that these "multitudinous" job titles function to further conceal the content moderation process.

The deliberate concealment of this process by target social media companies located in the Global North and complying supplier companies in India compels me to argue that the rules governing outsourcing relationships and the resulting labor processes of content moderators in India are designed to create opacity around content moderation practices. As we will see in the following section, social media companies outsource content moderation to India (our target location) through traditional business process outsourcing practices in which gig work online platforms do not play a major role. Most content moderation labor processes are organized and controlled by social media companies, their standards, and software infrastructures.

4 Exploring the labor process of content moderation

Over one-tenth of moderation workers worldwide are located in India, which is one of the main destinations of content moderation outsourcing.[5] In my research, I identify content moderation as a back-end, non-voice business process that is supplied as a service to their clients, including to social media companies, by information technology business process outsourcing (IT BPO) sector companies in India.

4 The suppliers referred to here are information technology business process outsourcing (IT BPO) companies who provide a range of services and technological solution to their clients located around the globe. A motley assortment of clients requires content moderation services for their social media platforms, e-commerce websites or simply their user-content hosting websites. The Indian IT BPO companies supply content moderation services to these different clients.

5 The estimation was made by Himanshu Nigam, former chief security officer at the social media platform MySpace and a former security executive at Microsoft (Chaudhuri et al., 2014). There are no publicly available statistics which could show the impact of content moderation services on the Indian labour market.

The analysis presented in this chapter is based on research fieldwork in India, and grounded theory methodology guided both the data collection and data analysis processes. The research was undertaken between January 2019 and April 2019, and constituted nine interviews with content moderators and three interviews with content operators[6], six interviews with management at the supplier companies in India, two interviews with domestic social media companies, seven interviews with trade unions, and eight interviews with civil society organizations. Moreover, informal meetings were held with experts in the fields of labor law, technology, and free speech to achieve more insights into this service work.

Gaining access to the participants was extremely challenging and explained the absence of representation from international social media companies. Most of the workers were approached on an international networking website for professionals, and the rest were contacted using the snowball sampling technique. Considering the sensitivity of the subjects, great care was taken in protecting the identities of all participants, both during and after the data collection process.

With content moderation practices treated as industrial secrets, as has been described before, the brief description of content moderation outsourcing mechanisms presented here is influenced by the vast body of literature on Indian call center companies (within the IT BPO sector) and the work organization and management strategies of these companies. The discussion on labor processes in call center companies highlights the subordinate position of Indian companies in global value chains and the subsequent vulnerability of the workforce (Batt et al., 2005).

Global content moderation value chains are facilitated by service level agreements (SLAs), which are established in this case between social media companies in the Global North and content moderation suppliers in India. Depending on the terms of the agreement, specific tasks are allocated to social media companies and their suppliers. Training, developing content moderation policies, and other product-oriented tasks are managed by social media companies.[7] By contrast, tasks such as managing wages, leave of absence, workplace

6 *Content operator* is an official designation at domestic and regional social media firms, wherein the workers are assigned content-related tasks, such as user acquisition and retention, along with either moderating the content themselves or overseeing the moderation tasks done by external freelance moderators.

7 Product-oriented task refers to social media platform as a product which is designed by its proprietor, the respective social media company.

conflict, and other human-resource related tasks are handled by the supplier companies. These agreements are mostly project based and are determined by measurable standards, such as quantity targets (the amount of user content moderated) and time. Such factors enable flexibility and scaling-up opportunities (regarding the volume of their outsourced content moderation business) for social media companies.

As observed, these content moderation value chains are characterized by high power asymmetries between social media companies on the one hand and supplier companies on the other.[8] These, I argue, have an influence on the content moderation labor process. The different kinds of content moderation value chains and the types of governance of these value chains will not be elaborated on here, thereby allowing readers to focus on the particularities of the labor process. Using the data collected from the research fieldwork, three main aspects of the content moderation labor process are highlighted: the recruitment process, the organization of work, and the conditions of work. These and other aspects of the moderation labor process have been expanded in further detail by Ahmad and Krzywdzinski (2022).

In terms of recruitment, the suppliers undertake most of the processes according to the SLAs, which specify the project details, including the number of workers to be hired by the supplier company. Depending on the agreements established between the two parties, some social media companies could directly participate in the recruitment process. The skills required for this work are mostly generic and allow applications from a diverse range of backgrounds, such as engineering and technology, media, and communications, management studies, and others. Opacity around content moderation production already starts from the recruitment process, where the moderators are required to sign non-disclosure agreements, thereby disallowing them from disclosing any details about the client and work process to a third party. Many of those who are selected and have agreed to exchange their labor for low wages and few benefits are "freshers," whose first job is content moderation. Overlooking the lack of work information provided to them, the moderators noted that they were

8 The analysis on the outsourcing relationships presented here and the resulting power asymmetries, is informed by an extensive literature on global value chains, most notably by Gereffi et al. (2005) and Ponte and Sturgeon (2014).

attracted to the possibility of working for *global brands* (popular social media companies) and saw it as their entry job into the IT sector.

The aspect of work organization can be explained according to the different types of content moderation. *Proactive moderation* before the content is published on the platform and *reactive moderation* after the content is published on the platforms are the two categories provided by Grimmelmann (2015) to explain the segmentation of the global content moderation market. A crucial point to note across both of these moderation types is the deployment of technical resources by social media companies. Extreme content, such as child sexual abuse and non-consensual porn, which impedes the public image of the company above the broad threshold, requires automatic detection before or very quickly after it is published on the social media platform. Interviews with both the content moderators and the representatives from the supplier companies revealed that such content does not enter the manual queues. Through the last years, many big social media companies have invested in or acquired the use of automated technologies to proactively moderate content. However, noting the large scale of content generated by users on their platforms, proactive moderation can be difficult. Reactive moderation depends on the users or third parties flagging or reporting content on the platform, and the content is sent to both automated and manual moderation processes.

Depending on the requirements of the social media companies, suppliers invest in basic filters or advanced technology, which constitutes the first part of the moderation process. Thereafter, content that has not been moderated by automated technology enters the queues of the moderators. These content queues can be identified as hate speech, spam, and others that are assigned to the moderators on mostly an arbitrary basis, following a mandatory training period. Depending on the terms of the SLAs, moderators review the user-generated content and make prescribed decisions according to the policies of the respective social media platforms.

The decision-making capabilities of the moderators vary from one moderation value chain to the other, where, on the one hand, the moderators are allowed to delete the content and even ban the user, and on the other, the moderators are allowed to simply tag the flagged content with the respective policies. Again, depending on the arrangement made between the social media company and the supplier, there exist other teams of quality analysts and team leaders that constitute fewer members and are higher up in the process hierarchy. Their work

comprises controlling the performance of moderators and may even include the task of making final decisions on the already tagged content by the lower-level of moderators. The work of content moderators is organized through moderation software and assistive technologies, which are either developed in-house by the social media company or have to adhere to stringent standards.

The organization of work has a multidimensional impact on the working conditions of the moderators. First, the content moderation work process is highly controlled, with specific targets assigned to each moderator every month, depending on their content queues, content format (videos, text, images, etc.), and team size. If they are unable to complete their targets on time, the management at the supplier company penalizes the moderators using gradually-depraving disciplinary measures. In the beginning, they are issued statutory warnings, following which they are shifted to elementary levels of content moderation work, or a simpler project. Granting all these steps, if the moderators are still unable to improve their performance, they are eventually expelled from the supplier company and are required to serve their notice period.[9] This creates a lot of psychological stress for the moderators and intersects with other reasons for resentment against the management, including low wages, long working hours, work-shifts[10], and lack of skill development.

The second trying element of this work is the distinct characteristic of the user content on which moderators have to review, tag, and or make decisions on. Content involving violence, assault, animal abuse, and other distressing material is visible to the content moderators, although the frequency of its visibility depends on their content queues. This means that queues with content on hate speech, violence, and nudity, etc., have a higher prevalence of distressing content than other queues, especially in the electronic commerce (e-commerce) section (such as Facebook Marketplace etc.). Regardless of the rate of occurrence

9 The notice period usually spans between one to three months and allows the moderator to apply for another project in the respective supplier firm. While their employment contract is still valid during the notice period, they are not paid their usual wages. Depending on the policies of the supplier company, the management might only support the health insurance of the moderator and even their families, which amounts to a small sum.

10 Content moderation service constitutes a 24-hour work cycle with three to four shifts running throughout day and night.

of distressing content, conversations with moderators during this research revealed that watching harmful content can have a lasting impact on the mental health of the respective moderators.

Considering the deplorable working conditions presented here, the reader might expect the emergence of collective resistance by the content moderators, especially since the Indian IT sector provides us with increasing examples of unionizing activities.[11] Instead of engaging in explicit forms of resistance, many of those who participated in this study exercised resilience and were of the view that they had to adapt to watching distressing content if they wanted to continue working in the content moderation process. Further, some echoed the opinion (by the management at the supplier companies) that their work was necessary to "guard the world against harmful content on social media." In terms of negotiations for wages and skill development, moderators approached the management individually, hoping to succeed on the basis of their personal relationships. However, the supplier management was often dismissive of these demands. Correspondingly, the social media companies played no role in managing conflicts between the moderator and the supplier company. Lacking possibilities for better career opportunities at the supplier company and the non-likelihood of employment at the respective social media companies (which they had initially aspired for), moderators design their own "career staircases" (James & Vira, 2012, p. 3; Ahmad & Krzywdzinski, 2022, p. 90) across the expanding labor market for content moderation in India.[12]

11 Indian IT trade unions, such as Union for IT-enabled Services (UNITES) professionals and Forum for IT Employees (FITE) have been formed in the last few years, mostly as a response to rising layoffs in the sector. Further, the IT and IT-enabled services sector is increasingly becoming a focus of interest for many central trade unions in the country.

12 There is a growing content moderation market in India with domestic and regional social media companies, including Chinese companies, sourcing content moderation services from the Indian suppliers.

5 Essential to social media but invisible to the world: Turning the spotlight on content moderation labor

Against the background of increasing public pressure to regulate social media platforms, this chapter presses for additional attention to the production processes of content moderation. This includes identifying the outsourcing practices that social media companies design to obtain content moderation services for their platforms. To this end, this chapter focuses on the labor process and the resulting working conditions of the moderators. The focus of existing scholarship on the Global North is expanded here to include India, where a growing number of content moderators are located. Much of this narrow focus can be attributed to the hidden outsourcing practices of content moderation, which veil the high power asymmetries between social media companies based in the Global North and content moderation supplier companies located in India.

The relationship between the two stakeholders has important consequences for the content moderation labor process. Social media companies outsource content moderation work to suppliers in India on a project basis and set standards for moderation policies, technology, and other product-related tasks. The companies in India are mostly tasked with employing the content moderators, controlling their performance and managing their wages, skill development, and other human-related aspects. The resulting labor processes have been described in this chapter under three main parameters: the recruitment process, organization of work, and working conditions. Lack of explicit forms of collective resistance by the content moderators is accompanied by their resilience and individual strategies for change.

This essay does not seek to provide an overarching picture of the outsourced content moderation practice to India. For starters, there is no single practice of content moderation outsourcing that can be delineated here. Instead, there are different content moderation value chains taking shape through agreements formed between social media companies and supplier companies in India. The governance of these value chains differs, leading to different levels of coordination mechanisms and power asymmetries. We can however note that most of the standards of content moderation are set by social media companies, thereby leaving the suppliers at less powerful positions and the moderators with even lower control over their labor. Further research is required to determine if there are more stakeholders involved in these chains, the mobility of workers across these

chains, and potential new forms of resistance. Additionally, domestic and international public policies must be aimed at improving the working conditions of moderators who supply commonly used social media platforms with essential labor.

Sana Ahmad is a doctoral candidate at the Freie Universität Berlin and an associate researcher at the Weizenbaum Institute for the Networked Society in Berlin, Germany.

References

Agence France-Presse (2017). Social media sites face heavy hate speech fines under German proposal. *The Guardian*. https://theguardian.com/media/2017/mar/14/social-media-hate-speech-fines-germany-heiko-maas-facebook

Ahmad, S., & Krzywdzinski, M. (2022). Moderating in obscurity: How Indian content moderators work in global content moderation value chains. In M. Graham & F. Ferrari (Eds.), *Digital work in the planetary market* (pp. 77–95). International Development Research Centre and MIT Press.

Batt, R., Doellgast, V., Kwon, H., Nopany, M., Nopany, P., & da Costa, A. (2005). The Indian call centre industry: National benchmarking report strategy, HR practices, & performance. *CAHRS Working Paper Series*. https://hdl.handle.net/1813/77401

boyd, d. m., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Buni, C., & Chemaly, S. (2016). The secret rules of the internet. *The Verge*. https://theverge.com/2016/4/13/11387934/internet-moderatorhistory-youtube-facebook-reddit-censorship-free-speech

Burgess, J., Marwick, A., & Poell, T. (Eds.) (2018). *The Sage handbook of social media.* SAGE. https://doi.org/10.4135/9781473984066

Chen, A. (2012). Inside Facebook's outsourced anti-porn and Gore brigade, where 'camel toes' are more offensive than 'crushed heads. *Gawker*. https://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads

Chaudhuri, P., Chatterjee, A., & Verma, V. (2014). Guardians of the internet. *The Telegraph India*. https://www.telegraphindia.com/7-days/guardians-of-the-internet/cid/1669422

Dyer-Witheford, N. (2010). Digital labor, species-becoming and the global worker. *Ephemera: Theory and Politics in Organization*, *10*(3), 484–503.

Fuchs, C. (2010). Labor in informational capitalism and on the internet. *The Information Society*, *26*(3), 179–196. https://doi.org/10.1080/01972241003712215

Fuchs, C. (2014). *Digital Labour and Karl Marx*. Routledge.

Gandini, A. (2020). Digital labor: An empty signifier? *Media, Culture & Society, 43*(2), 369–380. https://doi.org/10.1177/0163443720948018

Gereffi, G., Humphrey, J., & Sturgeon. T. (2005). The governance of global value chains. *Review of International Political Economy*, *12*(1), 78–104. https://doi.org/10.1080/09692290500049805

Gillespie, T. (2018a). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2018b). Regulation of and by platforms. In J. Burgess, A. Marwick, & T. Poell, (Eds.), *The SAGE handbook of social media* (pp. 254–278). SAGE.

Gottlieb, B., & Karatzogianni, A. (2018). *Digital materialism: Origins, philosophies, prospects*. Emerald Group Publishing.

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology*, *17*(1), 42–109.

Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media + Society*, *1*(2). https://doi.org/10.1177/2056305115603080

Herman, A. (2014). Production, consumption, and labour in the social media mode of communication and production. In J. Hunsinger & T. Senft (Eds.), *The social media handbook* (pp. 30–44). Routledge.

James, A., & Vira, B. (2012). Labour geographies of India's new service economy. *Journal of Economic Geography*, *12*(4), 841–875. https://doi.org/10.1093/jeg/lbs008

Medeiros, B. (2017). Platform (non-)intervention and the "marketplace" paradigm for speech regulation. *Social Media + Society*, *3*(1), 1–10. https://doi.org/10.1177/2056305117691997

Ponte, S., & Sturgeon. T. (2014). Explaining governance in global value chains: A modular theory-building effort. *Review of International Political Economy*, *21*(1), 195–223. https://doi.org/10.1080/09692290.2013.809596

Roberts, S. T. (2016). Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. *Wi: Journal of Mobile Media*, *10*(1), 1–18.

Roberts, S. T. (2017). *Content moderation*. In L. A. Schintler & C. L. McNeely, (Eds.) Encyclopedia of big data (C: 1–4). Springer.

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Stone, B. (2010). Policing the web's lurid precincts. *The New York Times*. https://www.nytimes.com/2010/07/19/technology/19screen.html

Wong, J. C. (2021). Banning Trump won't fix social media: 10 ideas to rebuild our broken internet – by experts. *The Guardian*. https://theguardian.com/media/2021/jan/16/how-to-fix-social-media-trump-ban-free-speech

Digital
Communication
Research.de

Recommended citation: Schemer, C., & Reiners, L. (2023). Challenges of comparative research on hate speech in media user comments: Comparing countries, platforms, and target groups. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 127–139). Digital Communication Research. https://doi.org/10.48541/dcr.v12.8

Abstract: Hate speech is a phenomenon studied in numerous disciplines by many researchers. This research has produced a variety of findings, e.g., with regard to the prevalence of hate, common targets or differences between platforms or countries. However, previous research also comes with conceptual and methodological challenges, e.g., definitions or operationalizations of hate speech in empirical studies. The present chapter focuses on the issue of equivalence in previous hate speech research—a well-known problem of comparative research in general. To compare research findings relating to hate speech across different contexts scholars need to consider the equivalence with respect to definitions, methods, measurements, procedures, and also the sampling communication content. We provide an overview about potential pitfalls and biases that can be due to a lack of equivalence and point to strategies on how to address them.

DOI 10.48541/dcr.v12.8

Christian Schemer & Liane Reiners

Challenges of Comparative Research on Hate Speech in Media User Comments

Comparing countries, platforms, and target groups

1 Introduction

A vast body of research on hate speech in user comments is dispersed across disciplines, such as communication, political science, computer linguistics, and linguistics. From a comparative perspective, one major challenge exists: It is difficult to compare and make sense of results from different studies because they differ in terms of their definitions, sampling strategies and units, and measurements of hate speech (for a recent overview of comparative studies, see Pamungkas et al., 2021 and also Fortuna et al. in this volume). Therefore, it is often difficult to argue that hate speech prevalence is higher in one country compared to another one. This problem also arises when researchers compare platforms, the comment sections of different news outlets, and so on. A downstream consequence of biased estimates of hate speech is also that the prediction of hate speech across different contexts cannot be compared. For comparative researchers, this is a well-known problem.

Basically, a comparison or summary of results across contexts requires assumptions that relate to the equivalence of definitions, methods, and procedures

that are used in empirical research (for a more in-depth look at comparative research methodology, see Rössler, 2012; van de Vijver & Leung, 1997; Wirth & Kolb, 2012, 2014). This also holds true for single studies that annotate hate speech in user comments in different contexts, including platforms (Olteanu et al., 2018), media outlets (Paasch-Colberg et al., 2021; Zannettou et al., 2020), countries (Hanzelka & Schmidt, 2021; Ruiz et al., 2011), and targets or authors of hate speech (ElSherief et al., 2018). The problems related to the analysis of contexts, such as platforms or media outlets that host user comments, are often not easier to solve than those linked to cross-cultural analysis.

Most studies on hate speech are not explicitly comparative in nature, but may nevertheless be plagued by equivalence issues. This chapter aims to raise awareness among researchers of these methodological issues to encourage research that can be used for comparative purposes. To this end, this chapter emphasizes the role of equivalence at different levels and responds to some equivalence issues that occur in the first part of this edited volume. It demonstrates what the equivalence of definitions of key concepts, sampling, and measurements means and how violations of equivalence can bias the comparison of findings across contexts.

2 Equivalence of definitions, measurements, and procedures

Comparisons across contexts, such as actors, platforms, or cultures, require that a construct of interest, such as hate speech, can be considered as a single unitary construct that is manifest (i.e., located and observable) in user comments. If we start from an etic position and the existence of a universal phenomenon called "hate speech," which we can describe as a theoretical concept, the crucial question relates to whether this phenomenon can be assessed with measures that are specific to a context or not (Triandis & Marin, 1983). If we assume that manifestations of hate speech differ across contexts (e.g., users rely on different ethnic slurs for social groups in different contexts), an emic measurement strategy is required. Research that aims at comparisons of hate speech across such contexts would need to argue that different ethnic slurs of social groups are functional equivalents (for a detailed discussion of these issues, see Wirth & Kolb, 2012). Without this assumption we cannot know whether a particular group is more often the target of hate

speech, whether a particular event elicited more hate than another, or whether hate speech is more prevalent in some countries than others.

If we consider previous definitions of hate speech (see, for an overview, Paasch-Colberg et al., 2021; Reiners & Schemer, 2020; Siegel, 2020), it becomes clear that researchers frequently start with different conceptions of the construct of interest. This is sometimes guided by pragmatic considerations (e.g., the processing of large quantities of user-generated content). Additionally, ideographic aspects of an event or a culture motivate how researchers approach hate speech (e.g., user-generated content after Islamist terror attacks). The issue of the *equivalence of definitions* is complicated by the use of different labels when talking about hate speech. This can vary from abusive language to verbal aggression, toxic or dangerous speech, extremism, and many more (e.g., Schmidt & Wiegand, 2017; Siegel, 2020; see also the "Theoretical Perspectives" section in this volume).

Some researchers also include an effects dimension of hate speech (i.e., speech that incites hate or violence; see, e.g., Gagliardone et al., 2015). This complicates the assessment of hate speech even further because research then has to specify not only the content that is typical of hate speech but also the effects on users that may be difficult to observe. Thus, if definitions of central theoretical concepts differ across studies, then comparisons across these studies or contexts become difficult to interpret at best and meaningless at worst (Rössler, 2012). Therefore, a basic requirement of comparisons across contexts is that at least the functional equivalence of measures of hate speech exists.

Narrow theoretical conceptions of hate speech can simplify the task of achieving the *equivalence of measurements*. However, they are likely to underestimate the amount of hate that circulates on social media. Broad definitions are likely to result in overestimation. For instance, Silva et al. (2016) assume that hate speech is an expression of a user that describes a negative stance toward a social group (e.g., "I hate [or don't like or other expressions by users] some member of a social group"). This is a narrow conception of hate speech because other expressions, such as explicitly assigning negative attributes to social groups or using ethnic slurs, can frequently occur (Siegel, 2020). This definition also ignores subtle forms of hate speech (Schmidt & Wiegand, 2017). Implicit notions, such as humor or the use of specific metaphors as hate speech devices, are real challenges for equivalence (see Szczepańska & Marchlewska in this volume). Specifically, the authors demonstrate the diversity of slogans that a Polish protest movement uses against

the government, ranging from the outright derogation of the ruling party to subtle and humorous appeals, which are less explicit and negative but are meant to ridicule the governing elite.

Therefore, the amount of hate speech that researchers relying on a narrow definition of the same can find is likely an underestimation (i.e., 20,305 tweets out of 512 million, which is around 0.004 per cent; Silva et al., 2016). Burnap and Williams (2015) started with a broader definition of hate speech as offensive or antagonistic in terms of race, ethnicity, or religion. They found a prevalence of 11 per cent of hateful tweets. The broader definition of hate speech is likely to result in a higher prevalence estimate. In this study, hateful comments may include expressions that other authors would not consider hateful, but rather criticism or disagreement. Another study defines "hateful speech as discourse practiced by communities who self-identify as hateful towards a target group" (Saleem et al., 2016, p. 4). This means that every post in such a community is automatically classified as hate speech (for a similar approach, see Albuquerque & Alves in this volume). In this study, the authors focus on a pro-Bolsonaro network on Brazilian social media, which is labeled the "Office of Hate" and is considered a spreader of hate speech against social groups and established institutions. Although these studies on the structure of notorious hate nets offer important insights, ignoring heterogeneity in their communication is a limitation. Saleem et al. (2016) also acknowledge that some of this communication may be "non-hateful chatter." Thus, not all communication in the "Office of Hate" should be automatically categorized as hate speech if parts of the conversation there do not attack or derogate individuals or social groups.

This discussion on the heterogeneity of definitions and a quick look at operationalizations of hate speech in previous studies demonstrate that extant research is far from achieving functional equivalence, let alone the strict invariance of measures for the detection of hate speech. However, having unequivocal definitions of hate speech would produce truly valuable findings. For instance, research could provide evidence of which platforms, outlets, or sites are more likely to be plagued by hate speech. This can be helpful for practitioners and political authorities to tailor interventions or policies that aim to reduce hate speech. Research would also benefit from unequivocal and comparable definitions. So far, most research is concerned with the detection of hate speech and less so with the prediction of the same. If researchers can agree on definitions

of hate speech, predictive studies also become comparable, and we would learn more about the causes of hate speech at the levels of the technical infrastructure, the authors, and the specific situations and contexts within which this communication emerges.

The equivalence of definitions and (functionally) equivalent measures to assess hate speech in different contexts are a necessary condition of comparisons but not a sufficient one. *Procedural equivalence* is another issue that researchers need to be aware of. For instance, this refers to potential differences in how annotators apply a coding scheme to a given corpus. Ross et al. (2016) demonstrate that even providing detailed guidance for annotators can result in the low reliability of hate speech annotations. If the application of annotation guidelines varies across annotators or cultures, then comparisons across these contexts can be severely biased. There are also practical issues in multicultural studies that can emerge from common language guidelines and the use of translations for annotations in a given language (see Rössler, 2012 for a discussion of such procedures in content analysis). When it comes to translations, researchers need to be aware of instrument bias, which means that translations of measures and guidelines result in different interpretations by annotators or different applications of the instrument for a given corpus. Consequently, the assumption of (functional) equivalence is violated, and comparisons across these contexts are also biased. There are also means to quantify whether measurement invariance truly holds by comparing the reliability of annotations or accounting for differences in reliability when analyzing comparative data. However, in cross-cultural content analytic work, this is more complicated than in survey research (for an overview of this problem, see Rössler, 2012; Wirth & Kolb, 2012).

3 Sampling equivalence

Hate speech is frequently a moving target, and sampling strategies need to account for these dynamics. The comparison of studies is frequently hampered by differences in sampling. Similarly, single studies that compare user-generated content across outlets or platforms encountering issues, such as different publication and registration policies, moderation frequency and style, and many more, can threaten sampling equivalence (e.g., Ruiz et al., 2011). There are at least two

sources of bias that can occur and challenge comparability: first, bias that is due to the researchers' motivation and focus, and second, bias that is due to platform hosts or providers or community managers in comment sections.

Sampling bias due to the focus of a researcher refers to the sampling of user comments that are specifically tied to an event; a specific group; keywords, such as hashtags; or a specific time frame (e.g., Burnap & Williams, 2015; Chaudhry, 2015; see also Harb and Szczepańska & Marchlewska in this volume). For instance, Szczepańska and Marchlewska (this volume) study hate speech in the context of the "All-Poland Women's Strike" against the ruling government. It is unclear how the amount and quality of hate speech found in this context compare to other protests or other targets of hate within Poland. In a similar vein, Harb (this volume) focuses on hate speech by Lebanese journalists targeting the Shia community, among others. However, it is difficult to know how this compares to hate speech by other actors (e.g., ordinary users) or how the findings compare to less exceptional situations.

Prevalence estimates of hate speech based on these selected samples cannot be compared to each other nor to representative samples from platforms, websites, or comments sections without any further assumptions about the data generation process. Siegel et al. (2021) compared a representative sample of random tweets to samples related to Trump and Clinton from the election campaign and found considerable differences between daily occurrences of hate speech that were difficult to predict. Thus, research findings based on samples generated in the context of specific events or related to specific keywords or hashtags cannot be generalized to other contexts or routine communication situations. Other studies demonstrate that moderators and platforms behave differently in times of crises than in routine periods (Mladenović et al., 2020). These differences in moderation behavior are another issue that threatens the generalization of findings based on event-specific samples.

Bias due to providers or hosts of user comments result from different policies of countries, providers, platforms, or outlets that affect the deletion rate of hateful comments. Specifically, some platform providers filter hateful comments before they get published and before researchers can capture them. These policies may be platform-specific or result from legislation that is specific to a country (e.g., the liability of Holocaust denial in different countries; Kennedy et al., 2018). In addition to such interventions, comment moderators or lay community

managers can actively intervene in discussions. The potential interventions by all these actors are likely to reduce the amount of hate that researchers can obtain from comment sections on news websites or networking sites.

However, it is important to consider how actors from different platforms or news sites differ in terms of their intervention strategies. For instance, Facebook, YouTube, and Twitter differ in their policies with regard to dealing with hateful content (for an overview, see Fortuna & Nunes, 2018; Siegel, 2020). To complicate matters even more, the same platform can even differ in its treatment of hate speech attacking specific targets. On Facebook, hateful comments addressing protected groups, such as Muslims, violate community policies, while migrants do not qualify as a protected group (Fortuna & Nunes, 2018). Thus, researchers would consider "Fucking Muslims" and "Fucking Migrants" as instances of hate speech. However, given that Facebook automatically deletes the former, comparisons across such groups will return biased results.

The use of the same platforms for sampling user comments across different countries does not guarantee equivalence either. Comparisons can be biased by different legislations and the populations that use these platforms. For instance, Twitter is more widely used by the populations of the United States or the United Kingdom, but less so in Germany. In this case, not only the populations differ but maybe also the functions of such a service. Algorithmic treatment of user comments may also differ across countries when algorithms are tuned for a specific language but perform poorly in others. Any difference in the prevalence of hate speech on such platforms between countries can be due to different populations using the platform, different intervention policies, algorithms working differently, or true cultural differences. However, it is impossible to disentangle these sources of variance in observational data.

One option for avoiding this problem involves studying comment sections without any moderation or intervention. However, this is difficult to know beforehand despite some platforms having few restrictions (Strippel & Paasch-Colberg, 2020). Another option is to account for differences in moderation practices by observing moderation or checking for differences in moderation policies. However, Ahmad (in this volume) suggests that moderation practices can vary across moderators and within moderators over time. Similarly, researchers can take into account differences in populations that communicate on specific platforms. This informed approach can result in weighting procedures to reduce sampling bias.

If specific sampling strategies are chosen, it is important to discuss the findings against this background (Rössler, 2012). Otherwise, findings from comparative studies are difficult to interpret.

For instance, Ruiz et al. (2011) compared user comments on newspaper websites in five countries. They sampled posts from one single quality newspaper in each country, most of which had a liberal leaning. Obviously, a single outlet with a specific political leaning cannot represent a whole media system or culture. Nevertheless, the authors present their results as if this was the case and as if the cultural context can explain the findings. Specifically, Ruiz et al. (2011, p. 482) state that the "results of this study suggest that the cultural context is relevant to the democratic quality of the debates we analyzed." So, if research only looks at variation across countries without any variation across outlets within a country, inferences with respect to cultural differences are always confounded by differences across outlets. Other research that examined single cases across countries produced similarly problematic findings that are difficult to interpret (e.g., the comparison of the anti-Islam Facebook group Pegida in Germany and initiatives against Islam in the Czech Republic by Hanzelka & Schmidt, 2017). However, avoiding these pitfalls is important to secure sampling equivalence and to draw valid inferences with respect to differences across countries, platforms, sites, or targets of hate speech. At the very least, a thorough discussion on how the sampling strategies may have affected the given findings should be included in any research report (Rössler, 2012).

4 Equivalence of context

Securing equivalence is a prerequisite for comparisons. However, researchers also need to be aware of the broader context in which hate speech occurs. This context can be essential for understanding and interpreting research findings. From the perspective of public discourse in liberal democracy, where the freedom of expression is not an issue, hate speech is easily condemned when it is observed since it can be harmful to substantive debates. However, hate speech or elements of hate speech can also occur in other contexts. There are subcultures and minority groups, for example, that use offensive and sometimes hateful language in a positive sense to build and preserve a common ingroup identity without

devaluing their own or other social groups (see Davidson et al., 2017). The use of the n-word among the people of colored communities is one prominent example. On the other hand, incivility and hate speech are frequently an option for expressing one's opposition to corrupt or authoritarian regimes when offline opposition is impossible or dangerous. In this vein, hate speech is considered as a means of self-defense against oppressive actors (see Szczepańska & Marchlewska in this volume). For instance, according to Szczepańska and Marchlewska, protesters in the "All-Poland Women's Strike" relied on hate speech as a last resort to fight against the abortion policies of the ruling government. In the present chapter, we cannot discuss the legitimacy of hate speech as self-defense. However, it is important to distinguish hate speech that comes from oppressed minorities or from actors that aim at silencing oppositional forces.

Therefore, it is important to consider the sociopolitical context in which hateful speech is embedded (see Litvinenko in this volume). This context also matters for the normative evaluation of hate speech and policies designed to avoid, reduce, or moderate it. These differences in functions of hate speech within and across societies considerably complicate the regulation of the phenomenon at the national and global levels (see Litvinenko and Ilori in this volume). For instance, harsher restrictions to regulate hate speech on social network sites in Western democracies have inspired authoritarian regimes to copy more restrictive policies, but with the goal of banning or censoring any oppositional voices. Thus, as Ilori (this volume) points out, any regulation, be it legal or non-legal (e.g., by exerting social pressure on haters in social networks), needs to balance the civility of political discourse against the freedom of speech.

5 Agenda for future comparative research

Research on hate speech has increased in the past decade and has improved considerably with respect to the methods that are used and breadth of phenomena and outlets that are studied. Making sense of all these studies requires comparing the findings from different studies or the results across contexts within single studies. Otherwise, we end up with idiosyncratic explanations for the emergence and dynamics of hate speech. The present chapter demonstrates how a basic requirement for comparisons is equivalence with respect to definitions,

methods, measurements and procedures, and sampling. Equivalence with respect to context matters for the substantive interpretation of comparisons.

Research reviews in the field raise awareness of some of these issues by discussing problems of narrow versus broad definitions of hate speech (Schmidt & Wiegand, 2017; Siegel, 2020), issues of reliability (Ross et al., 2016), or the generalization of classification algorithms (Fortuna et al., 2021). However, most primary research rarely accounts for the problem that violations of equivalence assumptions invalidate comparisons across studies or across contexts within a given study. Therefore, future research needs to take issues of equivalence and potential bias more seriously. Specifically, reasoning about equivalence should inform the design of a study, the sampling and collection of data, the measurements of hate speech, and, finally, the analysis of data. Ideally, equivalence should be quantified and used in weighting procedures in the analysis of data to account for potential bias. At the very least, researchers need to show awareness of bias due to violations of equivalence and discuss their findings against this backdrop.

Christian Schemer is Professor for Communication at the Department of Communication at Johannes Gutenberg-Universität in Mainz, Germany. https://orcid.org/0000-0002-7808-2240

Liane Reiners is a research assistant at the Department of Communication at Johannes Gutenberg-Universität in Mainz, Germany.

References

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Chaudhry, I. (2015). #Hashtagging hate: Using Twitter to track racism online. *First Monday*, *20*(2). https://doi.org/10.5210/fm.v20i2.5450

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In O. Varol, E. Ferrara, C. A. Davis, F. Menczer, & A. Flammini (Eds.), *Proceedings of the 11th International AAAI Conference on Web and Social Media – ICWSM 2017* (pp. 512–515). AAAI. https://arxiv.org/pdf/1703.04009.pdf

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). *Hate lingo: A target-based linguistic analysis of hate speech in social media.* https://arxiv.org/abs/1804.04257

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), article 85, 1–30. https://doi.org/10.1145/3232676

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, *58*(3), 102524. https://doi.org/10.1016/j.ipm.2021.102524

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. UNESCO Publishing.

Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-Muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*, *11*(1), 143–160. https://doi.org/10.5281/zenodo.495778

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Koombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Olmos, G., Omary, A., Park, C., Wang, C., Wang, X., Zhang, Y., & Dehghani, M. (2018). *Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale*. https://psyarxiv.com/hqjxn/

Mladenović, M., Ošmjanski, V., & Stanković, S. V. (2020). Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Computing Surveys*, *54*(1), article 1. https://doi.org/10.1145/3424246

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). *The effect of extremist violence on hateful speech online*. https://arxiv.org/abs/1804.05704

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180. https://doi.org/10.17645/mac.v9i1.3399

Panmungkas, E. W., Basile, V., & Patti, V. (2021). Towards multidomain and multilingual abusive language detection: A survey. *Personal and Ubiquitous Computing.* Advance online publication. https://doi.org/10.1007/s00779-021-01609-1

Reiners, L., & Schemer, C. (2020). A feature-based approach to assess hate speech in user comments. *Questions de Communication*, *38*, 529–548. https://doi.org/10.4000/questionsdecommunication.24808

Rössler, P. (2014). Comparative content analysis. In F. Esser & T. Hanitzsch (Eds.), *The handbook of comparative communication research* (pp. 459–468). Routledge.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *Proceedings of NLP4CMC*, *17*, 6–9. https://arxiv.org/abs/1701.08118

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, K., & Masip, P. (2011). Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. *International Journal of Press/Politics*, *16*(4), 436–487. https://doi.org/10.1177/1940161211415849

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). *A web of hate. Tackling hateful speech in online social spaces.* Text analytics for cybersecurity and online safety (TA-COS 2016). https://arxiv.org/pdf/1709.10159

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/W17-1101

Siegel, A. A. (2020). Online hate speech. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy. The state of the field and prospects for reform* (pp. 56–88). Cambridge University Press.

Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., … Tucker, J. A. (2021). Trumping hate on Twitter? Online hate speech in the 2016 U.S. Election campaign and its aftermath. *Quarterly Journal of Political Science*, *16*(1), 71–104. https://doi.org/10.1561/100.00019045

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM 2016).* https://arxiv.org/abs/1603.07709

Strippel, C., & Paasch-Colberg, S. (2020). Diskursarchitekturen deutscher Nachrichtenseiten [Discourse architectures of German news sites]. In V. Gehrau, A. Waldherr, & A. Scholl (Eds.), *Integration durch Kommunikation (in einer digitalen Gesellschaft): Jahrbuch der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft 2019* (S. 153–165). DGPuK. https://doi.org/10.21241/ssoar.68129

Triandis, H. C., & Marin, G. (1983). Etic plus emic versus pseudoetic: A test of the basic assumption of contemporary cross-cultural psychology. *Journal of Cross-Cultural Psychology*, *14*, 489–500. https://doi.org/10.1177/0022002183014004007

Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Sage.

Wirth, W., & Kolb, S. (2012). Securing equivalence: Problems and solutions. In F. Esser & T. Hanitzsch (Eds.), *The handbook of comparative communication research* (pp. 469–485). Routledge.

Zannettou, S., Elsherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and characterizing hate speech on news websites. *12th ACM Conference on Web Science*, 125–134. https://doi.org/10.1145/3394231.3397902

II. THEORETICAL PERSPECTIVES: TERMS, CONCEPTS, AND DEFINITIONS

Digital
Communication
Research.de

Recommended citation: Sponholz, L. (2023). Hate speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 143–163). Digital Communication Research. https://doi.org/10.48541/dcr.v12.9

Abstract: Hate speech—communication that attacks a person or a group on the basis of identity factors, such as gender, race, or religion—is one of the main digital threats to democracy. Hate speech has manifold, empirically evidenced consequences for targeted individuals and groups experiencing systematic discrimination and for social cohesion as a whole. Yet, while the upheaval of social media has put the concept in the spotlight, such attention has also structurally transformed its meaning, turning hate speech from a concept with clear defining properties into a family resemblance comprising all kinds of online abuse. This process is far from causing only academic issues. It also sidesteps historical oppression as a defining property and as the reason for which one is targeted by hate speech. Thus, the process has been belittling public animosity against historically oppressed groups, reducing hate speech merely to a matter of offensive language on social media. This chapter shows how and why this conceptual change has taken place and the consequences it unleashes. It specifically addresses the problems of concept stretching, concept shrinking, and the inflation of concepts. Finally, it concludes that such conceptual issues jeopardize the potential that digital media research on hate speech has to provide guidance to a broad range of social actors.

DOI 10.48541/dcr.v12.9

Liriam Sponholz

Hate Speech

1 Hate speech: What is the concept actually for?

Berlin, 2021: The artist Prince Ofori goes to a supermarket and is called the N-word. Customers, workers, security guards—none of them defends him. To the contrary, they start to collectively disparage Ofori, a Black man. Eventually, even the supermarket manager accuses him of being a security risk and throws him out of the store (Amjahid, 2021). Vienna, 2019: A woman wearing a headscarf is spat on in a train station. The woman is called a w***re, pig, and dog and told to go back to the place where she is supposed to belong. "The FPÖ [Austrian far right "Freedom Party," L.S.] will take you all," shouts her harasser ("Alltagsrassismus. Angespuckt und beschimpft," 2019).

In the 1980s, law scholars associated with historically oppressed groups sought to tackle this kind of situation by coining the concept of hate speech to describe the communication of animosity against, or disparagement of people of a historically oppressed group on the basis of identity factors (cf. Matsuda, 1989; Stone, 2000; Delgado & Stefancic, 2004, among others). These scholars were involved with critical race theory and had similar experiences on US-American campuses. To tackle the problem, they proposed that severe cases of hate speech (cf. Matsuda, 1989), such as those in Berlin and Vienna, should be outlawed.

At that time, the concept raised concerns about freedom of speech and was highly criticized. Thirty years later, the debate changed (cf. Tontodimamma et al., 2021). People from historically oppressed groups continue to experience the same experiences, but the concept of hate speech, instead of being rejected, has now been stretched and made ambiguous, leading to a downplaying of the problem.

Nowadays, digital communication enables everyone to gain insight into what it means to be publicly disparaged. By doing so, the digital transformation of the public sphere put the concept of hate speech in the spotlight but has also led to concept stretching (Collier & Mahon, 1993), that is, to the application of the term "hate speech" to cases that do not match its defining properties (cf. Sponholz, 2020).

Researchers on digital communication have been particularly guilty of damaging the clarity of the concept, often without realizing that they are doing so. For instance, they mention the original concept of hate speech in the theoretical part of their studies and then apply the term to cases of online harassment against journalists, online incivility, online abuse, and other forms of conflict that do not match the defining properties of the concept they have just mentioned (cf. Tontodimamma et al., 2021).

However, incidents such as those in Berlin and Vienna have been framed as other than hate speech, with a resultant downplaying of their severity. Therefore, the concept was appropriated by the same patterns of power asymmetry that it was intended to counter.

This chapter sheds light on the problem by analyzing how a concept coined by critical race theory came to be equated with online harassment, what role academic research has played in this development, and why this equation is a problem. As will be shown, hate speech is not a catch-all term for all conflict-related issues involving online communication, nor can it be replaced by catch-all terms such as "online hate." While on the one hand there are serious conceptual issues, such as concept stretching and shrinking, on the other hand, there is a broad consensus among different social actors about what hate speech is. Thus, communication and media scholars should play a significant role in overcoming these issues since hate speech is the key to understanding, explaining, empirically assessing, and tackling extreme forms of symbolic discrimination, one of the most severe digital threats to democracy and social cohesion.

2 Why do concepts matter?

Concepts are not merely a matter of abstract discussion among academics. They constitute a symbolical resource. They are deployed not only to determine how a research subject is assessed empirically but also to evaluate a situation, to define a problem (Thielmann 2004, p. 292, p. 310), to examine the way in which that problem should be tackled, which policies should be employed to tackle it (Palonen, 1999), which statistics should be used to underpin those policies, and—in the case of conflict regulation—who and what should be outlawed and how. When so many rides on a concept, the process of defining the concept becomes a struggle over a resource (Cobb & Elder, 1972), with politicians, governments, and digital platform companies fighting for a definition that best suits their political or economic interests.

"Hate speech" paradigmatically illustrates the struggle over concepts as symbolic resources. Far right actors build their media capital by making disparaging statements against Black people, Indigenous people, Jewish people, LGBTQ people, women, and Muslims and present themselves as victims of hate speech when they face offensive language in response to these statements. This is the case, for instance, when the Austrian far-right leader Heinz Christian Strache complained of hate against his party (Strache sieht in FPÖ-Hass, 2015).

In the sociotechnical realm, digital platform companies, whose economic model is based on interactions, have managed to establish the idea that "the best remedy against bad speech is more speech" (Brändlin, 2016). In line with this principle, Facebook launched the Online Civil Courage Initiative, a project encouraging people to speak up against hate speech. Its CEO, Mark Zuckerberg, also defended the right of Holocaust deniers "to be wrong" (Levin & Solon, 2018). The company only agreed to ban Holocaust denial content under pressure in 2020 (Bickert, 2020). In this context, hate speech has been treated as a matter of uncivil comments (see Coe et al., 2014, and Bormann & Ziegele in this collection for a discussion of the incivility concept), although incivility is not necessarily linked to identity factors, and hate speech, whether online or not, is not restricted to comments or content. However, by turning hate speech into a matter of incivility, digital platform companies a) veil their own role in triggering hate speech (for instance, through scoring or recommendation algorithms); b) enable haters to continue generating interactions and building networks around discriminatory content; c) induce individual

users and collective actors such as high-profile, well-intentioned organizations from civil society to work for them by producing content against hate speech; d) increase interactions not only through hate speech but also through counter speech; and e) polish their images by promoting such initiatives while tolerating hate speech. In taking this course, they fail to tackle the problems that individuals and societies have been suffering as a consequence of group-targeting, offensive, and inflammatory speech on social media, as the genocide in Myanmar, the riots in Chemnitz in Germany, and the online mobilization that led to the storming of the Capitol in the US illustrate.

3 What actually is hate speech?

Defining hate speech pose a particular challenge for research on digital communication, specifically with regard to online content moderation and automated detection of hate speech. On the one hand, researchers complain that there is no "universally accepted" concept of hate speech (MacAvaney et al., 2019, p. 2). On the other hand, they not only fail to make contributions that tackle this problem but even create more ambiguity by associating the term with different classes of objects, such as:

> *Abusive messages*, *hostile* messages, or *flames*. More recently, many authors have shifted to employing the term *cyberbullying* (Xu et al., 2012; Hosseinmardi et al., 2015; Zhong et al., 2016; Van Hee et al., 2015; Dadvar et al., 2013; Dinakar et al., 2012). The actual term *hate speech* is used by Warner and Hirschberg (2012), Burnap and Williams (2015), Silva et al. (2016), Djuric et al. (2015), Gitari et al. (2015), Williams and Burnap (2015), and Kwok and Wang (2013). Further, Sood et al. (2012a) worked on detecting (personal) *insults*, *profanity*, and user posts that are characterized by *malicious intent*, while Razavi et al. (2010) referred to *offensive language*. Xiang et al. (2012) focused on *vulgar language* and *profanity-related offensive content*. (Schmidt & Wiegand, 2017, p. 2-3)

However, the question remains: What is the problem with the concept of hate speech? Answering this question requires an understanding of what a concept is and what it is made up of.

Concepts are basically a matter of word and meaning (intension) and meaning and things (extension) (Sartori, 1984). This is the classical structure of a concept

(Marsteintredet & Malamud, 2020, p. 1025). In the academic context, concepts are applied by researchers to identify, describe, classify, understand, or explain what they observe (Sellars, 2016, p. 4). The intension of a concept consists of defining properties, that is, *criteria* that delimitate the scope of the term. The extension, in turn, determines the *class of objects* to which this meaning applies. Intension and extension are indirectly proportional: the fewer defining properties a concept has, the more abstract it is. The more abstract it is, the greater the number of objects that match it (Sartori 1984, p. 45).

Deficiencies in the intension and extension of a concept create different issues. Problems with intension create ambiguity. This is the case when the meaning of a term is not anchored in defining properties. Problems with the extension of a concept create vagueness. This is the case when a concept is too abstract, which makes the class of objects it applies unclear (Sartori, 1984, p. 27).

Hence, the question of what the problem with the concept of hate speech actually is can be answered. First of all, the problem does not lie in the intension of the concept.

The term "hate speech" is drawn by the following defining properties (DP): attacks (DP1) based on an identity factor (DP2), which are symbolic in nature (DP3) (Matsuda, 1989; Stone, 2000; Delgado & Stefancic, 2004; among others). Hate speech—whether online or not—is also a matter of communication in places of public space (cf. Sellars, 2016; Delgado & Stefancic, 2004). Nonetheless, this is not a classical defining property, as it may also apply to other cases of communication of disparagement, such as online incivility (cf. Sponholz, 2020).

There is a broad consensus, from international organizations to digital platform companies, about the linkage of the term hate speech with these defining properties, as follows:

- United Nations: Any kind of communication in speech, writing or behavior [DP3] that attacks or uses pejorative or discriminatory language [DP1] with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor [DP2]. (United Nations, 2020, p. 8)

- Facebook Company: We define hate speech as a direct attack [*DP1*] against people on the basis of what we call protected characteristics: race, ethnicity,

national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease [*DP2*]. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation [*DP3*]. (Facebook, 2021)

- Twitter: You may not promote violence against or directly attack or threaten other people [*DP1, DP3*] on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease [*DP2*]. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories. (Twitter, 2020)

- Council of Europe: The term "hate speech" shall be understood as covering all forms of expression [*DP3*] which spread, incite, promote or justify [*DP1*] racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin [*DP2*]. (Weber, 2009, p. 3)

These definitions are not identical. Nevertheless, as Sartori (1984, p. 29) asserts, a single concept can yield several conceptualizations. The same concept may, for instance, yield both denotative and operational definitions, but as long as different definitions possess the same defining properties, they still constitute the same concept.

4 What is not hate speech?

The intension of the concept of hate speech not only enables a determination of what hate speech is but also what it is not.

First, hate speech is not negative stereotypes or misrepresentation but a matter of attacks. Defining hate speech as "negative speech that targets individuals or groups" or even as "statements of disagreement, such as indications that the

group is wrong, what they claim is false or what they believe is incorrect" (Bahador & Kerchner, 2019, pp. 4–6) downplays the severity of the problem.

Negative stereotypes may be used by hate speakers, but they alone are not enough to constitute an attack. To be considered hate speech, it should be applied consciously or intentionally (Delgado & Stefancic, 2004). What precisely constitutes an attack is delineated in, for instance, General Recommendation Nr. 35 of the Committee on the Elimination of Racial Discrimination (2013): incitement of hatred, contempt, exclusion, or violence; threats or expressions of insults, ridicule, or slander (for an overview, see Table 1).

In this context, "conscious" means that the speaker is aware of the disparaging potential of the content, as in the case of identity-targeting offensive speech. "Intentional," in turn, means that symbolic disparagement is a way of achieving a goal. This goal may be hurting someone or derogating a group due to the ideological convictions of the speaker (prior intention), or it may be something other, such as gaining media attention or attracting voters in an election (subsidiary intention) (cf. Searle, 1980; Sponholz, 2018).

Second, not all disparagements of groups qualify as hate speech (cf. Sellars, 2016), only those based on identity factors in correlation with historical oppression (Matsuda, 1989) or systematic discrimination (Gelber, 2021). Hate speech represents the communicative ring on a chain of manufacturing human inferiority (Sponholz, 2018, p. 48), in which antinomies (Marková, 2003) on collective features such as race, gender, origin, religion, and sexual orientation are intentionally activated through communication. It works as another layer in the long-standing process of subordination (Matsuda, 1989). This is why "not everyone has known the experience of being victimized by racist, misogynist, or homophobic speech, and we do not share equally the burden of the societal harm it inflicts" (Lawrence III, 1993, p. 56).

A definition that takes a broader view of groups, contending that theoretically any group can become the target of hate speech, is exactly what critical race theorists were fighting against. It means erasing the power asymmetry that Lawrence III (1993) referred to. This does not mean that people can be attacked symbolically only if they possess one of the designated identity factors, but it clearly means that if there is not an identity factor involved, this kind of abuse or harassment should not be classified as hate speech.

Table 1: Defining properties of hate speech

Who	What		Where
Collective feature corresponding to an identity factor (e.g., gender/race) related to an unprivileged position (e.g., Women/Black people)	Dissemination of	Discriminatory ideas	Places of public life (workplace, school, university campus, media)
	Incitement of	• Hatred • Contempt • Exclusion • Violence	
	Incitement through	• Public denial of genocide and crimes against humanity	
	Threat		
	Justification of	• Genocides and crimes against humanity	
	Expression of	• Insults • Ridicule • Slander	

Source: Own illustration, based on Committee on the Elimination of Racial Discrimination (2013), Delgado and Stefancic (2004), and Matsuda (1989), among others

Third, hate speech is not necessarily a matter of (offensive) language but of communication (Stone, 2000; United Nations, 2020). This is particularly important when it comes to social media, where the media logic is not based on content—as with the mass media—but on interactions (van Dijck & Poell, 2013). As a result, digital communication is not only a matter of media objects (such as online comments) but also of other digital objects (Langlois & Elmer, 2013): network objects, such as hashtags (Poole et al., 2021); and phatic objects, that is, the networks generated by interactions on social networking digital platforms (Chaudhry, 2015). For this reason, hate speech cannot be detected solely by content analysis but also requires social network analysis and social media metrics analysis (Sponholz, 2021). As a consequence, countering the problem should not be limited to content moderation, but should include debates on de-platforming (Ali et al., 2021), cross-platform approaches (Johnson et al., 2019), and broader concepts of platform governance.

It is worth noting that neither hate (as an emotion) nor illegality are defining properties of the concept of hate speech. Brown (2017a) labels this first misunderstanding "the myth of hate," that is, the idea that emotions, feelings, or attitudes of hate or hatred are part of the essential nature of hate speech (cf. also Tetrault, 2019). The roots of hate speech are ideologies of inequality, such as racism, sexism, homophobia, and Islamophobia, not affective action. These ideologies gather enough empirical evidence that they can be expressed "rationally"; that is, they are underpinned by arguments (Sponholz, 2018; Meddaugh & Kay, 2009). To put it briefly, David Irving's denying the Holocaust is not an emotional response.

With regard to legality, although law scholars coined the concept, they also made it clear from the beginning that only a very strict range of cases could be regulated (Matsuda, 1989). Further, the upsurge of the concept in public and academic debate has not been triggered by legal issues but by the rise of social media (Paz et al., 2020; Sponholz, 2019b). Media and communication researchers have dominated this research agenda since the 2010s and apply the concept first to observe conflict dynamics, and not to matters of conflict regulation.

As seen above, the concept of hate speech has a clear definition (intension), with broad consensus on the application of the term anchored in the same defining properties. However, in spite of a clear intension, the concept is highly abstract and does not provide a clear extension, that is, an explicit scope for the class of objects to which it applies to (Sartori, 1984). Such vagueness lends it flexibility, but it also poses a challenge for empirical research.

However, it should be highlighted that flexibility in this context does not mean that the concept of hate speech can be applied to all kinds of wrongdoing in online communication, as often happens in digital media research. It actually means that the concept can be applied to a broad range of empirical manifestations, such as online firestorms or hashtag activism, as *long as they match the defining properties.*

By applying the term "hate speech" to offensive language in general, research on digital communication not only fails to tackle the concept's vagueness but also generates new conceptual issues. This is concept stretching—that is, the application of a concept to cases that do not fit its defining properties. In other instances, researchers on digital communication, particularly those working in automated detection, are shrinking the concept by applying it only to identity-targeting

group derogatory labels, such as racial slurs. In the third scenario, an inflation of concepts has taken place, failing to solve old concerns and generate new ones.

5 Conceptual issues within academic research

Epistemologically, definitions are pivotal to increasing and even enabling the efficiency of science (Potthof, 2017). They allow social scientists to avoid talking at cross-purposes when addressing a research subject, which means that empirical findings can be compared and knowledge can be accumulated. This, in turn, allows for the development of theories. Theorizing is imperative to understanding and explaining the puzzle of social reality.

However, when scholars expand the comparative perspective among research areas, they also tend to broaden the meaning of the concepts to incorporate a larger realm of observations under expanded rubrics (Sartori, 1984). The result is a conceptual travelling. This happened to hate speech, a concept coined by law scholars in the 1980s, when the rise of social media in the early 2000s turned the term into an interdisciplinary research subject (Paz et al., 2020; Sponholz, 2019b).

Although conceptual traveling may result in a concept being more relevant, it may also feed a conceptual stretching (Collier & Mahon, 1993). By not considering the intension of hate speech, researchers on media and communication have been applying the concept to a class of objects that do not possess the same defining properties, such as online harassment against journalists (Obermaier et al., 2018).

Different concepts might have the same, contingent, or accidental characteristics, but if they do not possess all the defining properties, they are not the same (Sartori, 1984). By applying the concept to classes of objects that do not belong under the same umbrella, conceptual stretching creates ambiguity and hinders the comparability of empirical evidence, which harms the accumulation of knowledge and makes it harder to provide qualified guidance when policies are formulated to tackle the problem.

Concept stretching jeopardizes hate speech research by erasing not only a defining property of the concept but also the concept's very reason for being: catching disparaging communication that is based on a collective feature linked to historical oppression or systematic discrimination.

6 Concept shrinking

When attempting to identify hate speech by automated means, researchers on digital media often try to solve the problem of vagueness by reducing the concept to a matter of racial slurs or symbols or open threats ("kill," "rape") (cf. Schmidt & Wiegand, 2017). By defining a hate speech message as a message that contains "hate words" (Silva et al., 2016), lexicon- and keyword-based approaches cannot identify cases that do not contain any "hateful" words (e.g., cases that use figurative or nuanced language) but that still deliberately discriminate symbolically against a group (MacAvaney et al., 2019).

Reducing hate speech to a matter of insults, for instance, would mean coming to the conclusion that racist groups are not libeling or inciting discrimination against historically oppressed groups in instances where they target people due to race, origin, or religion but refrain from writing the N-word or displaying a swastika.

Furthermore, lexicon-based approaches, including those based on offensive language, such as group derogatory labels, are capable of empirically assessing only a small proportion of cases. In the case of group libels, they are also considered a less severe form of hate speech (United Nations, 2020). Such approaches also fail to make visible those collective actors and public figures who apply hate speech as a kind of strategic communication, because such actors tend to avoid blatantly discriminatory language (Gerstenfeld et al., 2003; Kleinberg et al., 2021). Even in the case of hate speech during genocides, overt messages, such as open threats ("Go out and kill!"), constitute an exception (Benesch, 2004, p. 67; Straus, 2007, p. 612).

For this reason, several stakeholders have underlined that hate speech is not a matter of language but of communication (Stone, 2000; United Nations, 2020). While language refers to a system of codes (Lewandowski, 1994), communication involves speakers, messages, means of dissemination, audience, and historical, social, and political context. This is particularly relevant for research into hate speech on social media, as reducing the problem to content does not fit the media logic of such digital platforms, as analyzed earlier.

7 Inflation of concepts

Researchers have also tried to sidestep conceptual issues with hate speech by replacing it with, for instance, the concepts of "online hate" and "extreme speech."

"Online hate" is one of the catch-all terms that scholars have been using in their attempts to capture empirical developments in digital communication. "Online hate" incorporates issues such as "online toxicity," "online abusive language," "cyberbullying," "online harassment," and "online firestorms." That is, it comprises much more than online hate speech (Waqas et al., 2019). Hence, "online hate" is a family resemblance rather than a classical concept:

> One distinctive feature of family resemblance concepts is the fact that everything that falls under the concept shares at least one similar quality, feature, or descriptive property with at least one other thing that falls under the concept, even if there is no single quality, feature, or descriptive property that is common to all things that fall under the concept. (Brown, 2017b, p. 596)

Changing the concept structure from classical to family resemblances has been a successful strategy to capture empirical developments. Nonetheless, at least three problems can be caused by such conceptual change. First, it inhibits the recognition of hate speech. Second, it creates the danger of causal and conceptual confusion. Third, it may have serious political consequences.

Transforming hate speech into a matter of family resemblance means applying the same term based on different criteria depending on the context (Wennerberg, 1998, p. 64), opening the door to all kinds of political instrumentalization. Such adaptions entail many risks, including freedom of speech. The concept may even be turned against historically oppressed groups seeking to speak out about their situations of oppression (cf. Benesch, 2014; Gagliardone et al., 2015).

Applying "hate speech" and "online hate" interchangeably also means erasing discrimination and power asymmetries as the roots of the problem (cf. also Matamoros-Fernández & Farkas, 2021). By doing so, researchers, instead of investigating, assume that insults, threats, or incitement against Black people, women, Jewish people, Muslims, or other groups are the same as any other kind of disparaging communication, such as individual insults and slanders. In assuming that, they fail to look for empirical evidence that would, in the case of hate speech, prove its true nature.

This failure on the part of researchers is particularly problematic, given that it is often hate speech—and not other forms of online abuse—that plays a pivotal role in political developments, such as the rise of the far right and its linkage to digital communication (Art, 2020; Froio & Ganesh, 2019; Sponholz, 2019a). In a nutshell, family resemblance concepts such as "online hate" are successful at capturing empirical developments in digital communication but cannot replace the concept of hate speech.

Relabeling hate speech, in turn, might raise new issues, as is the case with the concept of "extreme speech," as applied within digital communication research (the term was applied before, at the end of the 2000s, by law scholars such as Hare and Weinstein (2009) in another context).

In digital communication research, "extreme speech" aims to provide an alternative, non-regulatory approach to hate speech since:

> The use of hate speech (...) embodies the colonial logic of "yet-to-be modern" societies prone to "emotions," manipulation, and public frenzy, which have to be tested against the high values of calm rationality of Western liberal democracy. (Udupa & Pohjonen, 2019, p. 3055)

To overcome the "Western bias" of the concept of hate speech, the authors propose the concept of "extreme speech," a framework to "capture digital cultures that push and provoke the limits of legitimate speech along the twin axes of truth-falsity and civility-incivility" (Udupa & Pohjonen, 2019, p. 3051).

This is particularly striking because, in the case of political incivility, the concept is deeply ingrained in the US legal and political debate on civil discourse. Moreover, as it is defined in the framework of deliberative theories, political incivility also relies heavily on values such as rationality (Massaro & Stryker, 2012, p. 379, p. 414). Regarding civility in general, it acts even as a further mechanism of discrimination, such as when the language used by members of historically oppressed groups use is classified as offensive language (Sap et al., 2019). This happens because the concept is supposed to be a high value of Western societies and intimately connected to social rank, class status, political hierarchy, and relations of power (cf. Harcourt, 2012). Hate speech, in turn, requires neither incivility nor irrationality, as it constitutes a matter of discrimination.

8 Why working on the concept of hate speech?

This chapter sheds light on the conceptual change the term "hate speech" has experienced, what role academic research has played in this, and what problems the change causes.

Hate speech is a theoretically sound scientific concept with a clear intention: it is anchored in three defining properties (DP), which work as criteria to disambiguate it: attacks (DP1) based on an identity factor (DP2) and that are symbolic in nature (DP3). Further, hate speech is a matter of communication in public life.

The clear intension of the concept provides a first approach to determining what is not hate speech. Yet, the concept is also highly abstract, which makes it difficult to identify to which class of objects it applies. Not having a clear extension lends it flexibility but also poses a challenge to empirical research.

However, by trying to overcome conceptual challenges in empirical research, digital media research has been creating new conceptual issues, such as: a) concept stretching—applying the concept to cases that do not match the defining properties of hate speech; b) concept shrinking—reducing the problem to a matter of content, as in the case of lexicon-based approaches; and c) an inflation of concepts—using the term interchangeably with "online hate" or replacing it with new terms, which creates its own conceptual issues.

This is problematic because concepts are not merely a matter of abstract discussions among academics. Concepts constitute a symbolic resource for defining problems and determining how they are going to be tackled. As a consequence, erasing the defining properties of hate speech creates many political issues. Transforming hate speech into a matter of family resemblance means that the concept can be "adapted" to any context, opening the door to all kinds of political instrumentalization. Applying the term "hate speech" and other forms of online abuse interchangeably downplays the problem as merely a matter of bad behavior among users in online conversations.

Replacing the concept of hate speech with that of online hate erases discrimination and power asymmetry from digital media research as the roots of this specific but highly harmful kind of communication.

So, what is the purpose of the concept of hate speech? Why should digital media researchers work on a concept that raises so many conceptual issues? First, of all the concepts applied to analyze threats in digital communication, hate speech is

probably the one with the longest research tradition. Many issues being discussed in the field of platform governance, for instance, have been analyzed for decades in hate speech research. Second, in spite of vagueness with regard to which cases the concept can be applied to, the concept is unambiguous: there is a broad consensus among different social actors about what hate speech means. Third, hate speech is a much more severe digital threat than insulting people on social media.

By abandoning, making ambiguous, or shrinking the concept, digital media research is jeopardizing its potential to tackle one of the most socially relevant problems in its field.

Liriam Sponholz is a postdoctoral researcher at the German Centre for Integration and Migration Research (DeZIM) in Berlin, Germany. https://orcid.org/0000-0001-7875-4273

References

Ali, S., Saeed, M. H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Understanding the effect of deplatforming on social networks. *WebSci '21*, June 21–25, Virtual Event, United Kingdom. https://seclab.bu.edu/people/gianluca/papers/deplatforming-websci2021.pdf

Alltagsrassismus. Angespuckt und beschimpft: Junge Muslima wird auf offener Straße angegriffen [Everyday racism. Spat on and verbally abused: Young Muslim women are attacked on the street]. (2019, April 3). *Stern*. https://www.stern.de/neon/wilde-welt/gesellschaft/alltagsrassismus--junge-muslima-wird-in-wien-auf-der-strasse-angespuckt-8650518.html

Amjahid, M. (2021, April 29). Neuköllner über Alltagsrassismus: "Nur Zufall, dass es bei Aldi war" [Neuköllner on everyday racism: "It was just chance that it was at Aldi"]. *Taz*. https://taz.de/Neukoellner-ueber-Alltagsrassismus/!5762911/

Art, D. (2020). The myth of global populism. *Perspectives on Politics*, 1–13. https://doi.org/10.1017/S1537592720003552

Bahador, B., & Kerchner, D. (2019). *Monitoring hate speech in the US media* (Working Paper). The George Washington University. https://cpb-us-e1.wpmucdn.com/blogs.gwu.edu/dist/8/846/files/2019/03/Monitoring-Hate-Speech-in-the-US-Media-3_22-z0h5kk.pdf

Benesch, S. (2004). Inciting genocide, pleading free speech. *World Policy Journal,* 21(2), 62–69.

Benesch, S. (2014). Defining and diminishing hate speech. In P. Grant (Ed.), *State of the world's minorities and indigenous peoples 2014* (pp. 19–25). http://minorityrights.org/wp-content/uploads/old-site-downloads/mrg-state-of-the-worlds-minorities-2014-chapter02.pdf

Bickert, M. (2020, October 12). *Removing Holocaust denial content. Facebook.* https://about.fb.com/news/2020/10/removing-holocaust-denial-content/

Brändlin, A.-S. (2016, February 26). Facebook's Zuckerberg to tackle hate speech. *Deutsche Welle*. https://www.dw.com/en/facebooks-zuckerberg-to-stamp-out-hate-speech-in-germany/a-19078185

Brown, A. (2017a). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, *36*, 419–468.

Brown, A. (2017b). What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, *36*(5), 561–613.

Chaudhry, I. (2015). #Hashtagging hate: Using Twitter to track racism online. *First Monday*, *20*(2). https://doi.org/10.5210/fm.v20i2.5450

Cobb, R. W., & Elder, C. D. (1972). *Participation in American politics: The dynamics of agenda building*. Allyn and Bacon.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Collier, D., & Mahon, J. E. (1993). Conceptual "stretching" revisited: Adapting categories in comparative analysis. *The American Political Science Review*, *87*(4), 845–855. https://doi.org/10.2307/2938818

Committee on the Elimination of Racial Discrimination. (2013, September 26). *General recommendation No. 35*. http://tbinternet.ohchr.org/_layouts/treatybodyexternal/Download.aspx?symbolno=CERD/C/GC/35&Lang=en

Delgado, R., & Stefancic, J. (2004). *Understanding words that wound*. Westview Press.

Facebook. (2021). *Hate speech*. https://www.facebook.com/communitystandards/recentupdates/hate_speech

Froio, C., & Ganesh, B. (2019). The transnationalisation of far right discourse on Twitter. Issues and actors that cross borders in Western European democracies. *European Societies*, *21*(4), 513–539. https://doi.org/10.1080/14616696.2018.1494295

Gagliardone, I., Gal, D., Alves, T., & Martínez, G. (2015). *Countering online hate speech*. UNESCO Series on Internet Freedom. http://unesdoc.unesco.org/images/0023/002332/233231e.pdf

Gelber, K. (2021). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, *24*(4), 393–414. https://doi.org/10.1080/13698230.2019.1576006

Gerstenfeld, P. B., Grant, D. R., & Chiang, C.-P. (2003). Hate online: A content analysis of extremist internet sites. *Analyses of Social Issues and Public Policy*, *3*(1), 29–44. https://doi.org/10.1111/j.1530-2415.2003.00013.x

Harcourt, B. (2012). The politics of incivility. *Arizona Law Review*, *54*(2), 345–374. https://scholarship.law.columbia.edu/faculty_scholarship/638

Hare, I., & Weinstein, J. (Eds.). (2009). *Extreme speech and democracy*. Oxford University Press.

Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*(7773), 261–265. https://doi.org/10.1038/s41586-019-1494-7

Kleinberg, B., van der Vegt, I., & Gill, P. (2021). The temporal evolution of a far-right forum. *Journal of Computational Social Science*, *4*, 1-23. https://doi.org/10.1007/S42001-020-00064-X

Langlois, G., & Elmer, G. (2013). The research politics of social media platforms. *Culture Machine*, *14*, 1–17. https://culturemachine.net/wp-content/uploads/2019/05/505-1170-1-PB.pdf

Lawrence III, C. R. (1993). If he hollers let him go: Regulating racist speech on campus. In M. J. Matsuda, C. R. Lawrence III, R. Delgado, & K. W. Crenshaw (Eds.), *Words that wound: Critical race theory, assaultive speech, and the First Amendment* (pp. 53–88). Westview Press.

Levin, S., & Solon, O. (2018, July 18). Zuckerberg defends Facebook users' right to be wrong – even Holocaust deniers. *The Guardian*. https://www.theguardian.com/technology/2018/jul/18/zuckerberg-facebook-holocaust-deniers-censorship

Lewandowski, T. (1994). *Linguistisches Wörterbuch* [Linguistic dictionary]. UTB für Wissenschaft.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, 14(8), Article e0221152. https://doi.org/10.1371/journal.pone.0221152

Marková, I. (2003). *Dialogicality and social representations: The dynamics of mind*. Cambridge University Press.

Marsteintredet, L., & Malamud, A. (2020). Coup with adjectives: Conceptual stretching or innovation in comparative research? *Political Studies*, *68*(4), 1014–1035. https://doi.org/10.1177/0032321719888857

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy, and emerging evidence on civility and effective democratic engagement. *Arizona Law Review*, *54*(2), 375–441. https://arizonalawreview.org/pdf/54-2/54arizlrev375.pdf

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205–224. https://doi.org/10.1177%2F1527476420982230

Matsuda, M. J. (1989). Public response to racist speech: Considering the victim's story. *Michigan Law Review*, 87(8), 2320–2381. https://doi.org/10.2307/1289306

Meddaugh, P. M., & Kay, J. (2009). Hate speech or "reasonable racism?" The other in Stormfront. *Journal of Mass Media Ethics*, *24*(4), 251–268. https://doi.org/10.1080/08900520903320936

Obermaier, M., Hofbauer, M., & Reinemann, C. (2018). Journalists as targets of hate speech: How German journalists perceive the consequences for themselves and how they cope with it. *SCM Studies in Communication and Media*, *7*(4), 499–524. https://doi.org/10.5771/2192-4007-2018-4-499

Palonen, K. (1999). Rhetorical and temporal perspectives on conceptual change. *Redescriptions: Political Thought, Conceptual History and Feminist Theory*, *3*(1), 41–59. http://doi.org/10.7227/R.3.1.4

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. *SAGE Open*, *10*(4). https://doi.org/10.1177/2158244020973022

Poole, E., Giraud, E. H., & de Quincey, E. (2021). Tactical interventions in online hate speech: The case of #stopIslam. *New Media & Society*, *23*(6), 1415–1442. https://doi.org/10.1177/1461444820903319

Potthof, M. (2017). Probleme von Begriffsbildung und -verwendung in der Kommunikationswissenschaft [Problems of concept formation and use in communication science]. *Studies in Communication | Media*, *6*(2), 95–127. https://doi.org/10.5771/2192-4007-2017-2-95

Sartori, G. (Ed.). (1984). *Social science concepts: A systematic analysis*. Sage.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech detection, in *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1668–1678), https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Spain, 1–10, https://www.aclweb.org/anthology/W17-1101.pdf

Searle, J. R. (1980). The intentionality of intention and action. *Cognitive Science*, *4*(1), 47–70. https://doi.org/10.1080/00201747908601876

Sellars, A. (2016). *Defining Hate Speech* (December 1, 2016). Berkman Klein Center Research Publication No. 2016-20, Boston Univ. School of Law. http://dx.doi.org/10.2139/ssrn.2882244

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). *Analyzing the targets of hate in online social media*. https://arxiv.org/pdf/1603.07709.pdf

Sponholz, L. (2018). *Hate Speech in den Massenmedien: Theoretische Grundlagen und empirische Umsetzung* [Hate speech in the mass media: Theoretical foundations and empirical implementation]. Springer Verlag. https://doi.org/10.1007/978-3-658-15077-8

Sponholz, L. (2019a). Hate Speech in Sozialen Medien: Motor der Eskalation? [Hate speech: Social media as trigger?]. In H. Friese, M. Nolden, & M. Schreiter (Eds.), *Rassismus im Alltag: Theoretische und empirische Perspektiven nach Chemnitz* [Racism in everyday life: Theoretical and empirical perspectives after Chemnitz] (pp. 157–178). transcript Verlag. https://doi.org/10.14361/9783839448212-009

Sponholz, L. (2019b). Hate Speech: Viel mehr als böse Wörter [Hate speech: Much more than bad words]. In E. Greif & S. Ulrich (Eds.), *Hass im Netz: Grenzen digitaler Freiheit. Linzer Schriften zu Gender und Recht*, 63 [Online hate: Limits to digital freedom. Linz writings on gender and law, 63] (pp. 1–30). Trauner Verlag.

Sponholz, L. (2020). Der Begriff „Hate Speech" in der deutschsprachigen Forschung. Eine empirische Begriffsanalyse [The term "hate speech" in German-language research: An empirical concept analysis]. *SWS-Rundschau*, *60*(1), 43–65.

Sponholz, L. (2021). Hass mit Likes: Hate Speech als Kommunikationsform in den Social Media [Hate with likes: Hate speech as communication in social media]. In S. Wachs, B. Koch-Priewe, & A. Zick (Eds.), *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen* [Hate speech – Multidisciplinary analysis and options for action] (pp. 15–37). Springer VS. https://doi.org/10.1007/978-3-658-31793-5_2

Stone, G. R. (2000). First amendment. In L. W. Levy, K. L. Karst, & A. Winkler (Eds.), *Encyclopedia of the American Constitution* (pp. 1055–1057). Macmillan.

Strache sieht in "FPÖ-Hass" Beleg für eigene Relevanz [Strache sees hate from "FPÖ" as evidence of own relevance]. (2015, July 14). *Standard.at.* https://www.derstandard.at/story/2000019117104/strache-sieht-in-fpoe-hass-beleg-fuer-eigene-relevanz

Straus, S. (2007). What is the relationship between hate radio and violence? Rethinking Rwanda's "Radio Machete". *Politics & Society*, *35*(4), 609–637. https://doi.org/10.1177/0032329207308181

Tetrault, J. E. C. (2019). What's hate got to do with it? Right-wing movements and the hate stereotype. *Current Sociology*, *69*(1), 3–23. https://doi.org/10.1177%2F0011392119842257

Thielmann, W. (2004) Begriffe als Handlungspotentiale [Concepts as potential for action]. In: *Linguistische Berichte*, *199*, 287–312.

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*(1), 157–179. https://doi.org/10.1007/s11192-020-03737-6

Twitter. (2020, December 2). *Updating our rules against hateful conduct*. https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html

Udupa, S., & Pohjonen, M. (2019). Extreme speech and global digital cultures: Introduction. *International Journal of Communication*, *13*, 3049–3067. https://ijoc.org/index.php/ijoc/article/view/9102/2710

United Nations. (2020). *United Nations strategy and plan of action on hate speech: Detailed Guidance on Implementation for United Nations Field Presences*. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf

van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, *1*(1), 2–14. http://dx.doi.org/10.17645/mac.v1i1.70

Waqas, A., Salminen, J., Jung, S.-G., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS One*, *14*(9), Article e0222194. https://doi.org/10.1371/journal.pone.0222194

Weber, A. (2009). *Manual on hate speech*. Council of Europe Publ.

Wennerberg, H. (1998). Der Begriff der Familienähnlichkeit in Wittgensteins Spätphilosophie [The concept of family resemblance in Wittgenstein's late philosophy]. In E. v. Savigny (Ed.), *Ludwig Wittgenstein* (pp. 41–69). De Gruyter.

Digital
Communication
Research.de

Recommended citation: Frischlich, L. (2023). Hate and harm. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 165–183). Digital Communication Research. https://doi.org/10.48541/dcr.v12.10

Abstract: From a psychological point of view, hate speech can be conceptualized as harmful intergroup communication. In contrast to other forms of incivility, hate speech is directed toward individuals because of their (perceived) social identity. This explains why the harm of hate speech can extend to entire social groups and societies. Hate speech therefore cannot be separated from pre-existing power structures and resource inequalities, as its harm is particularly severe when coping resources are already deprived. Psychological research on the perpetrators of hate speech links hate speech to a lack of empathy and the acceptance of, or even desire for social inequalities. In summary, hate speech jars the norms of democratic discourses by denying fellow humans basic respect and violating the democratic minimal consent of human equality. Overall, the chapter demonstrates the usefulness of a (social) psychological perspective on the harms of hate speech for both researchers and practitioners.

DOI 10.48541/dcr.v12.10

Lena Frischlich

Hate and Harm

1 Hate speech as "going against" a social identity

In this chapter, I understand hate speech as a specific form of *incivility*, a communication that violates norms (e.g., Kenski et al., 2020; Mutz, 2015; Papacharissi, 2004). Incivility is thereby a "notoriously difficult concept to define" (Coe et al., 2014, p. 660), not least because different perspectives have to be taken into account: the perspective of the perpetrator, the perspective of the attacked, and the perspective of the observer (O'Sullivan & Flanagin, 2003). Further, it is often unclear which norms exactly have been violated, and while explicit vulgar insults are relatively consistently rated as uncivil, more subtle norm transgressions, such as dehumanizing metaphors, or formal forms of incivility (e.g., the use of multiple exclamation marks or grammatically wrong expressions) are less agreed upon (for a discussion, see Chen et al., 2019; Bormann & Ziegele in this volume).

In the following, I focus on hate speech as a subtype of communication "going against" a target (Gagliardone et al., 2016, p. 6). Following Gagliardone et al. (2016), two types of targets can be distinguished in this context, although they may overlap (see also Rossini, 2020). The first type, which Gagliardone et al. (2016) label *offensive speech*, is directed toward individuals and is often studied under labels such as (*cyber-)bullying* (e.g., Festl, 2016) or *trolling* (for a comprehensive overview, see Phillips, 2012). Offensive speech violates interpersonal norms of

politeness (Mutz, 2015, p. 6), for example, by using swear words, and insults, or by mocking the target. The current chapter focuses on the second type of incivility—*hate speech*, which is directed against individuals because of their collective or social identity and reflects a biased attitude toward the targeted group rather than a personal dislike (Silva et al., 2016, p. 3). Although hate speech is not necessarily formally uncivil (i.e., detectable by exclamation marks) or offensive (e.g., using well-known swear words), it can have specific severe effects on those attacked and their social context, wherefore it is sometimes described as "harmful speech" (Bilewicz & Soral, 2020, p. 2).

In the following, I describe these harms for both the individual target and its social context from a predominantly psychological point of view. First, I will show how a social psychological perspective allows for describing the harmful "fallout" of hate speech compared to other types of incivility. Notably, this does not imply that offensive speech, such as cyberbullying, is harmless; however, as I will argue in the following, the fundamentally social nature of hate speech is unique and thus should be treated as such. Second, I will show how individual characteristics, such as personality, attitudes, and emotions, shape the spread of hate speech on the individual micro-level. I will close the chapter by arguing that the suggested perspective allows us to consider both the individual and the social levels when examining the harms of hate speech.

2 Hate speech is directed toward people's social identities

The fallout of hate speech can be explained by social identity and self-categorization mechanics. *Social identity theory* (Tajfel & Turner, 1979) and *self-categorization theory* (Turner et al., 1987) postulate that people not only own a *personal identity* that distinguishes them from others and makes them unique but also multiple *social identities* resulting from their social roles and memberships in social groups or categories. The *ingroups* to which people belong are perceptually and functionally distinct from *outgroups*, that is, groups or categories people do not belong to. The more people identify with their social group, the more they think, feel, and act on behalf of that group. For instance, people can feel nostalgic or guilty on behalf of their nation (Martinovic et al., 2017) and sad or joyful because of their sports teams' performances (for an overview of intergroup

emotions, see Smith & Mackie, 2015). Even random categorizations in artificial groups motivate a distinct treatment of ingroup versus outgroup members and change the neural processing of ingroup and outgroup members (Brewer, 1979; Crocker & Schwartz, 1985; Ratner & Amodio, 2013).

Due to the central role of social groups in ones' identity, people are motivated to see their ingroup(s) in a positive light and to perceive them as positively distinct from outgroups (Tajfel & Turner, 1979). Such positive ingroups are an important factor in psychosocial well-being (Haslam et al., 2016; Jetten et al., 2012) and a pillar of individuals' resilience in light of hardships (Muldoon et al., 2019). To preserve a positive group image, people are biased to perceive ingroup compared to outgroup members as being more trustworthy (Yamagishi & Kiyonari, 2000) and less flawed (Koval et al., 2012). In times of uncertainty (Hogg et al., 2007), when people feel socially ostracized (Pfundmair & Wetherell, 2019), or are reminded of their inevitable decay (Frischlich et al., 2015), ingroup biases can even extend to tolerate extremist and violent ingroup members more than outgroup members, as social identities can help individuals cope with these kinds of existential threats (Jonas et al., 2014).

People's social identities differ in their stability (or variability). While some groups are relatively easy to change through re-categorization processes (e.g., when changing one's employer), other social categories are more difficult to change and are repeatedly ascribed to individuals even without their intervention. This is especially true for membership in disadvantaged or visually marked groups (such as gender or ethnicity). Hate speech primarily attacks such stable identities (Bilewicz & Soral, 2020), often relying on century-old stereotypes and longstanding prejudice. To understand the damage caused by hate speech, it is therefore crucial to consider the perspective of socially marginalized groups (Dieckmann et al., 2018) and to understand hate speech as "harmful language" (Leets & Giles, 1999).

From this perspective, hate speech is more closely related to *hate crimes* (Walters et al., 2016) than to impoliteness. *Hate crime*, as a legal category in the UK, is defined as "any crime or incident where the perpetrator's hostility or prejudice against an identifiable group of people is a factor in determining who is victimised" (College of Policing, 2020). Typical hate crimes are incidents of discrimination or even violence against people who are interpreted as members of a certain social group, such as a religious or sexual minority.

Hate crimes are often a "stranger-danger" where perpetrator and victim are unknown to each other though physically (or virtually) existing in the same space (Mason, 2005). Hate crimes are often driven by the motivation to preserve the perceived "natural superiority"[1] of the perpetrator (Perry & Alvi, 2012). Through stereotypes and prejudices, hate speech is embedded in a specific socio-cultural system, with its specific power relationships and specific histories of intergroup conflicts. What makes hate speech specifically harmful speech are the resources for coping with the threat of hate speech, which are unequally distributed among the benefiters of human history and those struggling for their place at the table. A cross-country survey showed that hate speech varies in both reported frequency and attacked targets along socio-cultural lines and long-term narratives in a given context (Reichelmann et al., 2020). For instance, female (compared to male) journalists and politicians are disproportionally often overflooded with hate directed toward them (for media reports, see Carter, 2021; Gardiner et al., 2016).

3 Putting the harm in hate speech: Effects on victims and observers

Hate speech (and other types of incivility not in focal attention here) can have severe negative effects. For instance, a large German study (Geschke et al., 2019) found that among those who had experienced hate speech, only one-third reported no personal consequences, another third reported emotional distress, and 17% reported depression as a consequence of the attacks. The same study also showed that 46% of those who had experienced hate speech refrained from online discussions at least sometimes to avoid attacks, and 51% did not speak about their political orientation online. Similar silencing effects were reported by indigenous Australians in a study by Gelber and McNamara (2015).

Computational simulation studies warn that hate speech can over time erode norms for civil interactions via desensitization (Soral et al., 2018), leading to a "hate speech epidemic," as Bilewicz and Soral (2020, p. 3) termed it. Another computational simulation indicates that even subtle discrimination by a societal

1 The author distances herself from the idea of a natural order in which some human beings or social groups are supreme to others.

majority can cement prejudiced intergroup relationships over time by eroding trust in outgroups (Uhlmann et al., 2018). Hate speech thus diminishes social trust (Näsi et al., 2015), potentially contributing to "spirals of distrust" (Frischlich & Humprecht, 2021, p. 4) and endangering societal cohesion. Further, an experimental study by Hsue et al. (2015) showed that reading uncivil comments targeting minorities motivated more negative attitudes toward the targeted group. Once someone is classified as an outgroup member, people become less able to detect that person's pain (Ma et al., 2011), which in turn makes it less likely that they will respond with empathy in future interactions (Timmers et al., 2018). Not surprising, hate speech can also impact helping behavior. For instance, Ziegele et al. (2018) showed that reading hate speech reduced readers' pro-social intentions towards the attacked group. In summary, hate speech does jar the foundations of the democratic contract (Papacharissi, 2004) by denying human equality.

4 Interindividual differences and motivations for hate speech

Not all people are equally likely to spread hate. Interindividual differences in personality traits, attitudes, and emotions are all associated with a different likelihood of becoming a hate speech perpetrator. With regards to personality traits, different studies have indicated an association between incivility and the so-called *dark tetrad* (Međedović & Petrović, 2015). The dark tetrad describes four sub-clinical forms of offensive personalities, the so-called dark triad of *narcissism*, *Machiavellianism*, and *psychopathy* (Paulhus & Williams, 2002) plus everyday *sadism* (Buckels et al., 2013). Narcissists are characterized by grandiosity perceptions (Paulhus & Williams, 2002)—although their self-esteem can be brittle at the same time (Miller et al., 2017)—and social manipulativeness (Raskin & Hall, 1981). Machiavellianism involves manipulative and cold behavior, psychopathy describes impulsive and thrill-seeking behavior by individuals showing reduced levels of anxiety (Paulhus & Williams, 2002), and sadism describes the enjoyment of cruelty and others' harm (Buckels et al., 2013). People scoring higher on the dark triad are more likely to admit to engaging in uncivil (Frischlich et al., 2021) and aggressive online behavior (Buckels et al., 2014; Kurek et al., 2019), although the direct link to hate speech is unclear (Koban et al., 2018). One component that could link the dark triad and uncivil and hateful speech is empathy. People scoring higher on the

dark tetrad tend to have deficits in their empathy abilities (e.g., see Heym et al., 2021), and empathic people engage less in trolling (March, 2019) and hate speech dissemination (Bilewicz & Soral, 2020).

Hate speech is also associated with people's generalized attitudes—that is, their ideological evaluation frameworks across situations, time, and/or persons. Two of these generalized attitudes are particularly relevant with regards to discrimination and prejudice (Duckitt et al., 2002; Duckitt & Sibley, 2010): Individual's level of *right-wing authoritarianism* (RWA*)* and their *social dominance orientation* (SDO).

Right-wing authoritarianism reflects a general psychological tendency to submit to authorities, support conventional values, and punish those who transgress the rules (Altemeyer, 1988; Duckitt, 2015). Social dominance orientation (Sidanius & Pratto, 1999) reflects a preference for group-based inequalities in society (Ho et al., 2015), either such that powerful groups should forcefully oppress lower status groups or in a more subtle hierarchy-enhancing way, for instance, by endorsing policies that stabilize group-based inequalities.

Bilewicz et al. (2017) showed that individuals with a larger social dominance orientation were particularly likely to consider hate speech to be acceptable—mirroring the idea that hate crimes are often an attempt to restore the presumably "natural order" (Perry & Alvi, 2012). Authoritarians, by contrast, were eager to prohibit hate speech expressions in a study by Bilewicz et al. (2017)—likely because the norm deviant character of hate speech conflicts with authoritarians' preference for adherence to established norms. Of note, our own research found that high authoritarians are more open to hateful right-wing extremist propaganda (Frischlich et al., 2015; Rieger et al., 2013, 2017), suggesting that further research into the interplay between authoritarianism, norm perceptions, and hate speech is needed.

Research also points toward ideological *asymmetries* regarding the association between political attitudes and hate speech. Survey data from the US showed that conservatives, compared to liberals, evaluated hate speech as being less disturbing (Costello et al., 2019), and research from Germany showed that supporters of the right-wing populist *Alternative for Germany* (AfD) were particularly active in supporting hate speech in online media (Frischlich et al., 2021; Kreißel et al., 2018). Hate speech is also a prominent communication style in alt-right online circuits (Marwick & Lewis, 2017). Although ideological asymmetries between those leaning toward the right versus toward the left have been demonstrated

for a wide array of human characteristics (Jost, 2017), one aspect could be particularly relevant regarding hate speech: differences in moral evaluations across the political spectrum. Based on *moral foundation theory* (Graham et al., 2011; Haidt & Joseph, 2008), humans have an intuitive ethic that has evolved to fulfil specific adaptive needs and whose violation is disregarded. Five of these moral foundations are particularly well established: The intuition of (a) *care*, evolved through the adaptive challenge to care for humans' vulnerable offspring but also the larger tribe, as research on *parochial altruism* suggests (Bernhard et al., 2006; Choi & Bowles, 2007). Following Graham et al. (2013), care is assumed to be related to empathic responses to others' suffering, and its violation is described as *harm*. The other four dimensions are (b) *fairness/cheating* (evolved as humans' response to interaction partners' lack of reciprocity in interactions); (c) *loyalty/betrayal* (related to humans' devotedness to their ingroup or tribe); (d) *authority/subversion* (reflecting the social order within the tribe); and (e) sanctity or *purity* versus *degradation*, reflecting disgust toward devaluated behaviors.

Individuals with different political orientations differ with regard to the relevance they ascribe to the violation and upholding of these five moral foundations. While liberals value individualizing moral intuitions of care and fairness particularly highly, conservatives also uphold biding moral intuitions of loyalty, authority, and purity (Graham et al., 2009). This difference is even reflected in the extreme case of terrorists' self-explanations: Hahn et al. (2019) showed that right-wing terrorists and religious fundamentalists justified their deeds more with binding moral values, whereas left-wing terrorists and those acting for animal rights relied more often on individualizing moral foundations. People who highly value the individual moral foundations of care and fairness are also more likely to report hate speech, whereas those valuing loyalty, authority, and purity are less likely to do so (Wilhelm et al., 2020).

Hate speech and other forms of incivility are also associated with different negative emotions. Following the *appraisal theories of emotion* (for a comprehensive overview, see Scherer, 2005), emotions can be understood as a process that ranges from (1) the cognitive appraisal of a specific internal or external stimulus over, (2) the psychophysiological response to that stimulus, (3) a verbal or non-verbal response, and (4) a motivational activation specific to the given emotion, up to (5) a distinct feeling such as joy, fear, or awe (e.g., Scherer, 1987). For instance, evaluating a situation as unjust and someone guilty of this injustice triggers anger (Nabi, 2002). Anger is associated with increased blood pressure (Lindquist et al., 2016),

the motivation to change the anger-inducing condition, and subjective feelings of being annoyed or in rage (Harmon-Jones & Harmon-Jones, 2016).

Research on political incivility has shown that people who respond with anger to incivilities write more uncivil comments in response (Gervais, 2017, 2019). Although emotions of anger partially overlap with hate, Bilewicz et al. (2017) argued that most of the phenomena labeled as hate speech are actually driven by emotions of disgust and contempt rather than anger and hate. Both hate and contempt increase the willingness to harm the target; however, contempt is associated with perceiving the target as inferior, whereas hate often targets seemingly powerful targets. Consequentially, contempt is often a better predictor of hate speech than hate or anger (for an overview, see Bilewicz et al., 2017).

It is likely that the different personality, attitudinal, and emotional variables lead to different types of haters, as a study by Erjavec and Kovačič (2012) shows. Based on a series of interviews, the authors identified four distinct types. The first two types tap into the social identity and social dominance components of hate speech: (1) "the soldier" (p. 909), who is described as an active member of a political party or (nationalist) organization who engages in organized hate speech as part of a "contemporary war" (p. 909), and (2) "the believer[s]" (p. 911), who has a similar worldview but lacks the organizational affiliation. The third type, (3) the "player," is someone who derives pleasure from disturbing the discourse, implying that dark personality traits might play a role here. Lastly, (4), the "watch-dog[s]" uses hate speech to draw attention to what is perceived to be unjust, seemingly underlining the role of morality.

5 Equality and empathy against hate and harm

In summary, hate speech can be conceptualized as harmful intergroup communication. This harm is particularly severe when coping resources are low, for instance, when stable social identities such as gender or ethnicity are under attack and/or for those belonging to socially underrepresented and marginalized groups. The suggested social psychological perspective provides a solid social scientific base for legal-rooted terms, such as hate crime or hate speech, and allows for describing the fallout of this specific type of attacks. Hate speech not only harms those directly attacked but also the entire social group; it jars social trust

and contributes to lasting social frictions by fueling prejudice, reducing prosocial behavior, and endangering empathy for fellow humans.

Although civility very often lies "in the eye of the beholder" (Herbst, 2010, p. 3), hate speech in the narrower sense described in this chapter is bound to specific socio-cultural spaces and norms, often reflecting traditional stereotypes and power imbalances in a society. The tendency of hate speech to attack those already deprived of coping resources, and the fact that these attacks fall out toward larger social groups, underlines the specific harms of hate speech. Although offensive speech can also be harmful, hate speech denies fellow humans their right to equity, thus crossing the borders of "reasonable disagreement" in a normative sense (Nussbaum, 2011, p. 4). Consequently, counter-measures such as the moderation of online content—which always need to strike the fragile balance between the freedom of expression and the preservation of a reasonable democratic discourse—might not only refer to individual harms when it comes to deciding about hate speech but can, psychologically speaking, also take the broader intergroup and societal context into account (for an excellent fusion of legal and social science perspectives, see Leets & Giles, 1999).

The psychology-rooted perspective of this chapter also demonstrates that not all people are equally likely to engage in hate speech. Dark personality traits characterized by empathy deficits, binding moral foundations that weigh loyalty, authority, and purity at least as highly as caring for others and fairness, convictions that society is, and should be, composed of inequal groups with different rights, and emotions of contempt all are associated with a larger propensity to spread hate speech.

This observation has meaningful implications for prevention: fostering empathy (Miklikowska, 2017) and creating unified super-ordinated social identities within a society (Dovidio et al., 2007) or with all humankind (McFarland, 2017) can help reduce stereotypes and prejudice. Social-dominance orientation can be a barrier to such endeavors (Sidanius et al., 2013); thus, it is also necessary to address the larger context in which social dominance orientation thrives. For instance, meta-analyses have shown that social dominance orientation is larger among individuals perceiving the world as a competitive struggle (Perry et al., 2013) and living in more hierarchical societies (Fischer et al., 2012). Taking the psychological factors on the micro-level of the individual hater as well as on the

meso-level of social groups into account can help to understand the roots and harms of hate speech, and find new ways to heal them.

Lena Frischlich is a junior research group leader at the University of Münster, Germany. https://orcid.org/0000-0001-5039-5301

References

Altemeyer, B. (1988). *Enemies of freedom: Understanding right-wing authoritarianism.* Jossey-Bass.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912–915. https://doi.org/10.1038/nature04981

Bilewicz, M., Kamińska, O. K., Winiewski, M., & Soral, W. (2017). From disgust to contempt-speech: The nature of contempt on the map of prejudicial emotions. *Behavioral and Brain Sciences*, *40*, 1–63. https://doi.org/10/gg5dxr

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(1), 3–33. https://doi.org/10.1111/pops.12670

Bilewicz, M., Soral, W., Marchlewska, M., & Winiewski, M. (2017). When authoritarians confront prejudice. Differential effects of SDO and RWA on support for hate-speech prohibition. *Political Psychology*, *38*(1), 87–99. https://doi.org/10/bfk8

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*(2), 307–324. https://doi.org/10/dg5fn8

Buckels, E. E., Jones, D. N., & Paulhus, D. L. (2013). Behavioral confirmation of everyday sadism. *Psychological Science*, *24*(11), 2201–2209. https://doi.org/10.1177/0956797613490749

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10/f58bzw

Carter, L. (2021, March 13). Finland's women-led government targeted by online harassment. *Politico*. https://www.politico.eu/article/sanna-marin-finland-online-harassment-women-government-targeted/

Chen, G., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, *5*(3), 205630511986264. https://doi.org/10/gghcnh

Choi, J.-K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Nature*, *318*(October), 636–640. https://doi.org/10.1126/science.1144237

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10/f6dxrx

College of Policing. (2020, October 20). Major investigation and public protection. *Responding to Hate*. https://www.app.college.police.uk/app-content/major-investigation-and-public-protection/hate-crime/responding-to-hate/#agreed-definitions

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. *Sociological Inquiry*, *89*(3), 427–452. https://doi.org/10/gghcnc

Crocker, J., & Schwartz, I. (1985). Prejudice and ingroup favoritism in a minimal intergroup situation: Effects of self-esteem. *Personality and Social Psychology Bulletin*, *11*(4), 379–386. https://doi.org/10.1177/0146167285114004

Dieckmann, J., Geschke, D., & Braune, I. (2018). *Für die Auseinandersetzung mit Diskriminierung ist die Betroffenen Perspektive von großer Bedeutung* [For dealing with discrimination the perspective of the victims is of high relevance]. Institut für Demokratie und Zivilgesellschaft. https://doi.org/10.19222/201702/4

Dovidio, J. F., Gaertner, S. L., & Saguy, T. (2007). Another view of "we": Majority and minority group perspectives on a common ingroup identity. *European Review of Social Psychology*, *18*(1), 296–330. https://doi.org/10/bwt4jb

Duckitt, J. (2015). Authoritarian personality. In J.D. Wright (Ed.). *International Encyclopedia of the Social& Behavioral Sciences* (2nd. Ed). Elsevier: https://doi.org/10.1016/B978-0-08-097086-8.24042-7

Duckitt, J., & Sibley, C. G. (2010). Personality, ideology, prejudice, and politics: A dual-process motivational model. *Journal of Personality*, *78*(6), 1861–1893. https://doi.org/10/dq9ptj

Duckitt, J., Wagner, C., du Plessis, I., & Birum, I. (2002). The psychological bases of ideology and prejudice: Testing a dual process model. *Journal of Personality and Social Psychology*, *83*(1), 75–93. https://doi.org/10/cgq3sx

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920. https://doi.org/10/gfgnmm

Festl, R. (2016). Perpetrators on the internet: Analyzing individual and structural explanation factors of cyberbullying in school context. *Computers in Human Behavior*, *59*, 237–248. https://doi.org/10/f8hw28

Fischer, R., Hanke, K., & Sibley, C. G. (2012). Cultural and institutional determinants of social dominance orientation: A cross-cultural meta-analysis of 27 societies. *Political Psychology*, *33*(4), 437–467. https://doi.org/10.1111/j.1467-9221.2012.00884.x

Frischlich, L., & Humprecht, E. (2021). *Trust, democratic resilience, and the infodemic*. Public Policy Institute.

Frischlich, L., Rieger, D., Hein, M., & Bente, G. (2015). Dying the right-way? Interest in and perceived persuasiveness of parochial extremist propaganda increases after mortality salience. *Frontiers in Psychology: Evolutionary Psychology and Neuroscience*, *6*(1222). https://doi.org/10/f7n3q8

Frischlich, L., Schatto-Eckrodt, T., Boberg, S., & Wintterlin, F. (2021). Roots of incivility: How personality, media use, and online experiences shape uncivil participation. *Media and Communication*, *9*(1), 195–208. https://doi.org/10.17645/mac.v9i1.3360

Gagliardone, I., Pohjonen, M., Zerai, A., Beyene, Z., Aynekulu, G., Bright, J., Bekalu, M. A., Seifu, M., Moges, M. A., Stremlau, N., Taflan, P., Gebrewolde, T. M., & Teferra, Z. M. (2016). *MECHACHAL: Online debates and elections in Ethiopia–From hate speech to engagement in social media*. Oxford University.

Gardiner, B., Mansfield, M., Anderson, I., Hoolder, J., Louter, D., & Ulumanu, M. (2016). The dark side of Guardian comments. *The Guardian*. https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments

Gelber, K., & Mcnamara, L. (2015). Evidencing the harms of hate speech. *Social Identities, 22*(3), 234–341. https://doi.org/10.1080/13504630.2015.1128810

Gervais, B. T. (2017). More than mimicry? The role of anger in uncivil reactions to elite political incivility. *International Journal of Public Opinion Research*, *29*(3), 384–405. https://doi.org/10/gftpws

Gervais, B. T. (2019). Rousing the partisan combatant: Elite incivility, anger, and antideliberative attitudes. *Political Psychology*, *40*(3), 637–655. https://doi.org/10/gghcns

Geschke, D., Klaßen, A., Quent, M., & Richter, C. (2019). *Hass im Netz–Der schleichende Angriff auf unsere Demokratie [Hate on the net – the creeping attack on our democracy]*. Institut für Demokratie und Zivilgesellschaft.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55–130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10/fhfs36

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10/cq64hc

Hahn, L., Tamborini, R., Novotny, E., Grall, C., & Klebig, B. (2019). Applying moral foundations theory to identify terrorist group motivations. *Political Psychology*, *40*(3), 507–522. https://doi.org/10.1111/pops.12525

Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind Volume 3: Foundations and the future* (pp. 367–391). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195332834.003.0019

Harmon-Jones, E., & Harmon-Jones, C. (2016). Anger. In L. Feldmann Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed., pp. 774–791). The Guilford Press.

Haslam, C., Cruwys, T., Haslam, S. A., Dingle, G., & Chang, M. X. L. (2016). Groups 4 Health: Evidence that a social-identity intervention that builds and strengthens social group membership improves mental health. *Journal of Affective Disorders*, *194*, 188–195. https://doi.org/10/gf3hcv

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.

Heym, N., Kibowski, F., Bloxsom, C. A. J., Blanchard, A., Harper, A., Wallace, L., Firth, J., & Sumich, A. (2021). The dark empath: Characterising dark traits in the presence of empathy. *Personality and Individual Differences*, *169*, 110172. https://doi.org/10.1016/j.paid.2020.110172

Ho, A. K., Sidanius, J., Kteily, K., Jennifer, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO7 scale. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Hogg, M. A., Sherman, D. K., Dierselhuis, J., Maitner, A. T., & Moffitt, G. (2007). Uncertainty, entitativity, and group identification. *Journal of Experimental Social Psychology*, *43*(1), 135–142. https://doi.org/10.1016/j.jesp.2005.12.008

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research*, *41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Jetten, J., Haslam, A. S., & Haslam, C. (2012). The case for a social identity analysis of health and well-being. In J. Jetten, C. Haslam, & S. A. Haslam (Eds.), *The Social Cure: Identity, Health and Well-being* (pp. 3–21). Psychology Press.

Jonas, E., McGregor, I., Klackl, J., Agroskin, D., Fritsche, I., Holbrook, C., Nash, K., Proulx, T., & Quirin, M. (2014). Threat and defense: From anxiety to approach. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 49, pp. 219–286). Elsevier. https://doi.org/10.1016/B978-0-12-800052-6.00004-4

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political Psychology*, *38*(2), 167–208. https://doi.org/10/ggmpfr

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, *47*(6), 795–814. https://doi.org/10/gghcnf

Koban, K., Stein, J. P., Eckhardt, V., & Ohler, P. (2018). Quid pro quo in Web 2.0. Connecting personality traits and Facebook usage intensity to uncivil commenting intentions in public online discussions. *Computers in Human Behavior*, *79*, 9–18. https://doi.org/10/gf3gv4

Koval, P., Laham, S. M., Haslam, N., Brock, B., & Whelan, J. (2012). Our flaws are more human than yours: Ingroup bias in humanizing negative characteristics. *Personality and Social Psychology Bulletin*, *38*(3), 283–295. https://doi.org/10/bwt484

Kreißel, P., Ebner, J., Urban, A., & Guhl, J. (2018). Hass auf Knopfdruck– Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz [Hate on the press of the button: Right-wing extremist troll campaigns and the ecosystem of coordinated hate campaigns]. *Institute for Strategic Dialogue*, 28–28.

Kurek, A., Jose, P. E., & Stuart, J. (2019). 'I did it for the LULZ': How the dark personality predicts online disinhibition and aggressive online behavior in adolescence. *Computers in Human Behavior, 98*, 31–40. https://doi.org/10/ggft9d

Leets, L., & Giles, H. (1999). Harmful speech in intergroup encounters: An organizational framework for communication research. *Annals of the International Communication Association*, *22*(1), 91–137. https://doi.org/10.1080/23808985.1999.11678960

Lindquist, K. A., Gendron, M., & Satpute, A. B. (2016). Language and emotion putting words into feelings and feelings into words. In L. Feldmann Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed., pp. 579–594). The Guilford Press.

Ma, Y., Wang, C., & Han, S. (2011). Neural responses to perceived pain in others predict real-life monetary donations in different socioeconomic contexts. *NeuroImage*, *57*(3), 1273–1280. https://doi.org/10.1016/j.neuroimage.2011.05.003

March, E. (2019). Psychopathy, sadism, empathy, and the motivation to cause harm: New evidence confirms malevolent nature of the Internet Troll. *Personality and Individual Differences*, *141*, 133–137. https://doi.org/10.1016/j.paid.2019.01.001

Martinovic, B., Jetten, J., Smeekes, A., & Verkuyten, M. (2017). Collective memory of a dissolved country: Collective nostalgia and guilt as predictors of interethnic relations between diaspora groups from former Yugoslavia. *Journal of Social and Political Psychology*, 588–607. https://doi.org/10/gf3gzb

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. Data & Society Research Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online/

Mason, G. (2005). Hate crime and the image of the stranger. *The British Journal of Criminology*, *45*(6), 837–859. https://doi.org/10.1093/bjc/azi016

McFarland, S. (2017). Identification with all humanity: The antithesis of prejudice, and more. In C. G. Sibley & F. K. Barlow (Eds.). *The Cambridge handbook of the psychology of prejudice* (pp. 632–654). Cambridge University Press. https://doi.org/10.1017/9781316161579.028

Međedović, J., & Petrović, B. (2015). The dark tetrad: Structural properties and location in the personality space. *Journal of Individual Differences*, *36*(4), 228–236. https://doi.org/10.1027/1614-0001/a000179

Miklikowska, M. (2017). Empathy trumps prejudice: The longitudinal relation between empathy and anti-immigrant attitudes in adolescence. *Developmental Psychology*, *54*(4), 703. https://doi.org/10.1037/dev0000474

Miller, J. D., Lynam, D. R., Hyatt, C. S., & Campbell, W. K. (2017). Controversies in narcissism. *Annual Review of Clinical Psychology*, *13*(1), 291–315. https://doi.org/10/gghcm9

Muldoon, O. T., Haslam, S. A., Haslam, C., Cruwys, T., Kearns, M., & Jetten, J. (2019). The social psychology of responses to trauma: Social identity pathways associated with divergent traumatic responses. *European Review of Social Psychology*, *30*(1), 311–348. https://doi.org/10/gg4whk

Mutz, D. C. (2015). *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

Nabi, R. L. (2002). Anger, fear, uncertainty, and attitudes: A test of the cognitive-functional model. *Communication Monographs*, *69*(3), 204–216. https://doi.org/10/dchzh4

Näsi, M., Räsänen, P., Hawdon, J., Holkeri, E., & Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. *Information Technology & People*, *28*(3), 607–622. https://doi.org/10.1108/ITP-09-2014-0198

Nussbaum, M. C. (2011). Perfectionist liberalism and political liberalism. *Philosophy & Public Affairs*, *39*(1), 3–45. https://doi.org/10.1111/j.1088-4963.2011.01200.x

O'Sullivan, P. B., & Flanagin, A. J. (2003). Reconceptualizing "flaming" and other problematic messages. *New Media & Society*, *5*(2), 69–94. https://doi.org/10/b3txz4

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10/dz4rp6

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10/d2jxm9

Perry, B., & Alvi, S. (2012). "We are all vulnerable": The in terrorem effects of hate crimes. *International Review of Victimology*, *18*(1), 57–71. https://doi.org/10/fxsq7s

Perry, R., Sibley, C. G., & Duckitt, J. (2013). Dangerous and competitive worldviews: A meta-analysis of their associations with social dominance orientation and right-wing authoritarianism. *Journal of Research in Personality*, *47*(1), 116–127. https://doi.org/10/tvc

Pfundmair, M., & Wetherell, G. (2019). Ostracism drives group moralization and extreme group behavior. *The Journal of Social Psychology*, *159*(5), 518–530. https://doi.org/10/ggbhgv

Phillips, W. M. (2012). *This is why we can't have nice things: The origins, evolution and cultural embeddedness of online trolling*. ProQuest Dissertations Publishing.

Raskin, R., & Hall, C. S. (1981). The narcissistic personality inventory: Alternate form reliability and further evidence of construct validity. *Journal of Personality Assessment*, *45*(2), 159–162. https://doi.org/10/ch4b6b

Ratner, K. G., & Amodio, D. M. (2013). Seeing "us vs. them": Minimal group effects on the neural encoding of faces. *Journal of Experimental Social Psychology*, *49*(2), 298–301. https://doi.org/10/f4rvcr

Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., Oksanen, A., Räsänen, P., & Zych, I. (2020). Hate knows no boundaries: Online hate in six nations. *Deviant Behavior*, advanced online publication, https://doi.org/10.1080/01639625.2020.1722337

Rieger, D., Frischlich, L., & Bente, G. (2013). *Propaganda 2.0: Psychological effects of right-wing and Islamic extremist internet videos* (Vol. 44). Wolters Kluwer Deutschland.

Rieger, D., Frischlich, L., & Bente, G. (2017). Propaganda in an insecure, unstructured world: How psychological uncertainty and authoritarian attitudes shape the evaluation of right-wing extremist internet propaganda. *Journal for Deradicalization*, *10,* 203–229.

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, advanced online publication. https://doi.org/10.1177/0093650220921314

Scherer, K. R. (1987). *Toward a dynamic theory of emotion: The component process model of affective states* (Geneva Studies in Emotion, pp. 1–96).

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10/fwmgvv

Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Ho, A. K., Sibley, C., & Duriez, B. (2013). You're inferior and not worth our concern: The interface between empathy and social dominance orientation. *Journal of Personality*, *81*(3), 313–323. https://doi.org/10.1111/jopy.12008

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression.* Cambridge University Press.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *Cornell University Library*, *June*. http://arxiv.org/abs/1603.07709

Smith, E. R., & Mackie, D. M. (2015). Dynamics of group-based emotions: Insights from intergroup emotions theory. *Emotion Review*, *7*(4), 349–354. https://doi.org/10.1177/1754073915590614

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136-146. https://doi.org/10/gf3gx2

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In S. Worchel & W. G. Austin (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–47). Brooks-Cole. https://doi.org/10.1016/S0065-2601(05)37005-5

Timmers, I., Park, A. L., Fischer, M. D., Kronman, C. A., Heathcote, L. C., Hernandez, J. M., & Simons, L. E. (2018). Is empathy for pain unique in its neural correlates? A meta-analysis of neuroimaging studies of empathy. *Frontiers in Behavioral Neuroscience*, *12*. https://doi.org/10.3389/fnbeh.2018.00289

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory.* Basil Blackwell.

Uhlmann, E. L., Korniychuk, A., & Obloj, T. (2018). Initial prejudices create cross-generational intergroup mistrust. *Plos One*, *13*(4), e0194871. https://doi.org/10.1371/journal.pone.0194871

Walters, M. A., Brown, R., & Wiedlitzka, S. (2016). Causes and motivations of hate crime. *Equality and Human Rights Commission Research Report*, *102*, 61–61.

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research*, *47*(6), 921–944. https://doi.org/10.1177/0093650219855330

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly*, *63*(2), 116–132. https://doi.org/10.2307/2695887

Ziegele, M., Koehler, C., & Weber, M. (2018). Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting & Electronic Media*, *62*(4), 636–653. https://doi.org/10/gf8pn4

Digital
Communication
Research.de

Recommended citation: Benesch, S. (2023). Dangerous speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 185–197). Digital Communication Research. https://doi.org/10.48541/dcr.v12.11

Abstract: The concept of "dangerous speech," which I proposed in the early 2010s, illuminates a key fact that is often missed: hate speech (and related categories like toxic and extreme speech) affects people gradually, cumulatively, and by dint of repetition. Dangerous speech is defined based on the specific harm it engenders (inspiring intergroup violence) rather than its content alone or the intent of those who spread it, allowing for a more consistent definition and broader consensus that it should be addressed. In this article, I explain why this concept is useful; describe the five aspects of speech that must be evaluated in order to determine dangerousness; share examples of projects that have been conducted to monitor, evaluate, and counteract dangerous speech; and suggest future avenues for research.

DOI 10.48541/dcr.v12.11

Susan Benesch

Dangerous Speech[1]

1 Dangerous speech as a practical tool for research

Two facts about hate speech (and related categories, including toxic, extreme, and dangerous speech) are vital for understanding its effects, and should inform research. The first is that hatred is not innate; it must be taught. Hate speech should be conceptualized as a tool for doing that, in order to learn how to prevent or reverse the teaching process.

The second fact is that hate speech affects people gradually, cumulatively, and by dint of repetition.[2] This point is often neglected by scholars and content moderation policymakers who work on hate speech, but it is well known to those who suffer from it. One of the latter, a witness in a trial at the United Nations Tribunal for Rwanda, after the 1994 genocide, described it brilliantly. Explaining how a radio station had groomed its listeners to commit and condone unthinkable violence, the witness said, "In fact, what RTLM [Radio Télévision Libre des Mille Collines] did was almost to pour petrol, to spread petrol throughout the country

1 The author is very grateful to her colleague Tonei Glavinic for contributing invaluable research and editing to this chapter.

2 See e.g., Hasher et al. (1977), which coined the term "illusory truth effect" to describe the phenomenon of people coming to believe a falsehood after hearing it repeatedly stated as fact.

little by little, so that one day it would be able to set fire to the whole country" (Prosecutor v. Nahimana, 2003, para. 436).

The concept of "dangerous speech" (Dangerous Speech Project [DSP], 2021), which I proposed, incorporates this phenomenon of gradual norm change, allowing for study that more clearly depicts human experience than hate speech and similar categories, and permitting more sensitive monitoring for increased risk of violence. Dangerousness, in this formulation, is the capacity of a speech act[3] – as disseminated—to inspire violence against members of another human group. Dangerous speech is defined by the specific harm[4] it engenders, not by its content alone, nor by the intent or motives of the people who produce and spread it. This makes for more consistent definition, and more consensus against this kind of speech, since it is very difficult for people to agree on which content is offensive, but easy for them to agree that mass intergroup violence should be prevented.

To capture the variable impact of speech acts, or "drops of petrol," dangerous speech is not a binary concept; dangerousness falls on a spectrum. Speech (including words, sounds, and images) can be more or less dangerous, depending on characteristics including the means by which it was disseminated, the speaker or source, the audience, the message itself, and the social and historical context in which the speaker and audience find themselves. The social context includes previous dangerous messages, which slowly shift people's states of mind so that they become more susceptible to the next such message, which is therefore more dangerous.

It is also important to note that repeated exposure to dangerous speech can convince members of an audience that such ideas are widely accepted by the people around them, even if they do not believe or accept the messages themselves. In other words, dangerous speech can shift norms, and people eagerly comply with norms to maintain good standing in a group (Leader Maynard, 2014).

3 In language theory a "speech act" is any form of communication that brings about some sort of response or change in the world. The 20th-century British philosopher of language J. L. Austin (1962) pioneered speech act theory, in which he tried to capture and distinguish all the types of effects that language can have. "Perlocutionary force," Austin proposed, is the capacity of a speech act to provoke a response in its audience. Dangerous speech is defined by such force: its capacity to inspire violence.

4 For more on the wide variety of harms speech can engender, and an argument that for robust research and policymaking, it is important to categorize speech by harms, not only by content, see Benesch (2020).

2 Dangerous speech and hate speech

Dangerous speech is a narrower and more precisely bounded category than hate speech, the most prevalent term in academic literature and common discourse (see also Sponholz and Frischlich in this volume). Although some hate speech is explicit and all too easy to identify as such, as a category it is large and contested, with blurry boundaries. We lack consensus on how to define it in law,[5] scholarly literature, common parlance, and even in the rules under which internet companies prohibit some content—and permit the rest.[6]

The term hate speech itself presents important questions that have not yet been consistently answered. First, must hate speech express hatred, promote hatred, or make someone feel hated? For example, if asked whether a drawing of the Prophet Mohammed constitutes hate speech, should one consider the intention of the person who made the drawing, or of someone else who disseminated it, or its effect on some or all of the people who see it or hear about it?[7] If it is the intention of the author that is definitive, the state of another person's mind is not always easy to discover, especially when its expression is found online.

Moreover, if hate speech is related to hatred, what exactly is that? How strong or how durable must emotion be to count as hatred?

One point that is clear, paradoxically, is that "I hate you," no matter how vehemently or sincerely expressed, is generally not hate speech (European Commission, 2018, p. 2), since a common thread among definitions is that hate speech denigrates or attacks a person or people *due to some characteristic or identity that they share* with other people, such as race, religion, nationality, sexual orientation, gender, age, caste, immigrant status, or disability. Most definitions list some but not all of these characteristics, which has generated disagreement over which kinds of groups *count* as targets of hate speech. The United Nations wisely avoided

5 For details on the variety of definitions for hate speech, see Benesch (2014, p. 20); also Herz & Molnar (2012, p. 81).

6 See e.g., Facebook (2021); Google (2021); Twitter (2021).

7 For key relevant ideas, e.g., on the distinction between giving offense and taking offense, see George (2016). For description of the overlooked role of 'malevolent bridge figures,' or people who transmit content from one normative community in which it is offensive or controversial, to another in which it is highly inflammatory, see Benesch (2015).

this problem in a new definition of hate speech that it introduced in May 2019, by giving a non-exhaustive list of group characteristics—"religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" (2019, p. 2). Unfortunately the same definition is vague and overbroad in another way, by describing hate speech as "pejorative" language with no explanation or limitation of that term. The full UN definition is this:

> Any kind of communication in speech, writing, or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or group on the basis or who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor (United Nations, 2019, p. 2).

Definitional problems arise with other categories of speech as well. What may be called extremist content, for example, can be thus identified because it was produced by extremists, or because it depicts and endorses gore, or because—to the contrary—it is designed to recruit for extremist groups, by falsely describing a safe and satisfying life within them or simply by criticizing life outside those groups in ways that are compelling and convincing to certain audiences, such as lonely, frustrated youth. Though it might not be wrong to label all such content "extremist," it would be a mistake to use the same method to try to identify, study, or protect people from all of it.

3 Defining and identifying dangerous speech

I coined the term "dangerous speech" after noticing patterns in public speech during the months and years before mass violence happened in many parts of the world and in many historical periods. Political, religious, and cultural leaders' language tends to be similar during such times, from case to case, country to country, culture to culture, religion to religion, even from one historical period to another. If people could learn to identify the hallmarks of speech that seems to increase the risk of violence, then perhaps one could decrease that kind of violence. That is the most powerful reason for doing research on this kind of speech, and in my view it is reason enough.

Rhetoric alone cannot make speech dangerous, though; the context in which it is communicated is equally important. One can systematically capture many

features of that context, and analyze speech for dangerousness, by asking about five aspects of the speech as it was disseminated. This analysis can capture the cumulative effect of speech mentioned above, and the fact that dangerousness is relative; speech can be more or less dangerous. Here we mention the five aspects, with examples of the questions one would ask regarding each of them:

- Speaker: Did the message come from an influential speaker? What is the source of that influence: for example cultural, social, or religious status; public office; access to troops or another means of threatening force; charisma?
- Audience: Was the audience susceptible to inflammatory messages, e.g., because they were already fearful? Have they already been exposed to similar messages over time, like the drops of petrol described by the Rwandan witness?
- Message: Does the speech carry hallmarks of dangerous speech? These are the rhetorical patterns that my colleagues and I have identified in many examples of public speech before outbreaks of violence (DSP, 2021). A hallmark is not sufficient to identify dangerous speech on its own, since speech cannot be dangerous if an audience is not moved by it, and some audiences are resistant, fortunately. The hallmarks we have identified thus far include (DSP, 2021, pp. 12-19):
  - *Dehumanization*. Describing other people in ways that deny or diminish their humanity, for example by comparing them to disgusting or deadly animals, insects, bacteria, or demons. This makes violence seem acceptable.
  - *"Accusation in a mirror."* Asserting that the audience faces serious and often mortal threats from the target group—in other words, reversing reality by suggesting that the victims of, e.g., a genocide will instead commit it. The term "accusation in a mirror" was found in a mimeographed guide for making propaganda, discovered in Rwanda after the 1994 genocide (Des Forges, 1999, pp. 65-66). Accusation in a mirror makes violence seem necessary by convincing people that they face a mortal threat, which they can fend off only with violence. This is a very powerful rhetorical move since it is the collective analogue of the one ironclad defense to murder: self-defense. If people feel violence is necessary for defending themselves, their group, and especially their children, it seems not only justified but virtuous.

- *Assertion of attack on women/girls.* Suggesting that women or girls of the audience's group have been – or will be – threatened, harassed, or defiled by members of a target group. In many cases, the purity of a group's women is symbolic of the purity of the group itself, or of its identity or way of life.
- *Coded language.* Including phrases and words that have a special meaning, shared by the speaker and audience. The speaker is therefore capable of communicating two messages, one understood by those with knowledge of the coded language and one understood by everyone else. This can make the speech more dangerous in a few ways. For example, the coded language could be deeply rooted in the audience members' sense of identity or shared history and therefore evoke disdain for an opposing group. It can also make the people who use the term feel that they are more strongly bound to the group of other people who use the code, like a password or a special handshake. Finally, coded speech can be harder to identify and counter for those who are not familiar with it, including social media company staff. One example of coded language – or symbols – is the use of a pineapple to mock and denigrate Jews in France (Nelson, 2015). Another is the name of the mobile phone company "MTN" which has been used as a powerful slur against Dinka people in South Sudan because of the company's slogan "Everywhere you go," understood as a coded reference to the claim that the Dinka invaded the lands of other groups (Patinkin, 2017).
- *Impurity/contamination.* Giving the impression that one or more members of a target group might damage the purity or integrity or cleanliness of the audience group. Members of target groups have been compared to rotten apples that can spoil a whole barrel of good apples, weeds that threaten crops, or stains on a dress, for example.

- Context: Is there a social or historical context that has lowered the barriers to violence or made it more acceptable? Examples of this are competition between groups for resources and previous episodes of violence between the relevant groups.
- Medium: How influential is the medium by which the message is delivered? For example, is it the only or primary source of news for the relevant audience?

All five conditions need not be relevant for speech to be dangerous. For example, a message can be dangerous even when the speaker is anonymous. Only two conditions are necessary: the message must be inflammatory, and the audience must be susceptible.

4 Studying dangerous speech and efforts to counter it

The idea of dangerous speech can be useful for research of several kinds, of which the first is to collect and analyze examples of it, as a way of understanding whether and where violence may occur, and to inform violence prevention efforts.

This type of monitoring began in Kenya, leading up to that country's 2013 national elections—a tense moment since the previous national election campaign brought months of dangerous speech and was followed by mass violence in 2007, in which more than 1,200 people were killed and more than 600,000 were displaced from their homes. I worked with Ushahidi and iHub Research on their Umati project[8], in which teams of full-time monitors manually collected examples of vitriolic speech in six different languages. Their codebook built upon the contextual factors of dangerous speech outlined above by directing coders to consider each of the five factors for each speech act and then classify it as offensive, moderately dangerous, or (very) dangerous (Awori, 2013; DSP, 2016). The Umati codebook distilled that process for coders, and guided them through it, by asking them two questions about the speaker and about the content itself: "On a scale of 1 to 3 with 1 being little influence and 3 being a lot of influence, how much influence does the speaker have on the audience?" and "On a scale of 1 to 3, with 1 being barely inflammatory and 3 being extremely inflammatory, how inflammatory is the content of the text?" This method was inventive, and it may have increase inter-rater reliability, but it was of limited use when the speaker was unknown, which is quite often the case for online speech.

Coding questions arose frequently, and the team held regular meetings to consider and resolve them. The meetings were lessons in how varied hateful and inflammatory speech is, and how important context can be, for understanding it. In one example, a coder identified the sentence "I hate Raila" (Odinga, one of the

8 Umati means "crowd" in the Kenyan national language of Kiswahili.

leading presidential candidates) as dangerous speech. I said this was neither hate speech nor dangerous speech, since it was directed only at an individual, without reference to any group. The coder replied unequivocally that in the Kenyan context of that time, to say "I hate Raila" was also to say that the speaker hated Luos, the ethnic group of which Odinga was a leader.

In 2015, the Nigerian Centre for Information Technology and Development (CITAD, 2016) monitored online speech during and after Nigeria's 2015 election campaign, building on the Umati model.

Another promising body of research related to dangerous speech is studies on efforts to counter its harmful effects, focusing on whether and how they succeed. This is nascent, since only a few such projects have been carried out, so far without being rigorously studied. For example, Umati led to two efforts to counter dangerous speech during Kenya's national electoral campaign of 2013, by "inoculating" the public against such speech, i.e., teaching people that it is a tool used by unscrupulous leaders to manipulate them. One of those was studied. The first effort was called Nipe Ukweli, or "gimme truth" in Kenyan slang—a name reflecting the fact that much dangerous speech is also disinformation (DSP, 2016). This project consisted of flyers and community meetings that explained dangerous speech and encouraged people to report it to the Umati team. In the second effort, four episodes of a legendary, well-known Kenyan television courtroom drama called Vioja Mahakamani[9] focused on dangerous speech with a similar goal: to teach the audience to recognize it and to be skeptical of it. The Vioja Mahakamani project was independently evaluated for impact on audiences, by researchers from the University of Pennsylvania, with encouraging results. Focus groups conducted with young Kenyans from various ethnic backgrounds revealed that those who watched the dangerous speech episodes demonstrated a greater understanding of the origins, motivations, and consequences of incitement compared to those in control groups who watched unrelated episodes of the show (Kogen, 2013). More such studies are clearly needed, though they can be difficult to conduct, either because there are confounding third factors that make robust evaluation difficult, or because researchers cannot get sufficient data.

9 Vioja Mahakamani means "events in the courtroom." The four episodes were collectively designed by the actors who made them, after they all attended a workshop on dangerous speech.

Technology companies are one important type of stakeholder for such research, of course, since they have enormous power and capacity to experiment with methods to improve online behavior and norms. In response to public pressure and legal requirements, especially those with social media platforms, tech companies are increasingly trying out new techniques to try to more effectively identify and deal with hate speech, dangerous speech, and other harmful content on their online turf. While removing such content is the most visible remedy, it is a heavy-handed approach, and there are many alternatives that better protect freedom of expression but need to be better understood, including downranking content (reducing its algorithmic amplification), "nudging" users to reconsider their words before they are posted (Diaz, 2021), and proactively reminding users of the rules they must follow (Benesch & Matias, 2018). Yet it is rare that these efforts are A/B tested to see which is more effective—and virtually all research at companies is under nondisclosure agreements. Conducting independent, ethical, transparent, privacy-protecting research—either in cooperation with companies or in spite of them—and publishing it in reputable peer-reviewed journals would be a major step toward greater understanding of how to meaningfully address harmful online content.

There are also significant efforts in civil society to tackle hate speech and dangerous speech online, presenting many research opportunities that have not yet been seized. Anthropologist Cathy Buerger (2020) published the first in-depth, ethnographic exploration of #jagärhär, a thousands-strong network of volunteers in Sweden who launch coordinated responses to hate speech and dangerous speech on Facebook. That project is unusual not only for its large size (more than 70,000 people are members of the Swedish group alone) but for the fact that it is still going strong several years after its founding, and also for its replication in other countries (there are 16 groups operating in various countries, at this writing). Buerger (2020) interviewed 25 of the most active Swedish participants, many of whom said they observed favorable shifts in discourse norms in the spaces where they have intervened online.

We have also identified dozens of smaller anti-hate efforts in many countries. Activists, journalists, clergy, lawyers and others have been experimenting with quite a variety of methods, including some that deliberately amplify hateful or offensive content to force members of a society to accept that it is there and reckon with the racism and hatred it expresses. Of course technology plays a role in many of these efforts: just as new communications technologies are being used to amplify

inflammatory hate speech, they can also be marshalled to prevent and counter it. New technologies are also being employed to detect where dangerous speech may signal an increased risk of mass violence, and social media companies sometimes delete such content, or downrank it, as noted above (Facebook, 2020, p. 7).

Companies have so far missed other opportunities to detect dangerous speech, such as observing the way in which members of the public respond—in open online spaces—to the posts of unscrupulous leaders. This could give invaluable clues in many cases, without violating privacy or causing other harms. For example, in mid-December 2020, Donald Trump invited his followers to come to Washington, D.C. for a rally, on January 6, 2021. Though he wrote only that the rally "will be wild," many of his followers understood his ambiguous language as a call to violence, by telling each other that he wanted them to come with firearms, ready to use them. I have developed this idea in another article (Benesch, 2021).

In sum, dangerous speech is worth special attention from researchers for several reasons. First, it seems to be linked to intergroup violence, and therefore it may serve as a good early warning signal. Perhaps violence can be prevented, at least in part, if dangerous speech can be defanged or diminished without causing other harms (like infringing on freedom of expression). Second, dangerous speech is a more precise and less contested category than others like hate speech, so it should be possible to build comparable datasets of it from a variety of places or social groups. Transnational study is exceedingly rare in the literature on hate speech, and would be of great interest. Also, though the dangerousness of speech depends greatly on context, which cannot be detected and evaluated automatically, it may be possible to build classifiers for dangerous speech that operate by detecting similarities and patterns in it.

Finally, the concept of dangerous speech accommodates the fact that inciting language has a cumulative effect on people. This is key to understanding the capacity of speech to inspire behavior, but it has so far received scant attention. I hope the literature will soon grow in these areas.

Susan Benesch is founder and director of the Dangerous Speech Project, Faculty Associate of the Berkman Klein Center for Internet & Society at Harvard University, and Adjunct Associate Professor at American University's School of International Service, USA.

References

Austin, J. L. (1962). *How to do things with words.* Oxford University Press.

Awori, K. (2013, June). *Umati final report*. iHub Research. https://dangerousspeech.org/wp-content/uploads/2017/05/Umati-Final-Report.pdf

Benesch, S. (2015). Charlie the freethinker: Religion, blasphemy, and decent controversy. *Religion & Human Rights, 10*(3), 244–254. https://doi.org/10.1163/18710328-12341291

Benesch, S. (2020, June). *Proposals for improved regulation of harmful online content*. Dangerous Speech Project. https://dangerousspeech.org/wp-content/uploads/2020/06/Proposals-for-Improved-Regulation-of-Harmful-Online-Content-Formatted-v5.2.1.pdf

Benesch, S. (2021, March 4). *The insidious creep of harmful rhetoric*. Noēma. https://www.noemamag.com/the-insidious-creep-of-violent-rhetoric/

Benesch, S., & Matias, J. N. (2018, April 6). *Launching today: new collaborative study to diminish abuse on Twitter*. Medium. https://medium.com/@susanbenesch/launching-today-new-collaborative-study-to-diminish-abuse-on-twitter-2b91837668cc

Buerger, C. (2020, December 14). *The anti-hate brigade: How a group of thousands responds collectively to online vitriol*. Dangerous Speech Project. https://dangerousspeech.org/anti-hate-brigade/

Centre for Information Technology and Development. (2016). Traders of hate in search of votes: Tracking dangerous speech in Nigeria's 2015 election campaign. http://www.citad.org/download/traders-of-hate-in-search-of-votes/?wpdmdl=2493

Dangerous Speech Project. (2016, November 1). *Monitoring and evaluating inflammatory speech in Kenya.* https://dangerousspeech.org/kenya/

Dangerous Speech Project. (2021, April 20). *Dangerous speech: A practical guide*. Dangerous Speech Project. https://dangerousspeech.org/guide

Des Forges, A. (1999). *Leave none to tell the story: Genocide in Rwanda*. Human Rights Watch.

Diaz, J. (2021, May 6). *Want to send a mean tweet? Twitter's new feature wants you to think again*. National Public Radio. https://www.npr.org/2021/05/06/994138707/want-to-send-a-mean-tweet-twitters-new-feature-wants-you-to-think-again

European Commission. (2018, January 19). *Countering illegal hate speech online* [Fact sheet]. https://ec.europa.eu/commission/presscorner/api/files/document/print/en/memo_18_262/MEMO_18_262_EN.pdf

Facebook (2020, May 12). *Facebook response: Sri Lanka human rights assessment*. https://about.fb.com/wp-content/uploads/2021/03/FB-Response-Sri-Lanka-HRIA.pdf

Facebook (2021). Community Standards. https://www.facebook.com/communitystandards/

George, C. (2016). *Hate spin: The manufacture of religious offense and its threat to democracy*. MIT Press.

Google. (2021). YouTube policies: Hate speech policy. https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=2803176

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Herz, M., & Molnar, P. (Eds.). (2012). Interview with Kenan Malik. In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: Rethinking regulation and responses* (pp. 81–91). Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.008

Kogen, L. (2013, December 13). Testing a media intervention in Kenya: Vioja Mahakamani, dangerous speech, and the Benesch guidelines. Center for Global Communications Studies, University of Pennsylvania. https://dangerousspeech.org/testing-a-media-intervention-in-kenya-vioja-mahakamani-dangerous-speech-and-the-benesch-guidelines/

Leader Maynard, J. (2014). Rethinking the role of ideology in mass atrocities. *Terrorism and Political Violence, 26*(5), 821–841. https://doi.org/10.1080/09546553.2013.796934

Nelson, L. (2015, January 14). The quenelle: France's notorious anti-Semitic hand gesture, explained. *Vox*. https://www.vox.com/2015/1/14/7548289/quenelle-dieudonne-antisemitism-france

Patinkin, J. (2017, January 16). *How to use Facebook and fake news to get people to murder each other*. Buzzfeed News. https://www.buzzfeednews.com/article/jasonpatinkin/how-to-get-people-to-murder-each-other-through-fake-news-and

The Prosecutor v. Ferdinand Nahimana, Jean-Bosco Barayagwiza, Hassan Ngeze (Trial Judgment). (2003) ICTR-99-52-T, International Criminal Tribunal for Rwanda (ICTR). https://ucr.irmct.org/LegalRef/CMSDocStore/Public/English/Judgement/NotIndexable/ICTR-99-52/MSC26797R0000541998.pdf

Twitter (2021). *The Twitter rules*. https://help.twitter.com/en/rules-and-policies/twitter-rules

United Nations (2019). *United Nations strategy and plan of action on hate speech*. https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

Digital
Communication
Research.de

Recommended citation: Bormann, M., & Ziegele, M. (2023). Incivility. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 199–217). Digital Communication Research. https://doi.org/10.48541/dcr.v12.12

Abstract: Incivility is considered a significant challenge for democratic discourse and has been the subject of many studies in a variety of contexts. Although political incivility has a long research tradition, and scholarly attention toward the phenomenon has increased with the advance of social media, there is academic controversy regarding the concept and normative implications of incivility in political contexts. This chapter provides an overview of different incivility approaches in the extant literature, discusses key challenges in incivility research, and outlines normative implications. Further, we suggest future directions for incivility research and argue why an integrative, multidimensional concept of incivility offers great potential for incivility research in the field of political (online) communication.

DOI 10.48541/dcr.v12.12

Marike Bormann & Marc Ziegele

Incivility

1 Incivility in political communication—an established yet elusive concept

Incivility has been studied in a variety of contexts, ranging from workplace environments (e.g., Schilpzand et al., 2016) to political contexts (e.g., Jamieson, 2000; Papacharissi, 2004). For this chapter, we focus on incivility in political communication. Incivility in public political discourse is a recurring subject of concern across different countries. Recently, various speakers have feared a decline or even a "crisis of civility" (Boatright et al., 2019). Polls have shown that 68% of Americans think that incivility in political communication is a major social issue. Moreover, most Americans have reported personal encounters with incivility (Weber Shandwick, 2020). Surveys among German online users reveal a similar picture, with 73% of users reporting that they have already been exposed to uncivil or hateful comments (LfM, 2020). Even the German federal president urgently called for more "reason and civility" (Steinmeier, 2019) in online discussions.

Political incivility, similar to the general phenomenon of incivility, has been the subject of many studies in a variety of contexts. These include, for example, incivility in political news articles, political campaigns, and advertising, and in political debates in Congress, television, and radio talk shows or interviews. Studies in this field usually analyze uncivil portrayals of politicians or incivility

in the interactions between political elites, such as politicians, journalists, and experts (e.g., Ben-Porath, 2010; Jamieson, 2000; Mutz & Reeves, 2005). Besides incivility among political elites, scholars have become increasingly interested in studying incivility in online discussions among ordinary citizens on social media platforms or on the websites of traditional news media. Online incivility research has yielded significant output, including findings on the causes, determinants, and patterns of incivility (e.g., Coe et al., 2014; Rossini, 2020), the perceptions of incivility (e.g., Stryker et al., 2016), the effects of incivility (e.g., Rösner et al., 2016), and interventions against incivility (e.g., Kalch & Naab, 2017; Ziegele, Jost et al., 2018).

Although political incivility has a long research tradition and academic attention to the phenomenon has increased with the advance of the Internet, there is academic controversy regarding the concept, theory, operationalization, and normative implications of incivility in political contexts. In the following section, we first provide an overview of different approaches to the phenomenon of political incivility in the extant literature and argue for an integrative, multidimensional concept. We then discuss the challenges of different approaches and outline the normative implications of incivility. Lastly, we argue why an integrative approach offers great potential for incivility research in the field of political (online) communication.

2 Concepts of political incivility

Incivility is a broad phenomenon that encompasses a wide spectrum of communication in offline and online contexts. Owing to its Latin word stem *civis* (citizen) and *civitas* (citizenship), which historically refer to the civic role and the order of the polity (Simpson, 1960), the concept of incivility and much research on incivility explicitly focus on the political sphere and public political communication.

Incivility has a long tradition of research, but scholars are still having trouble finding an agreed-upon conceptual definition and operationalization. Herbst (2010) noted that the decision of where to draw the line between civility and incivility lies "very much in the eye of the beholder" (p. 3). Similarly, Coe and colleagues (2014) stated that "incivility is a notoriously difficult term to define, because what strikes one person as uncivil might strike another person as perfectly appropriate" (p. 660).

Benson (2011) pointed out that civility and incivility "are always situational and contestable" (p. 22). Hence, defining incivility is challenging, and a variety of approaches to the phenomenon can be found. Nevertheless, most definitions—at least implicitly—share the notion that *incivility is a violation of norms*. The majority of scholars approach incivility as a violation of *respect norms*, *democratic norms*, or *politeness norms*. These studies usually refer to normative theories of democracy or politeness theories. Additionally, recent studies have conceptualized incivility as a violation of *multiple norms*. Although these different perspectives are not always entirely clear-cut, it is helpful to briefly outline them in the following sections before proposing a new approach that integrates the different perspectives.

2.1 Incivility as a violation of respect norms

Studies analyzing incivility as a *violation of (deliberative) respect norms* usually refer to normative theories of democracy, mostly deliberation theory. Deliberation theory sketches a public sphere accessible to everyone in which citizens debate matters of public interest in a reciprocal, rational, and respectful manner (Gastil, 2008; Gutmann & Thompson, 2004; Habermas, 1996). Within this framework, civility is understood as mutual respect between discussants. Thus, studies have often defined incivility as *disrespectful behavior in public discussions* toward other participants, the forum, or specific topics (e.g., Anderson et al., 2014; Coe et al., 2014; Gervais, 2014, 2015; Sobieraj & Berry, 2011). It is important to note that such disrespectful behavior differs from mere disagreement. Disagreement, if voiced respectfully, is an inevitable characteristic of discussions with political opponents and is beneficial for deliberation (Herbst, 2010; Stromer-Galley, 2007). From this perspective, only disagreement (or negativity) combined with disrespect constitutes incivility (e.g., Hwang et al., 2018). Despite partly overlapping definitions, studies analyzing incivility as a violation of respect norms vary regarding their operationalizations of incivility. These operationalizations range from *name-calling*, *emotional displays*, and *ideologically extremizing language* (Sobieraj & Berry, 2011) to *lying* (Coe et al., 2014) and the *use of conspiracy theories* (Gervais, 2014).

2.2 Incivility as a violation of liberal democratic norms

Many scholars have also approached incivility as a *violation of liberal democratic norms* (e.g., Kalch & Naab, 2017; Oz et al., 2018; Papacharissi, 2004; Rowe, 2015). These studies often refer to Papacharissi's (2004) distinction between impoliteness and incivility. According to Papacharissi (2004), many earlier concepts of incivility have, in fact, measured impoliteness, which is "etiquette-related" and something that is not undesirable per se, as "adherence to etiquette [...] frequently restricts conversation" (p. 260), especially in political discussions. The author argued that incivility goes further than impoliteness, threatens democratic norms, and has negative implications for democracy. Consequently, impoliteness and incivility are operationalized differently, with the latter focusing on *threats to democracy*, *threats to individual rights*, and *antagonistic stereotypes*, such as *racism* or *sexism* (Papacharissi, 2004). This approach has since been used by various researchers. Rossini (2020), for example, similarly argued that violations of politeness norms cannot be equated with violations of democratic norms, and that only violations of the latter would be detrimental to democracy. Violations of democratic norms in Rossini's operationalization include discriminatory expressions and threats to individual liberty rights or denial of political participation. Contrary to Papacharissi (2004), however, Rossini defined violations of interpersonal politeness or respect norms as *incivility*, and norm violations that pose a threat to democracy as *intolerance*. Here, we clearly observe some inconsistencies in contemporary concepts of incivility. The resulting challenges will be discussed in more detail below.

2.3 Incivility as a violation of interpersonal politeness norms

Similar to Rossini (2020), various studies have analyzed incivility as a violation of *interpersonal politeness norms* (e.g., Ben-Porath, 2010; Chen & Lu, 2017; Chen & Ng, 2017; Mutz, 2007, 2015; Mutz & Reeves, 2005). These studies draw on politeness theories that deal with the rules of interpersonal interaction in public spaces, such as *social norm approaches* (Fraser, 1990) or *face theory* (Brown & Levinson, 1987; Goffman, 1959). Social norm approaches often follow a Western understanding of etiquette; within this understanding, incivility is usually defined as a

violation of the social norms of politeness for a given culture (e.g., Ben-Porath, 2010; Mutz, 2007; Mutz & Reeves, 2005). Against the backdrop of face theory, researchers have also conceptualized incivility as a threat to people's positive face, which is the socially desired and constructed public identity that people act out during a communication process (e.g., Chen & Lu, 2017; Chen & Ng, 2017). According to these approaches, incivility manifests, among others, in *insults*, *name-calling*, *yelling* (or using capital letters to indicate yelling in online communication), *interruption*, *profanity*, and *vulgarity* (Ben-Porath, 2010; Chen & Lu, 2017; Chen & Ng, 2017; Mutz, 2007; Mutz & Reeves, 2005).

2.4 Incivility as a violation of multiple norms

Contemporary theorizing about incivility has shifted to a constructionist perspective, suggesting that incivility is "multifaceted, individual, and context specific" (Wang & Silva, 2018, p. 73). Consequently, current research often approaches incivility as *perceived violations of multiple norms*. Muddiman (2017), for example, derived from the perceptions of participants in two experiments a two-dimensional model of perceived incivility. In this model, "personal-level incivility" includes violations of interpersonal politeness norms, and "public-level incivility" includes violations of deliberative norms, such as *ideological extremity* and *lack of comity*. Chen (2017) also approached incivility as a perceptual continuum, with impoliteness being on the mild end and hate speech being on the harmful end of the continuum. In their extensive survey, Stryker et al. (2016) found that besides violations of politeness and democratic norms, participants perceived *deception* as a third dimension of incivility. This dimension includes *lies* as well as *misleading* and *exaggerating claims*, which can be considered violations of honesty norms.

2.5 Toward an integrative concept of political incivility

In our own research, we propose a new concept of political incivility that incorporates previous concepts into an integrative framework, while following a bottom-up approach from the perspective of communication participants (Bormann et al., 2021). Based on theories on cooperation, communication, and

norms (e.g., Grice, 1975; Lindenberg, 2015; Tomasello, 2008, 2009), we suggest five communication norms that individuals can disapprove of violating. The five communication norms build on the central aspects of communication, namely, the substantial aspect (content; information), the formal aspect (mode), the temporal aspect (process), the social aspect (actors; relation), and the spatial aspect (context; Bormann et al., 2021; Lasswell, 1948; Schaff, 1962). Violations of the five norms potentially constitute incivility. The *information norm* refers to the substance of the information provided in a discussion. It can be violated when, for example, participants lie, spread conspiracy theories, or communicate misleading, irrelevant information. The *modality norm* concerns the formal aspect of communication and can be violated when participants communicate ambiguously, for example, by using sarcasm. The *process norm* refers to the interconnectedness of contributions and can be violated when, for example, participants deviate from the topic of the discussion or refuse to be responsive. The *relation norm* expresses the expectation of participants to be respectful and polite; it can be violated when, for example, participants use name-calling, insults, or vulgarity. Lastly, the *political context norm* encompasses the normative expectations of participants in political discussions to consider essential liberal democratic principles in their contributions. This norm can be violated when, for example, participants threaten the rights of other individuals, question the democratic constitution, or incite violence against democratic governments or minority groups. In our concept, incivility occurs when participants disapprove of an act of communication as severely violating one or several of these five communication norms.

In summary, it becomes clear that political incivility is a multi-faceted and complex phenomenon. A common denominator of the existing concepts that we can identify is that incivility refers to violations of norms. Depending on the research tradition, these norms include deliberative norms of mutual respect, liberal-democratic norms, or norms derived from politeness research. We also proposed an attempt toward an integrative concept of incivility in political communication. This concept describes incivility as a perceived violation of one or several of five basic communication norms, namely, the information norm, the modality norm, the process norm, the relation norm, and the political context norm. In the following sections, we discuss the challenges and perspectives related to these different approaches to political incivility.

3 Challenges of research on political incivility

3.1 Challenges related to inconsistent definitions and measures

A major challenge in research on political incivility is related to the difficulty of comparing the findings of different studies. Content analyses of online discussions, for example, have reported varying shares of incivility in user comments, ranging from 3% to more than 50% (e.g., Rowe, 2015; Santana, 2014). Some of these variations are clearly due to the fact that studies have analyzed different platforms and topics, among others. Yet, the *operationalizations of incivility* also vary significantly from study to study; thus, different phenomena are studied under the same term. Coe et al. (2014), for example, found that 22% of the user comments posted on a newspaper's website contained incivility, which the authors operationalized as name-calling, vulgarity, aspersion, pejoratives, or lying accusations. Rowe (2015) operationalized these norm violations as impoliteness and found that 32% of the comments posted on a newspaper's Facebook site and 35% of the comments posted on the newspaper's website were impolite. Incivility in terms of the assignment of stereotypes and threats to democracy or individual's rights was only visible in 3% of the Facebook comments and in 6% of the website comments (Rowe, 2015). Similarly, Santana (2014) compared incivility in anonymous and non-anonymous news website comments. Applying a broad operationalization of incivility as personal attacks, threats, vulgarities, abusive, foul, or hateful language, assignment of stereotypes, epithets, ethnic slurs, and racist or bigoted speech, Santana found that up to 53% of the comments were uncivil.

What renders these diverging findings particularly problematic is that they suggest different normative and practical implications for governing online discussion spaces. While policymakers or journalists may conclude that incivility is not a pressing issue based on studies that report low shares of incivility, research that has reported otherwise may justify calls for strong interventions. Future research should thus invest in reaching agreed-upon standardized operationalizations of incivility to increase the comparability of findings and to provide more reliable assessments of the development of incivility over time.

Diverging operationalizations of uncivil behavior are also problematic in experimental research (e.g., Chen & Ng, 2017; Gervais, 2015; Kalch & Naab, 2017; Rösner et al., 2016). Some studies on the effects of incivility, for example, have

operationalized incivility as a unidimensional construct or as a "monolith" (see Masullo in this collection). These studies mingle different types of uncivil behavior, such as name-calling, vulgarity, histrionics, and lies. Consequently, the distinct effects of the different types of incivility cannot be assessed (e.g., Gervais, 2015; Rösner et al., 2016). Yet, the few studies that have investigated people's perceptions of different types of incivility suggest that participants evaluate each type differently in terms of severity (e.g., Muddiman, 2017; Stryker et al., 2016), and that different types of incivility have varying effects on people's behavioral intentions (e.g., Kalch & Naab, 2017). Distinct forms of uncivil behavior should therefore not be viewed and investigated as unidimensional in future studies (see also Masullo in this collection for a similar appeal).

3.2 Challenges related to the reliable measurement of incivility in content analyses

As previously mentioned, many studies on political incivility have applied content analyses to investigate the patterns, determinants, and potential consequences of uncivil communication (e.g., Coe et al., 2014; Rowe, 2015; Ziegele, Jost et al., 2018). For these analyses, it is often challenging to achieve satisfactory levels of reliability and external validity for the measures that are used. Some manifestations of incivility, such as name-calling, can easily be recognized by all coders. However, when it comes to more subtle, culture-specific, or context-specific norm violations, such as implicit stereotypes, coders regularly struggle to detect these forms of incivility reliably. Similarly, it is difficult to detect norm violations in online discussions that perpetrators intentionally camouflage to circumvent algorithms and word filters, for example.

Ross et al. (2018) demonstrated that even among researchers who are familiar with incivility-related concepts, there is sometimes low agreement on what should be classified as civil and uncivil. Particularly for subtle norm violations, the coders' individual perceptions, knowledge, and experiences impact whether they classify a speech act as uncivil. Human speech is a rich and complex phenomenon, and so are the potential manifestations of political incivility. Although many studies provide clear coding instructions for various types of incivility, it is challenging or even impossible to consider all or even the most possible manifestations of these types in a coding scheme. Some researchers tackle this problem by coding

only incivility that is measurable on the level of words. This, however, reduces the validity of incivility measures. The problem is no less urgent in automated analyses of political incivility. Previous studies have already applied dictionary-based approaches (e.g., Muddiman & Stroud, 2017) and machine learning (e.g., Su et al., 2018) to study online incivility. Similar to manual content analyses, these methods work best for explicit forms of incivility that are clearly expressed through the use of specific words, such as offensive language or extreme forms of hate speech (e.g., Davidson et al., 2017). Automatically detecting subtle or ambiguous forms of incivility, such as covert racism or sarcasm, is far more challenging, and many automated measures suffer from high rates of misclassification (Stoll et al., 2020).

In understanding incivility as a perceptual construct and accepting that even the work of professional coders in content analyses will be, to some extent, affected by individual biases, we can think about alternative or complementary ways to classify incivility in content analyses. For example, each contribution in online discussions could be checked to determine whether it was visibly disapproved of by other participants. Disapproval here can be expressed, among others, through a sanctioning reply comment. If a comment has been visibly disapproved, coders can analyze it regarding the specific type(s) of norm violations (Bormann et al., 2021). Although this procedure will certainly work only for a small fraction of uncivil contributions, it would account for the fact that incivility is often a matter of the perceptions of the people involved in the respective communication.

3.3 Challenges related to the normative implications of incivility

Normative implications of incivility are controversial among scholars. This can be partly explained by the fact that studies have reported different consequences of incivility. Experimental research, for example, has found various negative effects of being exposed to uncivil content: incivility in political talk shows can reduce viewers' trust in politics and politicians (Mutz & Reeves, 2005). Uncivil online discussions have been found to increase readers' opinion polarization (Anderson et al., 2014), stimulate negative emotions and aggressive cognitions (Gervais, 2015; Rösner et al., 2016), and promote further incivility (Gervais, 2015; Ziegele, Weber et al., 2018). Moreover, uncivil comments can adversely affect the perceived quality of news articles (Prochazka et al., 2018) and increase prejudice

against social minorities (Hsueh et al., 2015). Beyond that, specific types of incivility, also known as *hate speech* (e.g., Ziegele, Koehler & Weber, 2018; see also Frischlich and Sponholz in this collection), have raised strong concerns among researchers, since these types are often used to further marginalize certain groups. Uncivil attacks against women in online discussions, for example, often aim to silence and exclude them from political discourse (e.g., Chen et al., 2020). However, various studies have also reported beneficial outcomes of incivility; exposure to uncivil content can, for example, increase people's interest in politics (Brooks & Geer, 2007) and their intentions to participate politically (Borah, 2014; Chen, 2017; Chen & Lu, 2017).

Taken together, empirical studies analyzing the consequences of incivility arrive at different conclusions regarding whether incivility is a good or bad thing. Overall, however, the prevailing claim in public discourse is that incivility is undesirable and needs to be eliminated (Chen et al., 2019). This claim is not only based on empirical findings but also on prescriptive theories. From a deliberation perspective, for example, incivility is mainly considered as undermining deliberative discourse, and from a politeness perspective, it is predominantly assessed as a negative threat to the constructed public self-image of individuals. These prescriptive theories, however, neglect an important argument: just as incivility itself can serve as a tool to silence minorities, calls for civility can also be used as silencing mechanisms (see also Litvinenko in this collection). As of today, various researchers have argued that democracy can endure heated discussions and that high demands for civil discourse can exclude certain social groups, such as educationally disadvantaged milieus (e.g., Bejan, 2017; Estlund, 2008; Garton Ash, 2016). Therefore, calls for "robust civility" (Garton Ash, 2016) or "mere civility" (Bejan, 2017) are being voiced—a civility that is robust and broad, tolerates disagreement, various language styles, and heated discussions.

In a similar vein, a large body of *critical studies* conceive of civility as a set of norms that a powerful elite establishes to suppress marginalized groups. From this perspective, calls for civility mainly serve as an instrument of the powerful to suppress the powerless and reinforce existing power relations and social inequality (e.g., Baez & Ore, 2018; Lozano-Reich & Cloud, 2009; Stuckey & O'Rourke, 2014). According to these studies, the powerful can decide what is considered (un)civil, perform social control, and thus exclude minority voices from political discourse.

When conceptualizing calls for civility as a strategy to exclude and suppress certain groups, the positive implications of incivility emerge. For example, critical studies have acknowledged incivility as an instrument of the powerless to express their identity. From this perspective, incivility is a powerful means of differentiating an oppressor from an oppressed, and thus an out-group from an in-group (e.g., Jamieson et al., 2017). Violations of civility norms can then demonstrate self-assertion and belonging to a marginalized group (e.g., Lozano-Reich & Cloud, 2009; Stuckey & O'Rourke, 2014). Further, marginalized groups can use incivility to draw attention to their problems and fight for their rights. In fact, incivility has been described as the weapon of the powerless (Scott, 1985) and as a strategic instrument of marginalized groups to denounce injustice and seek change. Incivility is then seen as an act of dissent and democratic activism and has important mobilizing functions (Edyvane, 2020; Jamieson et al., 2017). Thus, protest, threats, insults, and several other uncivil expressions against social injustice can sometimes be considered legitimate, and some scholars even plead for an "uncivil tongue" (Lozano-Reich & Cloud, 2009). Other scholars, however, explicitly call for "responsible incivility" (Edyvane, 2020, p. 105). From this perspective, incivility is legitimate only when its positive democratic consequences outweigh the negative ones.

Overall, the normative implications of incivility depend on various factors. An across-the-board evaluation of incivility as something bad seems inappropriate because such an evaluation neglects the sometimes positive effects of incivility and the sometimes legitimate use of an "uncivil tongue" (Lozano-Reich & Cloud, 2009) to fight inequality and injustice. Researchers should, therefore, withstand the temptation to justify the relevance of their own research solely by referring to the destructive effects of incivility. Thereby, they can help to promote a more differentiated perspective on the phenomenon.

4 Towards new perspectives on incivility in political communication

Incivility is a multi-faceted, dynamic, and, partly, elusive phenomenon. What we can say with some confidence is that incivility is mostly situated in the fields of politics and political communication. Moreover, studies are relatively consistent in conceptualizing incivility as a violation of norms, although the specific norms that

incivility violate cover a broad range and include interpersonal politeness norms, deliberative respect norms, liberal democratic norms, and communication norms. Further, an increasing number of studies agree that incivility is a matter of perceptions and, as such, often a violation of multiple norms.

In this chapter, we have outlined various conceptual, methodological, and normative challenges that arise from a multitude of approaches toward incivility. From these challenges, we have derived some potential directions for future research on incivility. More specifically, we recommend developing more consistent operationalizations of incivility, rethinking the ways in which perceived incivility can be measured in content analyses, and broadening the view on when and why incivility is a "good" or "bad" thing.

Despite the challenges related to the concept of incivility, we should not disregard its benefits. Most importantly, by broadly focusing on *norm violations*, incivility resonates with other concepts that investigate specific deviant communicative behaviors, such as *flaming*, *offensive speech, and hate speech* (see Sponholz and Frischlich in this collection). Compared to other concepts of deviant communication, such as *toxicity* (see Risch in this collection), incivility is a strongly theory-based construct that has a long research tradition. Research has provided far-reaching insights into the causes, patterns, and consequences of incivility in offline and online contexts, and future studies can build on established experiences and measurements. Incivility is also tailored to the analysis of political communication among elites and citizens. At the same time, the concept is flexible enough to be applied to non-political contexts, such as the analysis of social interactions in the workplace.

Nevertheless, to exploit the full potential of the incivility concept, we advocate a broad view of the phenomenon that integrates different previous approaches. More specifically, we sketched a perceptual and multidimensional model of incivility (Bormann et al., 2021). This model is built on fundamental concepts of human cooperation and communication, and includes five communication norms (information, process, modality, relation, and political context) that are largely compatible with the multitude of the norm concepts suggested in previous incivility research. Within our integrative approach, we conceive of incivility as disapproved violations of one or several of these communication norms. This concept offers various benefits for future research. First, although our concept is broad enough to cover most norm violations that previous research has identified, it

does not conceive of incivility as a monolith. Rather, the model specifies different types of norm violations in a distinctive way by systematizing them along the five communication norms. Second, owing to its roots in the fundamental processes of communication and cooperation, the concept can be applied to a variety of contexts, ranging from offline political interactions between politicians to online discussions among citizens. Lastly, the concept is based on perceptions or, more specifically, on the disapproval of those involved in the respective communication. Consequently, our concept allows for a less prescriptive and more differentiated perspective regarding which potential norm violations can actually be considered uncivil in specific contexts.

Social norms have always been in flux and are constantly being renegotiated among citizens and elites. The Internet and the social web have accelerated this development, as currently demonstrated by debates around *political correctness* or canceling culture, to name only a few. In these debates, we observe that the perceptions of civility and incivility clash among different camps and that the perceived civil behavior of one's own camp is disapproved of as uncivil by members of the other camp. Further, various communication and behavior that societies have evaluated as civil back in history may be considered uncivil today. For example, denying women the right to publicly raise their voice on political issues and to participate politically was not considered uncivil a few decades ago but certainly would be today. Similarly, in many societies, the use of racial stereotypes was widely perceived as appropriate for a long time but would today be evaluated as an act of incivility. Since incivility is—and will likely always be—subject to individual perceptions and zeitgeists, future research would benefit from paying more attention to the contexts of uncivil communication, such as time, culture, situation, social groups, or issues, for example. With these arguments in mind, we argue that future incivility research should investigate more comprehensively the circumstances under which different individuals and social groups perceive specific norm violations as civil or uncivil and evaluate them as (democratically) legitimate or harmful. Our multidimensional concept offers a fruitful starting point for such research in that it distinguishes between distinct norm violations, considers individual perceptions and evaluations of communication participants, and is applicable to a wide variety of contexts.

Marike Bormann is a postdoctoral researcher at the Department of Social Sciences at the University of Düsseldorf, Germany. https://orcid.org/0000-0002-2879-2471

Marc Ziegele is Assistant Professor at the Department of Social Sciences at the University of Düsseldorf, Germany. https://orcid.org/0000-0002-2710-0955

References

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication, 19*(3), 373–387. https://doi.org/10.1111/jcc4.12009

Baez, K. L., & Ore, E. (2018). The moral imperative of race for rhetorical studies: on civility and walking-in-white in academe. *Communication and Critical/ Cultural Studies, 15*(4), 331–336. https://doi.org/10.1080/14791420.2018.1533989

Bejan, T. (2017). *Mere civility: Disagreement and the limits of toleration*. Harvard University Press.

Ben-Porath, E. N. (2010). Interview effects: Theory and evidence for the impact of televised political interviews on viewer attitudes. *Communication Theory*, *20*(3), 323–347. https://doi.org/10.1111/j.1468-2885.2010.01365.x

Benson, T. W. (2011). The rhetoric of civility: Power, authenticity, and democracy. *Journal of Contemporary Rhetoric, 1*(1), 22–30.

Borah, P. (2014). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research*, *41*(6), 809–827. https://doi.org/10.1177/0093650212449353

Boatright, R., Shaffer, T., Sobieraj, S., & Young, D. G. (Eds.) (2019). *A crisis of civility? Political discourse and its discontents*. Routledge. https://doi.org/10.4324/9781351051989

Bormann, M., Tranow, U., Vowe, G., & Ziegele, M. (2021). Incivility as a violation of communication norms: A typology based on normative expectations toward political communication. *Communication Theory*. https://doi.org/10.1093/ct/qtab018

Brooks, D.J., & Geer, J.G. (2007). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science, 51*(1), 1–16. https://doi.org/10.1111/j.1540-5907.2007.00233.x

Brown, P., & Levinson, S. C. (1987). *Politeness. Some universals in language usage.* Cambridge University Press.

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Palgrave Macmillan.

Chen, G. M., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media*, *61*(1), 108–125. https://doi.org/10.1080/08838151.2016.1273922

Chen, G. M., & Ng, Y. M. M. (2017). Nasty online comments anger you more than me, but nice ones make me as happy as you. *Computers in Human Behavior*, *71*, 181–188. https://doi.org/10.1016/j.chb.2017.02.010

Chen, G. M., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media and Society*, *5*(3), 1–5. https://doi.org/10.1177/2056305119862641

Chen, G. M., Pain, P., Chen, V. Y., Mekelburg, M., Springer, N., & Troger, F. (2020). 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, *21*(7), 877–895. https://doi.org/10.1177/1464884918768500

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. https://arxiv.org/abs/1703.04009

Edyvane, D. (2020). Incivility as dissent. *Political Studies, 68*(1), 93–109. https://doi.org/10.1177/0032321719831983

Estlund, D. M. (2008). *Democratic authority: A philosophical framework.* Princeton University Press.

Fraser, B. (1990). Perspectives on politeness. *Journal of Pragmatics, 14*(2), 219–236. https://doi.org/10.1016/0378-2166(90)90081-N

Garton Ash, T. (2016). *Free speech: Ten principles for a connected world.* Yale University Press.

Gastil, J. (2008). *Political communication and deliberation.* Sage Publications.

Gervais, B. T. (2014). Following the news? Reception of uncivil partisan media and the use of incivility in political expression. *Political Communication*, *31*(4), 564–583. https://doi.org/10.1080/10584609.2013.852640

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics, 12*(2), 167–185. https://doi.org/10.1080/19331681.2014.997416

Goffman, E. (1959). *The presentation of self in everyday life.* Doubleday.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 41–58). Academic Press.

Gutmann, A., & Thompson, D.F. (2004). *Why deliberative democracy?* Princeton University Press.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy*. MIT Press.

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research, 41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Hwang, H., Kim, Y., & Kim, Y. (2018). Influence of discussion incivility on deliberation: An examination of the mediating role of moral indignation. *Communication Research*, *45*(2), 213–240. https://doi.org/10.1177/0093650215616861

Jamieson, K. H. (2000). *Incivility and its discontents: Lessons learned from studying civility in the US House of Representatives.* Allyn and Bacon.

Jamieson, K. H., Volinsky, A., Weitz, I., & Kenski, K. (2017). The political uses and abuses of civility and incivility. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford handbook of political communication.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793471.013.79_update_001

Kalch, A., & Naab, T. K. (2017). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *SCM Studies in Communication and Media*, *6*(4), 395–419. https://doi.org/10.5771/2192-4007-2017-4-395

Lasswell, H. D. (1948). The structure and function of communication in society. In L. Bryson (Ed.), *The communication of ideas* (pp. 37–51). Harper and Brothers.

LfM – Landesanstalt für Medien NRW (2020). *Ergebnisbericht der forsa-Befragung zu Hate Speech 2020*. https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/forsa_LFMNRW_Hassrede2020_Ergebnisbericht.pdf

Lindenberg, S. (2015). Solidarity: Unpacking the social brain. In A. Laitinen & A.B. Pessi (Eds.), *Solidarity. Theory and practice* (pp. 30–54). Lexington Books.

Lozano-Reich, N. M., & Cloud, D. L. (2009). The uncivil tongue: Invitational rhetoric and the problem of inequality. *Western Journal of Communication, 73*(2), 220–226. https://doi.org/10.1080/10570310902856105

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication, 11*, 3182–3202.

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication, 67*(4), 586–609. https://doi.org/10.1111/jcom.12312

Mutz, D. C. (2007). Effects of "In-your-face" television discourse on perceptions of a legitimate opposition. *American Political Science Review*, *101*, 621–635. https://doi.org/10.1017/S000305540707044X

Mutz, D. C. (2015). *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

Mutz, D. C., & Reeves, B. (2005). The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review*, *99*(1), 1–15. https://doi.org/10.1017/S0003055405051452

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society*, *20*(9), 3400–3419. https://doi.org/10.1177/1461444817749516

Papacharissi, Z. (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Prochazka, F., Weber, P., & Schweiger, W. (2018). Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism Studies, 19*(1), 62–78. https://doi.org/10.1080/1461670X.2016.1161497

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior*, *58*, 461–470. https://doi.org/10.1016/j.chb.2016.01.022

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2018). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. https://arxiv.org/pdf/1701.08118.pdf

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*. https://doi.org/10.1177/0093650220921314

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society*, *18*(2), 121–138. https://doi.org/10.1080/1369118X.2014.940365

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice, 8*(1), 18–33. https://doi.org/10.1080/17512786.2013.813194

Schaff, A. (1962). *Introduction to semantics.* Pergamon Press.

Schilpzand, P., Pater, I. E., & Erez, A. (2016). Workplace incivility: A review of the literature and agenda for future research. *Journal of Organizational Behavior, 37*(S1), 57–88. https://doi.org/10.1002/job.1976

Scott, J. C. (1985). *Weapons of the weak: Everyday forms of peasant resistance*. Yale University Press.

Simpson, D. (1960). *Cassell's new Latin dictionary*. Funk and Wagnalls.

Sobieraj, S., & Berry, J. M. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication, 28*(1), 19–41. https://doi.org/10.1080/10584609.2010.542360

Steinmeier, F.-W. (2019, May 6-8). Eröffnung der re:publica 2019 [Keynote], Berlin, Germany. https://www.bundespraesident.de/SharedDocs/Reden/DE/Frank-Walter-Steinmeier/Reden/2019/05/190506-Eroeffnung-Republica.html

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research, 2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. Journal of Public Deliberation, 3(1), Article 12. https://doi.org/10.16997/jdd.50

Stryker, R., Conway, B. A., & Danielson, J. T. (2016). What is political incivility? *Communication Monographs, 83*(4), 535–556. https://doi.org/10.1080/03637751.2016.1201207

Stuckey, M. E., & O'Rourke, S. P. (2014). Civility, democracy, and national politics. *Rhetoric and Public Affairs, 17*(4), 711–736.

Su, L. Y. F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society, 20*(10), 3678–3699. https://doi.org/10.1177/1461444818757205

Tomasello, M. (2008). *Origins of human communication.* MIT Press.

Tomasello, M. (2009). *Why we cooperate*. MIT Press.

Wang, M. Y., & Silva, D. E. (2018). A slap or a jab: An experiment on viewing uncivil political discussions on Facebook. *Computers in Human Behavior, 81*, 73–83. https://doi.org/10.1016/j.chb.2017.11.041

Weber Shandwick (2020). Civility in America 2019: Solutions for tomorrow. https://www.webershandwick.com/wp-content/uploads/2019/06/CivilityInAmerica2019SolutionsforTomorrow.pdf

Ziegele, M., Jost, P. B., Bormann, M., & Heinbach, D. (2018). Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *SCM Studies in Communication and Media, 7*(4), 525–554. https://doi.org/10.5771/2192-4007-2018-4-525

Ziegele, M., Koehler, C., & Weber, M. (2018). Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting & Electronic Media, 62*(4), 636–653. https://doi.org/10.1080/08838151.2018.1532430

Ziegele, M., Weber, M., Quiring, O., & Breiner, T. (2018). The dynamics of online news discussions: Effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. Information, *Communication & Society, 21*(10), 1419–1435. https://doi.org/10.1080/1369118X.2017.1324505

Ziegele, M., Quiring, O., Esau, K., & Friess, D. (2020). Linking news value theory with online deliberation: How news factors and illustration factors in news articles affect the deliberative quality of user discussions in SNS' comment sections. *Communication Research, 47*(6), 860–890. https://doi.org/10.1177/0093650218797884

Digital
Communication
Research.de

Recommended citation: Risch, J. (2023). Toxicity. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 219–230). Digital Communication Research.
https://doi.org/10.48541/dcr.v12.13

Abstract: In research on online comments on social media platforms, different terms are widely used to describe comments that are hateful or disrespectful and thereby poison a discussion. This chapter takes a theoretical perspective on the term toxicity and related research in the field of computer science. More specifically, it explains the usage of the term and why its exact interpretation depends on the platform in question. Further, the article discusses the advantages of toxicity over other terms and provides an overview of the available toxic comment datasets. Finally, it introduces the concept of engaging comments as the counterpart of toxic comments, leading to a task that is complementary to the prevention and removal of toxic comments: the fostering and highlighting of engaging comments.

DOI 10.48541/dcr.v12.13

Julian Risch

Toxicity

1 Toxic comments make readers leave a discussion

In computer science research in the broad field of social media analysis, *toxicity* is a collective term for a variety of phenomena. For several decades, comments have been denoted as toxic if they contain toxic language, including profanity, insults, and hate. As of today, there is no research in the computer science community that specifically addresses and discusses the term toxicity in this context. However, Fortuna et al. (2020) compared different terms across multiple datasets.

The term has become much more popular with the Kaggle Challenge on Toxic Comment Classification in 2018.[1] The general idea of such challenges or shared tasks is to stimulate research by having a competition, where all participants have access to the same training data, develop machine learning models in teams, and compare their model's performance with regard to a pre-defined machine learning task on the same test dataset and the same set of evaluation metrics. A machine learning task can be described as a set of given inputs, for example, social media comments, and expected outputs, for example, the two class labels *toxic* and *non-toxic*. A model that solves the task well can automatically map given inputs to the correct outputs, even for inputs it has not seen before. Common machine learning tasks besides

1 https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

classification are regression and clustering, where inputs are not mapped to previously defined class labels but to numeric values or to previously undefined groups of similar inputs. Kaggle is one of multiple web platforms where such tasks can be hosted as a competition and where the evaluation of the models is automated; for example, the number of correct predictions is calculated automatically.

Several series of shared tasks centered on toxic comment classification have recently emerged, with most of them having yearly or bi-yearly events. For example, HatEval deals with hate speech against immigrants and women (Basile et al., 2019), HaSpeeDe and HASOC with hate speech detection in general (Bosco et al., 2018; Mandl et al., 2019), IberEval assesses automatic misogyny identification (Fersini et al., 2018), GermEval and OffensEval cover offensive language (Struß et al., 2019; Zampieri et al., 2019), and TRAC focuses on aggression (Kumar et al., 2018; Bhattacharya et al., 2020). The main advantages of these competitions are the comparability of the results, the simultaneity of the efforts of the different teams, and thus, the intensive knowledge exchange at workshops typically following the competition. At these workshops, the final results are revealed, and the approaches are published in workshop proceedings.

This particular Kaggle Challenge on Toxic Comment Classification was organized by Google's subunit Jigsaw and allowed participants to compete at a shared task on a provided dataset. With more than 4,500 participating teams and a dataset of 150,000 hand-labeled comments, it was by far the largest shared task for toxic comment classification. The task defined toxic comments as comments that are likely to make a reader leave a discussion. Interestingly, this definition focuses on the effect of toxic comments on others instead of the linguistic features of the content.

The intention behind the definition becomes clearer with a closer look at the task's dataset, which is described in detail in a publication by Wulczyn et al. (2017). The dataset comprises comments that were posted on Wikipedia talk pages, where users discuss article page edits. Rude and disrespectful comments can arise in these discussions if users disagree on how an article page should be edited. In these situations, users might try to silence others and make users with other opinions leave the discussion to enforce their own views and end any further argument. The task definition of the Kaggle Challenge on Toxic Comment Classification includes not only a binary label for toxicity but also finer-grained labels: *toxic* and *severe toxic* as different severity levels of toxicity and a segmentation of toxic comments into

obscenity, threats, insults, and *identity hate* (each comment has been labeled by multiple crowd workers). Due to this fine-grained segmentation, the term *toxic* became established as a collective term for a variety of online comments.

What kind of content is considered toxic depends on the social media platform and its user community. Many platforms provide discussion guidelines that make transparent what rules users must adhere to and on what basis moderators remove content. However, in the end, it depends on the user community what kind of content makes them leave a discussion and is consequently considered toxic. Thus, the definition of toxicity depends on language use that is accepted by the community. For example, profanity might be allowed on some platforms and accepted by their users. While the wording of the definition leaves much room for interpretation, it is interpreted very similarly on many different platforms. The *netiquette,* the etiquette of the Internet, is a set of general guidelines that also applies to online discussions. For example, they are the basis for the discussion rules of online news platforms or Wikipedia.[2]

Due to its broad definition, the term toxicity can be applied to other comment datasets and platforms (van Aken et al., 2018), and this broadness can be seen as its main advantage. Other terms that are frequently used to denote hate speech in computer science research include offensive language, abusive language, and aggression. However, each of these terms describes only a subset of toxic comments. For example, vulgar or obscene language is not necessarily abusive, and benevolent sexism is not necessarily aggressive.[3] Toxicity as a higher-level concept, builds a bridge between the different lower-level concepts. As a consequence, models that need to be good at classifying a particular subset of toxic comments can be pre-trained on other similar subsets of toxic comments.

2 https://www.zeit.de/administratives/2010-03/netiquette/seite-2; https://www.welt.de/debatte/article13346147/Nutzungsregeln.html; https://de.wikipedia.org/wiki/Wikipedia:Wikiquette

3 Note that benevolent sexism is not necessarily perceived as benevolent by the recipient.

2 Toxic comment datasets

To show the diversity of toxicity and to give an overview of what falls under the definition of toxic comments, Table 1 lists the publicly available toxic comment datasets used in related work. While the term toxicity is rarely used in these datasets as a label, the labels used represent subclasses of toxicity. Most of the datasets have been labeled by the researchers themselves but a few of them by crowd workers. The respective publications contain descriptions of the individual datasets. Two recent surveys have compared and discussed the datasets (Poletto et al., 2020; Vidgen and Derczynski, 2020). A more detailed table that includes the number of comments per dataset and their language is also available (Risch, 2020). Table 1 makes clear that the majority of publicly available toxic comment datasets were collected on Twitter (26 out of 41). The set of class labels is more diverse. For example, there are datasets of comments from online news platforms where only one binary label is available, indicating whether the comment was published or removed from the platform (accept/reject). Further, there are different severity levels of toxicity (very toxic/mildly toxic), hate (strong hate/weak hate), and aggression (overtly aggressive/covertly aggressive). Many class labels focus on a particular subset of toxic comments, such as insults, profanity, cyberbullying, stereotypes, racism, and sexism.

Although the detection of toxic comments is challenging, the differentiation of subsets of toxicity is a difficult task on its own (van Aken et al., 2018). Davidson et al. (2017) and Kwok and Wang (2013) studied words that distinguish hate speech from offensive language. Some comments might fall into multiple subclasses, or they can happen to be at the borderline between two classes. An advantage of the term toxicity is that it does not require making a finer-grained and therefore more difficult classification. On the downside, the analysis of toxic comments is limited to a rather general level if no further fine-grained classification is used.

3 Toxic comments vs. engaging comments

Detecting and removing toxic comments prevents them from forcing readers to leave a discussion. Keeping more users engaged in an online discussion also matches the commercial interests of the providers of social media platforms.

They increase their revenue by maximizing the time that users spend on the platform. Thereby, they can show more ads and promote content to their users. An example is the shadow banning used by Twitter, where a toxic comment's visibility on the platform is reduced up to the point where it can only be seen by directly accessing the author's page.

An advantage of the term *toxic comment* over other terms is that it allows an elegant way of defining an opposite category of comments: while toxic comments make other users leave a discussion, engaging comments make other users join a discussion (Risch & Krestel, 2020).

The latter encourages users to actively join a discussion by replying to another user's comment or voting on a comment. Not only are these engaging comments thought-provoking, but they also stimulate users to express their opinions by posting a reaction. The concept of engaging comments has its roots in the concept of engaging, respectful and/or informative conversations (Napoles et al., 2017). A different definition considers constructiveness to be the opposite of toxicity (Kolhatkar & Taboada, 2017). However, constructiveness refers more to the content of the comment, whereas toxicity and engagement refer more to the effect of the comment. The two categories, toxicity and engagement, are not necessarily mutually exclusive. Comments can be rude and disrespectful, thereby making some users leave a discussion while at the same time, they can trigger some other users to join the discussion, either contributing counter-speech or, in the worst case, adding more toxic comments.

While social media platforms detect toxic comments to remove them, the detection of engaging (or constructive) comments would increase their visibility and highlight them on the platform. In the same direction, fostering engaging comments could help to have more diverse opinions in discussions, as it encourages more users to join a discussion.

4 Conclusion

Toxicity describes comments that make readers leave a discussion, for example, because of profanity, insults, threats, or hate speech. This chapter described the origins of this term and showed how it comprises the class labels used in various comment datasets. With this wide range being one main advantage of the term, the

Table 1: Toxic comment datasets (sorted by year of publication)

Study	Platform	Class labels
Kwok and Wang, 2013	Twitter	Racism
Djuric et al., 2015	News	Hate
Waseem, 2016	Twitter	Racism, sexism
Waseem and Hovy, 2016	Twitter	Racism, sexism
Badjatiya et al., 2017	Twitter	Racism, sexism
Davidson et al., 2017	Twitter	Hate, offense
Gao and Huang, 2017	News	Hate
Jha and Mamidi, 2017	Twitter	Benevolent/hostile sexism
Mubarak et al., 2017	News	Reject
Pavlopoulos et al., 2017	News	Reject
Schabus et al., 2017	News	Argument, discrimination, inappropriate, sentiment, off-topic
Vigna et al., 2017	Facebook	Strong/weak hate
Wulczyn et al., 2017	Wikipedia	Attack
Albadi et al., 2018	Twitter	Hate
Álvarez-Carmona et al., 2018	Twitter	Aggressive
Bosco et al., 2018	Facebook	Strong/weak hate
Bosco et al., 2018	Twitter	Aggression, hate, irony, offense, stereotype
Fersini et al., 2018	Twitter	Derailment, discredit, harassment, misogyny, stereotype, target
Founta et al., 2018	Twitter	Abuse, aggression, cyberbullying, hate, offense, spam
de Gibert et al., 2018	Forum	Hate
Kumar et al., 2018	Facebook	Aggression, covert, overt
Ljubešić et al., 2018	News	Reject

Sanguinetti et al., 2018	Twitter	Aggression, hate, irony, offense, stereotype
Wiegand et al., 2018	Twitter	Abuse, insult, profanity
Zhang et al., 2018	Twitter	Hate
Basile et al., 2019	Twitter	Aggression, hate, target
Fortuna et al., 2019	Twitter	Hate, target
Ibrohim and Budi, 2019	Twitter	Abuse, strong/weak hate, target
Kolhatkar et al., 2019	News	Very toxic, toxic, mildly toxic
Mandl et al., 2019	Twitter	Hate, offense, profanity, target
Mulki et al., 2019	Twitter	Abuse, hate
Ousidhoum et al., 2019	Twitter	Group, hostility, sentiment, target
Ptaszynski et al., 2019	Twitter	Cyberbullying, hate
Qian et al., 2019	Misc	Hate
Struß et al., 2019	Twitter	Abuse, insult, profanity, explicitness
Zampieri et al., 2019	Twitter	Offense, target
Bhattacharya et al., 2020	YouTube	Aggression, sexism, covert, overt
Caselli et al., 2020	Twitter	Abuse, explicitness
Çöltekin, 2020	Twitter	Offense, target
Pitenis et al., 2020	Twitter	Offense
Sigurbergsson and Derczynski, 2020	Misc	Offense, target

chapter also described the definition of its counterpart as another advantage. In contrast to toxic comments, engaging comments make readers join a discussion. Therefore, online platforms should detect both toxic comments and engaging comments to either increase or decrease their visibility. An interesting path for future work is to investigate the overlap of these two categories of comments.

Julian Risch is a senior machine learning engineer at deepset.

References

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 69–76.

Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y- Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. *Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval@SEPLN)*, 74–96.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *Companion Proceedings of the International Conference on World Wide Web (WWW Companion)*, 759–760.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. *Proceedings of the International Workshop on Semantic Evaluation (SemEval@NAACL)*, 54–63.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., & Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@LREC)*, 158–168.

Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the EVALITA 2018 hate speech detection task. *Proceedings of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, *2263*, 1–9.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 6193–6202.

Çöltekin, Ç. (2020). A corpus of Turkish offensive language on social media. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 6174–6184.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 512–515.

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, 11–20.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. *Companion Proceedings of the International Conference on World Wide Web (WWW Companion)*, 29–30.

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at IberEval 2018. *Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval@ SEPLN)*, 214–228.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 94–104.

Fortuna P., Soler, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? an empirical analysis of hate speech datasets. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 6786–6794.

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 491–500.

Gao, L., & Huang, R. (2017). Detecting online hate speech using context aware models. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, 260–266.

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 46–57.

Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. *Proceedings of the Workshop on Natural Language Processing and Computational Social Science (NLP+CSS@ACL)*, 7–16.

Kolhatkar, V., & Taboada, M. (2017). Constructive language in news comments. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 11–17.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2019). The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, *4*(2), 155–190.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@COLING)*, 1–11.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 1621–1622.

Ljubešić, N., Erjavec, T., & Fišer, D. (2018). Datasets of Slovene and Croatian moderated news comments. *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, 124–131.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in Indo-European languages. *Proceedings of the Forum for Information Retrieval Evaluation*, 14–17.

Mubarak, H., Kareem, D., & Walid, M. (2017). Abusive language detection on Arabic social media. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 52–56.

Mulki, H., Haddad, H., Bechikh Ali, C., & Alshabani, H. (2019). L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 111–118.

Napoles, C., Tetreault, J., Pappu, A., Rosato, E., & Provenzale, B. (2017). Finding good conversations online: The yahoo news annotated comments corpus. *Proceedings of the Linguistic Annotation Workshop (LAW@EACL)*, 13–23.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4675–4684.

Pavlopoulos, J., Malakasiotis, P., Bakagianni, J., & Androutsopoulos, I. (2017). Improved abusive comment moderation with user embeddings. *Proceedings of the Natural Language Processing meets Journalism Workshop (NLPmJ@EMNLP)*, 51–55.

Pitenis, Z., Zampieri, M., & Ranasinghe, T. (2020). Offensive language identification in Greek. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 5113–5119.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation, 55*(1), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Ptaszynski, M., Pieciukiewicz, A., & Dybała, P. (2019). Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval Workshop*, 89–110.

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP- IJCNLP)*, 4755–4764.

Risch, J. (2020). *Reader comment analysis on online news platforms* (doctoral thesis). Universität Potsdam. https://doi.org/10.25932/publishup-48922

Risch, J., & Krestel, R. (2020). Top comment or flop comment? Predicting and explaining user engagement in online news discussions. *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 579–589.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2798–2805.

Schabus, D., Skowron, M., & Trapp, M. (2017). One million posts: A data set of German online discussions. *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, 1241–1244.

Sigurbergsson, G. I., & Derczynski, L. (2020). Offensive language and hate speech detection for Danish. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 3498–3508.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval task 2, 2019 shared task on the identification of offensive language. *Proceedings of the Conference on Natural Language Processing (KONVENS)*, 352–363.

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, 33–42.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS One*, *15*(12), 1–32.

Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *Proceedings of the Italian Conference on Cybersecurity (ITASEC)*, 86–95.

Waseem, Z. (2016). Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. *Proceedings of the Workshop on NLP and Computational Social Science (NLP+CSS@EMNLP)*, 138–142.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. *Proceedings of the Student Research Workshop@NAACL*, 88–93.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. *Proceedings of the Conference on Natural Language Processing (KONVENS)*, 1–10.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. *Proceedings of the International Conference on World Wide Web (WWW)*, 1391–1399.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *Proceedings of the International Workshop on SemanticEvaluation (SemEval@NAACL)*, 75–86.

Zhang, Z., Robinson, D., & Tepper, J. A. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. *Proceedings of the Extended Semantic Web Conference (ESWC)*, 745–760.

Digital
Communication
Research.de

Recommended citation: Udupa, S. (2023). Extreme speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 233–248). Digital Communication Research. https://doi.org/10.48541/dcr.v12.14

Abstract: Extreme speech is a critical conceptual framework that aims to uncover vitriolic online cultures through comparative and ethnographic excavations of digital practices. It is not one more new definition or a term replaceable with *extremist speech*. Rather, it is a conceptual framework developed to foreground historical awareness, critical deconstruction of existing categories, and a grounded understanding of evolving practices in online communities, in ways to holistically analyze the contours and consequences of contemporary digital hate cultures. This framework suggests that the *close contextualization* of proximate contexts—of media affordances in use or situated speech cultures—should accompany *deep contextualization*, which accounts for grave historical continuities and technopolitical formations unfolding on a planetary scale. Through such elaborate forays into everyday practices and deeper histories, extreme speech theory proposes to nuance normative and regulatory efforts to classify and isolate hate speech and disinformation.

DOI 10.48541/dcr.v12.14

Sahana Udupa

Extreme Speech

1 Introduction

Extreme speech is a critical conceptual framework that has drawn attention to online vitriolic cultures by ethnographically analyzing digital practices and online user communities from a comparative, historically sensitive perspective.

The concept of *extreme speech* departs, in particular, from the dominant legal-normative definitions of *hate speech* and the discourse of securitization around *terrorism* and *political extremism*. These definitions approach *hate speech* primarily as a discourse of pathology by predetermining the effects of online volatile speech as vilifying, polarizing, or lethal. *Extreme speech* instead stresses the importance of holistic comprehension over classification by placing such practices in a broader context of contestations over power and allowing normative approaches and mitigation efforts to emerge from a grounded, historically aware analysis (Pohjonen & Udupa 2017; Udupa, 2015, 2017, 2019, 2020; Udupa et al., 2021; Udupa & Pohjonen, 2019). In this sense, the contributions of extreme speech research *qualify*, rather than seek to replace, the existing repertoire by highlighting areas that hate speech research has insufficiently explored as well as by drawing attention to the political consequences of hate speech discourse.

In terms of its definitional scope, extreme speech analysis focuses on derogatory speech forms aimed at any group (including groups that hold power) and

exclusionary discourses with hateful language and expressions dressed as *facts* that implicitly or explicitly exclude or harm a person or group on the basis of their group belonging. Derogatory extreme speech is particularly ambivalent since it represents online discourses that challenge the protocols of polite language to speak back to power, but it also constitutes a volatile slippery ground on which what is comedic and merely insulting could quickly slide down to downright abuse and threat. For content moderation, such derogatory expressions can serve as the earliest cultural cues to brewing and more hardboiled antagonisms. The analysis of exclusionary extreme speech builds on existing definitional standards of *hate speech* set up by the United Nations and Wardle and Derakhshan's (2017) distinction between *disinformation* ("when false information is knowingly shared to cause harm") and *malinformation* ("when genuine information is shared to cause harm") (p. 5). Extreme speech analysis covers *misinformation* (spreading false information without an intention to cause harm) so far as it is part of the social fields, where deliberate efforts to spread hate activate a variety of actors and networks, which ultimately spread hateful language that could harm vulnerable groups. The purpose of extreme speech analysis is, therefore, to exceed the legal focus on culpability and instead analyze—with ethnographic and historical depth—how different actors and actions animate one another and how new interventions must be crafted to address not only actors who deliberately engineer hateful language and disinformation but also those who succumb to it or do it to earn a livelihood. This approach allows researchers and policymakers to chart new analytical pathways and diverse fields of action, beyond intentionality-based investigations.

The focus on cultural practice is especially important for extreme speech analysis. In particular, it calls for ethnographic explorations of media cultural practice—that is, what people do that relates to media (Couldry, 2012) within particular structural conditions that shape and are shaped by such practices. The practice approach emphasizes that political configurations of discourses and inherited dispositions prefigure mediated action inasmuch as users' situated practices alter political discourse.

These analytical moves require situating the contemporary moment of online volatile speech within regional and historical contexts—ranging from the micro-contexts of online user cultures and the contradictory pulls of realpolitik to macro-historical formations of colonial imperialism—as a necessary corrective to the seeming universality of the normative basis of the hate speech discourse.

Coloniality is conceptualized as a global process that institutionalized and legitimized three sets of relations—market relations, nation state relations and racial relations—that constitute a composite structure of oppression with impacts beyond the actual geographies that the Empire colonized. For this reason, coloniality is a relevant critical framework to understand contemporary formations of inequality and repression, including those articulated through digital mediations and how they are especially shaped by digital capitalist logics that facilitate and exacerbate vectors of difference. Such a historically contextualized understanding calls for a comparative analysis that looks beyond the West, extending its focus into the rapidly expanding online worlds of the Global South. The comparative approach here is not based on a model with quantitative metrics that are tested across selected case studies; rather, it is rooted in ethnography of practice and historical anthropology (Van der Veer, 2016).

2 The limits of hate speech

Key interventions of the conceptual framework of extreme speech have emerged from highlighting the limits of the hate speech discourse while also recognizing its significance as a regulatory concept. The legal-regulatory terminology of *hate speech* draws on longer legal debates over speech restrictions (Nockleby, 2000; Udupa et al., 2020; Warner & Hirschberg, 2012). Although legal traditions and scholarly discussions differ, a common element throughout this discourse is the assumption that hate speech involves the disparagement of other groups, based on their belonging to a particular group with a collective identity. Waldron (2012) argues that this kind of speech has two key characteristics: the first is to dehumanize members who belong to another group, and the second is to reinforce the boundaries of the in-group against the out-group by attacking the other group's members. Hate speech discourse predefines the effects of hate speech as *negative* and *damaging*, and its regulatory rationale is, thus, of control and containment. The state is the largest actor in this effort, but internet intermediaries also increasingly monitor and restrict speech on their platforms. Responding to civil society concerns, governmental injunctions, and international conventions on hate speech, online forums and social networking sites have developed their own terms of service to detect, regulate, and prohibit hate speech.

As it jostles between state regulation, the capitalist market, and political fields, hate speech has become what Brubaker and Cooper (2000) would describe as a *thick concept* with a "tangle of meanings" and an evaluative load (p. 14). Moreover, these concepts become empirical objects in themselves; the researcher's task would be merely to discover the degree of variance or agreement between different kinds of online speech from this ideal object type. Extreme speech calls such contextual flattening into question.

Furthermore, as a form of power, the discourse of hate speech is inextricably tied to the state and its political economies of violence. Historically, it emerged from the projects of civility that coincided with (and partly constituted) the state's monopolization of violence (Giddens, 1987; Thirangama et al., 2018). The moral claims of liberal thought require that hate speech regulation protects substantive virtues, such as sympathy and understanding (at least in the procedural terms of decorum), in the interest of a common good. Liberal understandings premised on abstract principles of equality conceal multifarious and, at times, manipulative political agendas that have grown around the regulatory discourse of hate speech.

Moreover, the liberal moral principle of civility that partly informs the rationale of hate speech is "intimately tied up with class and race privilege," which consolidated the colonial and postcolonial state (Thirangama et al., 2018, pp. 153–155). Colonial histories have cemented the self-righteous schema of the liberal center (the self-understanding of the West) and extreme periphery (the rendering of the non-West), which is now manifest in diverse forms of political grandstanding and control not only between the (former) metropole and colony but also within the nation-states where similar structures of speech restriction, based on moral self-understandings, have taken root.

Under these conditions, the pressure to speak the *polite* language has been an act of domination—moral injunctions linked to assertions of privilege. Civility, thus, is an "effect of political recognition and of a responsive structure of authority" (Mitchell, 2018, p. 217). In other words, the implications of incivility—or the extremeness of speech more broadly—cannot be apprehended without analyzing particular forms of recognition and responsiveness to demands that are available to diverse groups.

The thick concept of *hate speech* comes with an evaluative load aimed at immediate action, raising the risk of glossing over historical trajectories, as well

as the ambivalence of extremeness within particular contexts of power. This is not merely a fine grained theoretical objection but also, more gravely, a political problem. Both historically and in the contemporary moment, the ambivalence of extreme speech is closed off when political actors who are pressured to do something about hate invoke the label of *hate speech* (Pohjonen, 2019), at times brutally using force to target marginalized groups. Examples abound of regimes misusing the hate speech discourse to squash dissent or target minoritized groups. In the context of India, currently ruled by a Hindu nationalist regime, selective application of state restrictions on online speech has cited the "law and order" rationale, invoking the legally imprecise term of *hate speech* in conjunction with colonial legislations around sedition or outraging religious feelings (Modh, 2015). Such restrictions on the national level have sought to quell dissenting voices, while regional governments with diverse ideological agendas, set in a multiparty system of competitive electoral politics, have mobilized similar efforts to frame political opposition as *hate speech*.

In Kenya's context, Katiambo (2021) has argued that "the polysemy of extreme speech is removed when *incivility* becomes known as hate speech, blocking us from ever knowing its alternative possibilities" (p. 49). In everyday conversational contexts, *hate speech* is often used as a charge or an accusation that closes off, rather than opens up, avenues for change and dialogue (Boromisza-Habashi, 2013).

Recognizing the limits of hate speech both as a regulatory value and a concept-in-use in everyday interactions, ethnographic sensibility advocated by extreme speech research insists that the moral charge around vitriol and disinformation should come from lived concepts and situated contexts, rather than frameworks imposed from the outside. This shift requires a critical approach that is sensitive to cultural variations in speech, including sanctioned forms of disrespect; political contexts where hate, as an order value of regulation, is assigned to speech acts; and historical conditions that implicate extreme speech with particular forms of power—subversive in some contexts and repressive in others.

3 Extreme speech as methodology

The conceptual framework of extreme speech comes with a set of methodological perspectives.

3.1 Comparative Practice

Extreme speech research proposes to map a critical typology of vitriol based on historical, cultural, and political variations, and a focus on media-cultural practice described in the preceding section. This methodological approach might be described as *comparative practice*, where interlocking factors in different national and regional scenarios are studied for their specificities and in relation to one another.

3.2 Everydayness and emic categories

Drawing from an anthropological emphasis on everyday cultures, extreme speech research draws attention to *emic* categories (i.e., categories derived from the perspectives of research participants than the observer), through which the complex use of language operates. Methodologically, it involves exploring the meanings that online users, as historical actors, attach to *vitriol* and the diverse practices that congeal around them.

Online *gaali*, in the Indian context, might illustrate such an emic category (Udupa, 2017). *Gaali* is a Hindi term for a complex amalgam of abrasive, abusive, or unabashed language seen as joking and disrespectful at the same time. It is a commonly invoked term to define the aggressive styles of online debating cultures. Online gaali has provided new avenues of participation for politically savvy internet users, especially among the educated middle-class groups in urban India and diverse class groups with access to mobile media who feel confident that they can trump legacy media and political authorities by engaging in social media discussions. While anti-establishment *gaali* does not always articulate progressive politics, gaali's performative spread has, nonetheless, brought new political voices to the fore of public debate. By online actors' own account, gaali—as rancorous rabble-rousing—has helped them thrust their voices into the public domain hitherto dominated by the state and organized commercial media. Consequently, gaali has sparked voluminous online contestations around the developmental, representational, and economic issues facing contemporary India.

At the same time, the blurred arena of online comedy, insult, and abuse that gaali represents has facilitated the perpetuation of religious majoritarian nationalism and

exclusionary discourses centering on assertions of Hindu-first India. Often, online gaali grows into a full-blown shaming punishment, articulating nationalism through the gendered trope of regulating sexuality and what Irvine (1993) calls "evaluative talk" (p. 106). Online gaali as gendered abuse has led to severe cases of intimidation and harassment against female online commentators.

Nested in digital culture but drawing on longer histories, gaali has spawned the interlocking practices of insult, comedy, shame, and abuse that unfold in a blurred arena of online speech. On this slippery ground of shifting practices, comedy stops and insult begins or insult morphs into abuse in mutually generative ways. Contextually sensitive analysis reveals, in this case, gaali's Janus-faced status as performance; while its routine detoxification opens up new lines of participation in political discourse to online users, it takes a menacing edge when they instantiate gendered discursive relations of nationalism.

3.3 Empathy and reflexivity

Other key methodological approaches of extreme speech research include reflexivity and empathy. It is difficult to develop access to complex ground realities that are rife with contradictions without sustained ethnographic engagement among communities even when such communities harbor despicable or less than ideal political views. Sustained engagement comes with a commitment to extend the same principles of honesty and openness that inform a sound ethnographic practice. Arguably, the foremost ethical principle in advancing such an ethnographic sensibility is empathy, which is guided by a commitment to learn and see insider views as a *working morality*. Empathy as a practical or working morality in ethnographic practice does not, in itself, entail an endorsement of the views expressed by online actors or claims to moral equivalence between different ideological positions. As researchers explore the political implications of digital practice in each case to its fullest possible detail, empathy as an ethical stance allows them to avoid a tendency for critique to precede understanding or for a moral-evaluative framework to predetermine what to expect. Empathy demands active dispositions on the researcher's part, foremost a commitment to listen to actors who inhabit the digital world through myriad expressions, aspirations, habits, and tactics, including those aimed at advancing politically problematic

ideologies. As we, as researchers, navigate our interlocuters' diverse narratives and life worlds, anthropological reflexive praxis is especially pertinent since our own positionalities are intricately interwoven with digital discourses, and our material, social, and political circumstances shape the ethical, affective, and political terms with which we approach online speech as problematic or otherwise (Udupa & Dattatreyan, 2023).

4 Global conjuncture and deep contextualization

These methodological moves are important in advancing a *conjunctural analysis* of varied forces, rather than assessing social and political worlds based on predefined normative categories (Mankekar, 1999). By the same token, inasmuch as extreme speech stresses the analytical value of highlighting ambiguity in online speech, it is methodologically equipped to examine the diverse factors that shape particular political formations. In this sense, extreme speech avows ethnographic specificities—but in ways that connect contexts and situate them within socio-technological transformations that are unfolding on a global scale and in relation to long-standing historical processes.

Gleaning from cases around the world, extreme speech analysis has highlighted that, over the last two decades, vitriolic cultures have precipitated a condition of violent exclusion based on "exacerbated fracture lines of difference that include race, gender, sexuality, religion, nation and class" in a context where "computational capital has built itself and its machines out of those capitalized and technologized social differentiations" (Beller, 2017).

We define this condition as the global conjuncture of affects, actors, and affordances that is driving contemporary forms of exclusionary extreme speech. The socio-technological mediations of internet-based media are particularly significant in this conjuncture; we argue that they constitute a context in themselves, rather than acting as mere channels for discourses external to them. In particular, exclusionary extreme speech rides on digital affordances of peer-to-peer mobilizations, continuous exchange, platform migration, and layered anonymity.

Through the lens of Ahmed's (2004) semiotic analysis of affect, it is possible to see digital mediation as mechanisms that materialize the surfaces of hateful bodies through association, alignment, displacement, and "stickiness" (p. 89). If

hate is part of the "production of the ordinary" (Ahmed, 2004, p. 56), digital exchange has realized hate by bringing hateful expressions closer to one's everyday conversational realities. Tagging on to the small-screen intimacy of digital exchange, hate passes to the ordinary in continuous loops, powered by the systematic channeling of affect—of anger, glee, envy, and the transgressive pleasures of vitriolic online exchanges—within the participatory condition of digital capitalism (Udupa, 2020). I have argued that fun is a particularly significant affective infrastructure in ramping up online extreme speech among right-wing ideological communities in digital environments (Udupa, 2019). From quasi-public forums such as Twitter to image boards such as 4Chan, hate sticks to bodies through signs that are constantly innovated upon in *creative funny* ways, allowing the affective economy of hate to spread laterally between peers in solidarity.

Yet, far from a media-centric argument and claims that online affordances have let loose humankind's most primal animosities, extreme speech analysis highlights interconnections and continuities underwritten by longer historical processes. Exclusionary online extreme speech is shaped by the longer *global* process of colonial modern relations that unfolds as both internal and external forces in different societies. Colonial relations could be traced along three interconnected lines: nation-state relations established by colonial power, which frames the boundaries of minority-versus-majority and inside-versus-outside; market relations institutionalized by colonial power, which now manifest as uneven data relations; and racial relations naturalized by colonial power, which dispose people as objects of hatred (Udupa, 2020).

This analysis is a corrective to not only liberal moral panics about digital communication but also certain strands of Western left intellectualism that anxiously term ongoing digital turbulences as a "strange brew of bellicosity, disinhibition and rancor" among people who have been pushed to the wrong side of economic liberalization (Brown, 2019, p. 61). Such analysis elides the grave history of systematic violence that installed unequal racialized relations through actions—past and present—that are orchestrated, directed, and economic inasmuch as they are helpless reactions of backbiting revenge.

Following De Genova (2010), these historical conditions could be defined as postcolonial *metastasis*. Assertions of aggrieved power, common among White supremacists, emanate not only from structural subordination under oppressive market conditions but also through a sense of dethronement—a product of

far-reaching global imperial legacies. Crucially, through nation-state relations canonized by colonialism, this aggression wrought by imaginary wounds unfolds *within* different national and subnational contexts as racialized relations of majoritarian belligerence. Hindu nationalists in India, Sinhalese nationalists in Sri Lanka (Aguilera-Carnerero & Azeez, 2016), Han nationalists online in China (de Seta, 2021), the Sunni majoritarian politics around blasphemy in Pakistan (Schaflechner, 2021), Duterte's trolls in the Philippines (Ong & Cabanes, 2018), and online nationalists in Nepal (Dennis, 2017) are some examples, and so are the meme makers in northern Chile who seize internet memes' mashup cultures to portray migrants from Bolivia and Peru as backward, dirty, uneducated plunderers of limited resources and contributors to cultural degradation (Haynes, 2019). Such exclusionary discourses against *immigrants* (a category that emerged from the nation-state distinction between inside and outside) and *minorities* (a category that emerged from the nation-state distinction between a majority and a minority) are rife with racialized portrayals. Colonialism reproduced hierarchy and difference as intrinsic features of the modern nation-state, and this process of racialization of social relations within the newly stabilized structure of the nation-state alongside market relations was *global* in scope (Shankar, 2020; Treitler, 2013).

The framework of extreme speech has, thus, emphasized that longer historical processes should be examined in relation to proximate contemporary contexts of digital circulation and practice—a kind of dual analysis that might be described as *decolonial thinking*. This kind of analysis is not a macrohistorical glossing of diverse power dynamics. Without doubt, affective energies that emanate from and animate internet spaces should be analyzed in relation to specific structures of animosities and interlocking systems of coercion and power along various axes—including race, class, gender, religion, caste, nationality and ethnicity—that have precipitated the current global conjuncture of exclusionary extreme speech. Intersectionality invites attention to structures of power that predated, comingled or remained rather independent of colonial occupation. However, conceptualizing colonialism as a *set of relations* (market, nation-state and race) is important in tracking the overarching frameworks and historical continuities that undergird contemporary forms of exclusionary extreme speech. We might call this analysis *deep contextualization*. Decolonial thinking suggests that the *close contextualization* of proximate contexts—of media affordances in use or situated speech cultures—

should accompany *deep contextualization* that accounts for grave historical continuities and technopolitical formations unfolding on a planetary scale.

5 People-centric models of moderation

Through such elaborate forays into everyday practices and deeper histories, extreme speech theory proposes to nuance normative and regulatory efforts at classifying and isolating hate speech and disinformation. In this regard, regulatory and policy approaches honed by extreme speech perspectives call for people-centric models that can account for cultural variation, ambiguities, and dynamic forms of vitriolic online exchange.

An illustrative case might be the AI4Dignity project, a European Research Council-funded project that I run as the principle investigator. The project has partnered with independent fact-checkers from the Global North and the Global South as critical community intermediaries in developing artificial intelligence-assisted models for speech moderation. Recognizing that human supervision is critical, the project has devised ways to connect, support, and mobilize existing communities who have gained reasonable access to the meaning and context of speech because of their involvement in online speech moderation of some kind. Building spaces of direct dialogue and collaboration between artificial intelligence (AI) developers and relatively independent fact-checkers who are not part of a large media corporation, political party, or social media company is a key component of AI4Dignity. Moreover, this dialogue has involved academic researchers specialized in particular regions as facilitators. Through this triangulation, AI4Dignity's process model has aimed to stabilize a more encompassing collaborative structure in which *hybrid* models of human-machine filters can incorporate dynamic reciprocity between critical communities, such as independent fact-checkers, AI developers, and academic researchers. These efforts offer pathways to ground big data and computational methods with the extreme speech framework's emphasis on critical sensibility to cultural difference, historical contexts, local practices, and meanings drawn by users themselves in everyday lived environments.

Importantly, such efforts offer ways to bring inclusive training data sets to AI models. These datasets are more inclusive because they are based on culturally coded, linguistically diverse, and dynamic expressions that critical communities—

such as fact-checkers—can locate, rather than based on corporate social media definitions or annotations that natural langauge processing (NLP) experts develop within their professional fields. AI4Dignity's labeling process has involved reflexive and active iterations between ethnographers, communities, and AI developers. These iterations have, at times, led to confusing twists and turns in the annotation process, but they have also strengthened efforts to bring cultural nuance to data sets. For instance, at the beginning of the annotation process, confusion arose around the distinction between the three labels *derogatory extreme speech* (defined as expressions that do not conform to accepted norms of civility within specific local or national contexts and targeted at any group but not explicitly excluding vulnerable and historically disadvantaged groups; it includes derogatory jokes and sobriquets; Udupa, 2020), *exclusionary extreme speech* (defined as expressions that call for or imply excluding disadvantaged and vulnerable groups; Udupa, 2020) and *dangerous speech* (defined as expressions that have reasonable chances to trigger or catalyze harm and violence against target groups; Benesch, 2013). We had drawn this distinction based on published work and after some internal discussions with team members, but when we invited collaborating fact-checkers to categorize social media passages under one of these labels, several questions came up. A partnering fact-checker remarked that all extreme passages they encountered were indeed *dangerous* in the broadest sense of negatively affecting society. This opinion was, indeed, completely legitimate, but I requested that he appreciate efforts to keep the categories more precise because, once machine learning (ML) models begin to categorize, these mapped data sets could have regulatory implications. In the next round of discussions, we observed more clarity around the term *dangerous speech*, but fact-checkers found the distinction between *derogatory extreme speech* and *exclusionary extreme speech* rather slippery and difficult to operationalize. These questions led us to clarify the definitions by listing target groups. (For *derogatory extreme speech*, we listed protected categories such as gender, caste, ethnicity, and national origin, as well as racialized categories, but also the state, legacy media, politicians and civil society representatives advocating for inclusive societies). Our objective was to capture the cultural patterns of speech forms that are seen as uncivil within specific linguistic, cultural contexts but also express diverse and ambivalent forms of political contestation, as mentioned at the beginning of this article. We did not include the state, legacy media and politicians as target groups under *exclusionary extreme speech* since this label was meant

to capture expressions that exclude marginalized and vulnerable groups. AI developers were keen to keep the labels as precise as possible, while participating fact-checkers were keen to see more target groups added to the list. After several iterations, the project has received annotated passages for the three categories in the English, Hindi, Swahili, German, and Portuguese languages from partnering fact-checkers. These fact-checkers have brought—with their keen understanding and involvement in the political discourses of the region and its lifeworlds—linguistically diverse, contextually rich datasets to the ML pipeline, allowing the automated detection of problematic online speech to acquire some degree of cultural knowledge and contextualization.

Aside from its efforts to bring contextually sensitive, inclusive datasets to ML models, AI4Dignity aims to develop a tool for fact-checkers, expanding the access to AI-related technological resources for communities who are actively involved in grounding digital discourse in democratic values in different regions of the world. AI4Dignity's collaborative process model and policy engagements around AI-assisted content moderation have directly emerged from the extreme speech framework and its emphasis on comparative ethnographic excavations of the complex politics surrounding online speech.

Thus, as a critical framework, extreme speech offers methodological, policy, and theoretical perspectives rooted in ethnographic sensibility and historical awareness, toward envisioning a (digital) world of dignity.

Sahana Udupa is Professor of Media Anthropology at the University of Munich, Germany.
https://orcid.org/0000-0003-3647-9570

Acknowledgments

The extreme speech concept has developed through various publications listed in the reference list below. I thank all the coauthors, reviewers, and different publication venues from which this contribution has drawn. I also acknowledge the generous funding of the European Research Council (under the European Union's Horizon 2020 research and innovation program: grant agreement No. 714285, Project ONLINERPOL, and No. 957442, Project AI4Dignity).

References

Aguilera-Carnerero, C., & Azeez, A. H. (2016). Islamonausea, not Islamophobia: The many faces of cyber hate speech. *Journal of Arab & Muslim Media Research*, *9*(1), 21–40. https://doi.org/10.1386/jammr.9.1.21_1

Ahmed, S. (2004). *The cultural politics of emotion*. Edinburgh University Press.

Beller, J. (2017). *The fourth determination*. E-Flux. https://www.e-flux.com/journal/85/156818/the-fourth-determination/

Benesch, S. (2013). Dangerous Speech: A Proposal to Prevent Group Violence. *Dangerous Speech Project*. Available at: https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf

Boromisza-Habashi, D. (2013). *Speaking hatefully: Culture, communication, and political action in Hungary*. The Pennsylvania State University Press.

Brown, W. (2019). Neoliberalism's Frankenstein: Authoritarian freedom in twenty-first century "democracies." *Critical Times, 1*(1), *60–79*. https://doi.org/10.1215/26410478-1.1.60

Brubaker, R., & Cooper, F. (2000). Beyond "identity." *Theory and Society*, *29*(1), 1–47. https://doi.org/10.1023/A:1007068714468

Couldry, N. (2012). *Media, society, world: Social theory and digital media practice*. Polity.

De Genova, N. (2010). Migration and race in Europe: The trans-Atlantic metastases of a post-colonial cancer. *European Journal of Social Theory*, *13*(3), 405–419. https://doi.org/10.1177/1368431010371767

de Seta, G. (2021). The politics of Muhei: Ethnic humor and Islamophobia on Chinese social media. In S. Udupa, I. Gagliardone, & P. Hervik (Eds.), *Digital hate: The global conjuncture of extreme speech*. Indiana University Press.

Dennis, D. (2017). Mediating claims to Buddha's birthplace and Nepali national identity. In S. Udupa & S. McDowell (Eds.), *Media as politics in South Asia* (pp. 176–189). Routledge.

Giddens, A. (1987). *The nation state and violence*. Polity Press.

Haynes, N. (2019). Writing on the walls: Discourses on Bolivian immigrants in Chilean meme humor. *International Journal of Communication*, *13*, 3122–3142.

Irvine, J. T. (1993). Insult and responsibility: Verbal abuse in a Wolof village. In J. H. Hill & J. T. Irvine (Eds.), *Responsibility and evidence in oral discourse* (pp. 105–134). Cambridge University Press.

Katiambo, D. (2021). It is incivility, not hate sppeech: Application of Laclau and Mouffe's discourse theory to analysis of non-anthropocentric agency. In S. Udupa, I. Gagliardone, & P. Hervik (Eds.), *Digital hate: The global conjuncture of extreme speech*. Indiana University Press.

Mankekar, P. (1999). *Screening culture, viewing politics: An ethnography of television, womanhood and nation in postcolonial India*. Duke University Press.

Mitchell, L. (2018). Civility and collective action: Soft speech, loud roars, and the politics of recognition. *Anthropological Theory*, *18*(2–3), 217–247. https://doi.org/10.1177/1463499618782792

Modh, Ketan. (2015). *Controlling hate speech on the Internet: The Indian perspective*. SSRN. https://ssrn.com/abstract=2783447

Nockleby, J. T. (2000). Hate speech. In W. L. Levy, K. L. Karst, & A. Winkler (Eds.), *Encyclopedia of the American Constitution* (pp. 1277–1279). Macmillan.

Ong, J. C., & Cabanes, J. V. (2018). *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines*. The Newton Tech4Dev Network. https://newtontechfordev.com/wp-content/uploads/2018/02/ARCHITECTS-OF-NETWORKED-DISINFORMATION-FULL-REPORT.pdf

Pohjonen, M., and Udupa, S. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, *11*, 1173–1191.

Pohjonen, M. (2019). A comparative approach to social media extreme speech: Online hate speech as media commentary. *International Journal of Communication*, *13*, 3088–3103.

Schaflechner, J. (2021). Blasphemy accusations as extreme speech acts in Pakistan. In S. Udupa, I. Gagliardone, & P. Hervik (Eds.), *Digital hate: The global conjuncture of extreme speech*. Indiana University Press.

Shankar, A. (2020). Primitivism and race in ethnographic film: A decolonial re-visioning. *Oxford Bibliographies in Anthropology*. Oxford University Press.

Thirangama, S., Kelly, T., & Forment, C. (2018). Introduction: Whose civility? *Anthropological Theory*, *18*(2–3), 153–174. https://doi.org/10.1177/1463499618780870

Treitler, V. B. (2013). *The Ethnic Project: Transforming racial fiction into ethnic factions*. Standford University Press.

Udupa, S. (2015). Abusive exchange on social media: The politics of online gaali cultures in India. Media Anthropology Network 52nd E-seminar, European Association of Social Anthropologists, July.

Udupa, S. (2017). Gaali cultures: The politics of abusive exchange on social media. *New Media and Society*, *20*(4), 1506–1522. https://doi.org/10.1177/1461444817698776

Udupa, S. (2019). Nationalism in the digital age: Fun as a metapractice of extreme speech. *International Journal of Communication*, *13*, 3143–3163.

Udupa, S. (2020). *Decoloniality and extreme speech*. Media Anthropology Network 65th E-Seminar, European Association of Social Anthropologists. https://www.easaonline.org/downloads/networks/media/65p.pdf

Udupa, S., & Dattatreyan, E.G. (2023). *Digital Unsettling: Decoloniality, Disposession, Rupture*. New York University Press.

Udupa, S., & Pohjonen, M. (2019). Extreme speech and global digital cultures. *International Journal of Communication*, *13*, 3049–3067.

Udupa, S., Gagliardone, I., & Hervik, P. (2021). *Digital hate: The global conjuncture of extreme speech*. Indiana University Press.

Udupa, S., Gagliardone, I., Deem, A., & Csuka, L. (2020). *Field of disinformation, democratic processes and conflict prevention*. Social Science Research Council. https://www.ssrc.org/publications/view/the-field-of-disinformation-democratic-processes-and-conflict-prevention-a-scan-of-the-literature/

Van der Veer, P. (2016). *The value of comparison*. Duke University Press.

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe. https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Langauge in Social Media* (pp. 19–26). Association for Computational Linguistics. https://www.aclweb.org/anthology/W12-2103

Digital
Communication
Research.de

Recommended citation: Quandt, T., & Klapproth, J. (2023). Dark participation: Conception, reception, and extensions. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 251–270). Digital Communication Research. https://doi.org/10.48541/dcr.v12.15

Abstract: While the new possibilities of online participation were initially described and analyzed from a mainly optimistic perspective, more recent work in communication studies draws a rather bleak picture of the state of communication in today's online world. The concept of "dark participation" (Quandt, 2018) picks up on this profound change of perspective. In addition to the systematization of negative participatory forms, the concept was also used as a rhetorical device to comment on the change in scientific perspective: the original publication was primarily meant as a call for balance in the analysis of online participation—something that was often neglected in the subsequent debate. Based on a brief summary of the core ideas and the context of the original publication, the current paper revisits the concept of dark participation by discussing its reception and potential extensions. Furthermore, a reassessment of its value and the limitations for analyzing (negative) forms of online participation is presented vis-à-vis related concepts.

DOI 10.48541/dcr.v12.15

Thorsten Quandt & Johanna Klapproth

Dark Participation

Conception, reception, and extensions

1 Scientific construction of a changing mosaic

Numerous scientific articles analyzing online communication start with overarching statements about "all-encompassing and unprecedented media change" and suggest that "the Internet revolutionized not only the media system but also how we live as a society." Typically, these studies illustrate statements with cogent arguments and middle-range empirical work on aspects of communication that support the idea of a "media revolution," incrementally contributing pieces to a grand mosaic of what public communication in the current era looks like. Indeed, it has been argued that normal science can be considered a laborious and collaborative process of piecing together such a mosaic image based on existing patterns of thinking about the world (Kuhn, 1962). In that sense, the grand picture depends not only on its object, but also on the concept, tone, and style of the representation, as well as the arrangement of elements and even the individual tessera.

Judging from recent work in communication studies, one might get a rather bleak impression of the state of communication in the online world, as if the mosaic is full of pitch-black pieces and the overall atmosphere is dark and depressive. Researchers have identified "toxic talk" (Anderson et al., 2018) and

"partisan incivility" (Muddiman & Stroud, 2017) in online discussions and comment forums, even going so far as to declare a "cyberspace war" that uses "propaganda and trolling as warfare tools" (Aro, 2016). Online communication seems to be pervaded by "hate speech" (Silva et al., 2016) and "fake news" (Bennett & Livingston, 2017, 2018; Lazer et al., 2018; Wardle & Derakhshan, 2017) that are assumed to be a serious danger to societal coherence. As a protection against this, scholars propose interfering by "moderation" (Ziegele et al., 2018), "deplatforming" (Chandrasekharan et al., 2017; Rogers, 2020), "counter speech" (Bartlett & Krasodomski-Jones, 2015; Garland et al., 2020), or other means of "controlling the conversation" (Santana, 2016). Indeed, further inspection of current communication journals and conferences would most likely strengthen this rather dismal impression of today's online world. In that sense, even the current volume is a reflection of this and may add further tessera to the mosaic.

However, turning back the pages of said journals and checking the volumes from just a few years ago would reveal a completely different, uplifting, and much more positive picture. One and a half decades ago, scientists described the online world using bright colors, and there was a lot of hope and optimism in their analyses. In contrast to the depictions of today, scholars were hoping for a "communicative democracy in a redactional society" (Hartley, 2000) in which users were empowered to become part of the production process (labeled "produsage" by Bruns, 2008), leading to the "future of news and information" via "we media" (Bowman & Willis, 2003). "The people formerly known as the audience" (Rosen, 2006) would become actively engaged in the information flows, leading to "an age of participatory news" (Deuze et al., 2007). There was a spirit (and expectation) of revolution in many of these works, not only for information flows, media, and journalism, but for society as a whole. The new options of online participation were also regarded as a rejuvenation of—somewhat congealed—media democracies by means of an "online agora" as the ideal space for a digital assembly of the people.

Naturally, the inconsistency of these two totally different depictions of online communication leads to an important question: Has the world changed so much in such a short time—or just the scientific perspective?

This is a difficult question to answer since the observer may have changed in tandem with the object being observed. Naturally, even long-term empirical data are subject to (re)interpretation, but certainly some of the forms or participation

heralded by communication scholars briefly after the millennium still exist—and one could even argue that the options for participation have dramatically improved since then. However, these positive spaces are often overlooked in light of the negative aspects so prominently featured in today's research and public discussion. This may be partially due to frustrations with the empirically observable world not following the normative ideas and expectations espoused back then (both in science and society). In line with this assumption, Peters and Witsche argued that we came "from grand narratives of democracy" and ended up with "small expectations of participation" (2014). Usher and Carlson even identified a "midlife crisis of the network society" (2018).

This profound change in the perspective and tone of the discussion about online participation was also the motivation of one of this chapter's authors to introduce a concept called "dark participation" (Quandt, 2018). On the surface level, the original article was a reflection and systematization of the negative or even sinister forms of participation scholars seem to witness these days—at first sight, another dark tessera added to the overall picture. However, on a second, more subtle level, the original article was also used as a rhetorical device to comment on the change in perspective. It included a call for balance in the discussion instead of overpronouncing dark aspects in favor of more positive ones (or vice versa). In that sense, the article and concept were something of an academic conjuring trick: by presenting the audience with a dark tessera and discussing it in detail, the author enticed the audience to follow his argument and the idea of an overly negative, depressing mosaic—only to reveal that this was done on purpose and that caution is necessary when arguments appear one-sided.

The dark participation concept quickly developed a life of its own, with a notable—and sometimes critical—reception. Further work also embraced the systematization of dark participation. It needs to be noted, though, that some of the discussion overlooked the more complex nature of the original publication, while others extended it beyond what the author hoped for (or even considered), partially transforming it into something else (e.g., Kowert, 2020).

Therefore, the current paper will revisit the concept of dark participation by briefly summarizing its core ideas and the context of its original publication, discussing its reception and potential extensions, and finally re-assessing its value—and limitations—for analyzing current (negative) forms of online communication vis-à-vis other related concepts.

2 From "Participation" to "Dark Participation"

2.1 The reversal of the participation concept

As noted above, participation in online media was a highly relevant concept in many theoretical and empirical works in communication studies at the beginning of the millennium. The options online communication offered, in contrast to the traditional system of societal information distribution (primarily via media and journalism), were considered promising for both socio-political and economic reasons. Some scholars argued that online communication would turn the information and news flow from a lecture to a "conversation" (Kunelius, 2001), while media businesses and journalists more often perceived the new influx of user-generated content as a valuable resource for exploitation (Vujnovic et al., 2010). Accordingly, the understanding of participation at that time ranged from the limited contribution of raw material to the production processes in journalism and enclosed debates in "walled gardens" of forums provided by media companies (Domingo et al., 2008; Hanitzsch & Quandt, 2012) to the influential and decisive role of citizens in public communication as active „produsers" (Bruns, 2008).

As a reflection of this range of ideas and the empirical work on the topic, Domingo et al. (2008) proposed the conceptualization of participation as a continuum along an analytical grid consisting of five stages of news production (access and observation, selection/filtering, processing/editing, distribution, and interpretation) that may or may not be (partially) open for citizen participation. This is in line with more general conceptualizations of citizen participation in relation to other aspects of societal life that preceded the discussion of online communication. In such early works (in political science and sociology, for example), it was noted that participation can take multiple forms and may reach various levels, ranging from non-participation and placebo forms of tokenism to decisive citizen power in societal processes (Arnstein, 1969).

Despite this potential variance, the general expectations regarding participation in online media were high. Scholars hoped such participation would have a positive effect on journalistic businesses (which were already struggling), public communication, and society in general. However, many of these works at the beginning of the millennium suffered from notable limitations. They modeled participation as an enhancement to or extension of the existing system of information flows in society,

with citizens contributing to the well-known information production processes enabled by (journalistic) institutions and actors. Social media as we know it today were in their infancy—the forerunner SixDegrees.com had economically failed and closed in 2001, and Mark Zuckerberg only started to work on what would become Facebook in 2003. In that sense, many communication and journalism scholars approached participation from the perspective of the previous traditional system, and this viewpoint presented natural limitations to the visions of the future since institutionalized media and journalism in particular were still regarded as the most relevant references for the understanding of information flows. In line with this understanding, many scientific (empirical) works dealt with participation in the news-making process or contributions to forums provided by journalistic media.

Furthermore, the early conceptualizations often implicitly understood the participating citizens as intrinsically motivated members of liberal democracies; thus, they were following normative ideas of how an ideal society should communicate. The options for online communication were regarded as the key to a door opening to a free and mutual exchange of ideas that, more often than not, was perceived as a solution to many societal problems (such as hegemonial structures and the neglect of minorities). In that sense, the previous system of limited access to information distribution and control via journalistic gatekeepers was regarded as suppressing an existing motivation to communicate and participate, and online communication was a liberating force for this will to participate.

The subsequent developments, and especially the success of social media, did not necessarily follow the expected path. While the number of users communicating in (more or less) publicly accessible online spaces began to grow, their motivations and contributions were often much different from the normative concepts of participation that had been implicitly projected as serving democracy and the public interest. Naturally, there were motivated onliners contributing to societally relevant information production and public discourse, sometimes even in the expected narrow sense but more often in a much wider sense and not necessarily reflecting a traditional (journalistic) definition of "relevant news." However, an active contribution to information flows regarding issues of public interest—even in a rather broad sense—was not necessarily what most people regarded as their main interest in online communication. As a result, the unfolding new communication system was, essentially, quite different from early expectations and was certainly not just an extension of a traditional news-centered media system.

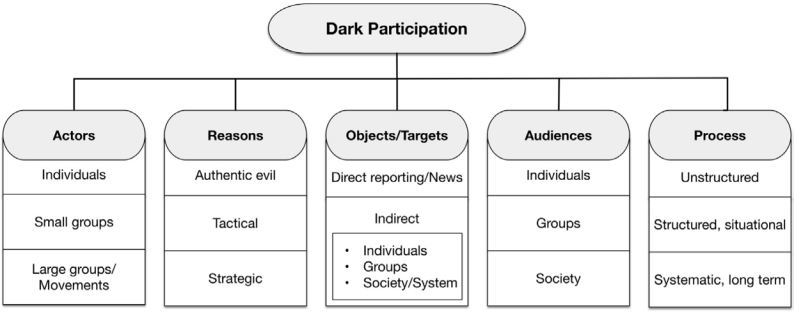
Indeed, one may argue that large parts of social media communication today are tied to the individual experiences of users and are private in nature. And on some platforms, only a fraction of users actively contribute content (Springer et al., 2015). Thus, some of the academic discussion on the problems of today's online communication can be understood as a reaction to a violation of expectations (as also noted by Peters & Witschge, 2014). To put it more precisely, the actual active user was a disappointment when judged against the normative ideal of a highly engaged online citizen fully motivated to serve democracy via participation in information flows and valuable contributions to public discourse.

However, when judging from the early, quite hopeful perspective, the situation is probably even worse: not only does a rather limited fraction of users participate, and often in a different way than expected, but some of these users do not follow the principles of constructive, positive participation. Instead, they spread lies or hate and act in a destructive or manipulative way, as also discussed in the current volume. These rather "sinister" forms of participation were not only disappointing; they also seemed to offer a glimpse at the dark heart of society, in stark contrast to the hopeful promises that followed the new millennium. Accordingly, many communication scholars switched their perspective by 180 degrees and fully embraced the research on manipulation (as discussed above), negativity, and hate fueled by a fear of the individual and social damage these may cause. And perhaps also by a slight fascination with evil and darkness.

2.2 Systematization of dark participation

The dark participation concept introduced by Quandt (2018) addresses this debate in the field and offers, on the surface level, a universal "umbrella" concept for the various forms of negative, manipulative, or destructive participation. The initial article introduces the concept based on a reflection of the situation in the field and then systematizes the various strands of debate and the corresponding sub-concepts into a general model. This model delineates variants of dark participation (see Figure 1). It includes five main dimensions on which variations of dark participation may occur: the actors (i.e., participators), the reasons for their behavior, the targets or objects of their participation, the intended audiences, and the structure of the process.

Figure 1: Dark participation umbrella model



Source: Quandt, 2018, p. 42

Actors are differentiated according to size and complexity, ranging from individuals to large movements (since forms of dark participation are often carried out by coordinated groups or ideological movements). The *reasons* for dark participation can be classified as tactical or strategic, since they are often intended to achieve a situational or long-term goal (such as orchestrated hate campaigns). There are also purely destructive actions that do not follow goals beyond the destruction itself; in that sense, the actions are self-serving (trolls often claim that they just do it "for the sake of it" or "for fun"; Buckels et al., 2014). It needs to be noted that this differentiation already refers to the fact that, despite being perceived as "sinister" from the outside and when judged against societal norms, forms of dark participation may serve a function for the actors. Such functions range from signaling a standpoint or exerting social influence and control to emotional gratifications (Erjavec & Kovačič, 2012).

The third category refers to *objects or targets* of participation. As noted in the original publication with reference to participation in journalistic forums or social media, actors "may attack specific articles or topics, and they can also divert content-driven hate to actors mentioned in the article or the journalists themselves" (Quandt, 2018, p. 38). In that sense, participation in such contexts may directly attack the communication of others, the authors of said communication, or third parties (that may or may not be addressed in said communication). Even a mix of direct and indirect targets may occur. For example, during the "refugee

crisis" in Europe of 2015/16, right-wing participators targeted press articles on refugees in the comment sections of journalistic media and typically criticized the journalists for not telling the "full truth." Thus, "the press and journalism in general became representative of an adverse system and the intended target of the negativity" (Quandt, 2018, p. 42).

Audiences must not be confused with the former category. For example, by bullying others or starting hate campaigns against specific societal groups, actors often try to address an "overhearing" audience or third groups that are not directly involved. These actors want to "convey a message" to these groups (such as showing how relevant or powerful the actors are, where they stand politically, or who they oppose in order to attract supporters for their cause or new followers for a movement, etc.). The intended audience can even extend to the whole of society, such as when groups try to position themselves according to their political/ideological standpoint via dark participation.

Finally, the *process* category refers to the structure and planning of the process. As discussed in the original publication, some forms of dark participation may be "unstructured and random," some "structured, but still bound by the specifics of the situation" and others "systematic and long-term processes" (Quandt, 2018, p. 43). These variations are not fully independent from reasons and motivations since large-scale strategic disinformation campaigns are typically planned and systematic long-term processes, whereas individual outbreaks of emotion-driven, situational trolling may not be following a clearly defined, structured process (incidentally, such a process does not equal behavioral patterns as observed by scientists).

The original model, as summarized here, is deliberately broad and all-encompassing. It is meant to offer a rather universal system of categorizing all potential forms of dark participation according to the main categories. While the original publication presents several examples and references to empirical research, they are primarily meant to illustrate the more frequent variants. Naturally, some combinations are more likely than others: as outlined above, long-term strategic actions of co-ordinated groups will be typically planned and structured, whereas individual tornados of rage will be most likely unstructured, episodic, and not follow a long-term strategy (as noted above). This does not rule out divergent options, though; for example, the latter can be part of a larger plan if groups use highly emotional trolls in an instrumental way. Other empirically less frequent

and therefore less "typical" combinations are also easily conceivable and underline the spectrum of possibilities the dark participation model offers.

2.3 The reversal of the reversal: Dark participation as a mirror trick

The concept and model were quickly picked up in the field and were subject to numerous reactions, from embrace to rejection (see below). However, it has often been overlooked that the original article has a dual message and uses the concept of dark participation as a tool to illustrate the fallacies of normative, one-sided approaches. By introducing the concept and developing it in a way that is similar to earlier works on participation, Quandt tries to lure the reader to his side of the argument, only to reveal in the last few sections of the paper that the construction of a convincing, one-sided argument solely in favor of dark participation was a "mirror trick" (Quandt, 2021, p. 85) meant to evoke a reflection on normativity and empirical balance in the research on participation: "If you now believe that the future is all doom and gloom, then you have stepped into a trap I intentionally set" (Quandt, 2018, p. 44). So, the article deliberately misleads the reader about its goals, and it is designed as an "experienceable" warning. In the final sections, the author argues that embracing the concept and model of dark participation without considering other forms of participation would be as wrong as the earlier works were in their overwhelmingly positive (and therefore uncritical) approaches to participation:

> (...) the current wave of apocalyptic analyses of media and society are partially born out of the same fallacies that plagued the early enthusiastic approaches. (...) The issue here is not the (most relevant) topic of dark participation itself, but a growing lopsidedness that repeats the earlier failings in approach, just with an inversed object of interest. (Quandt, 2018, p. 44)

Thus, dark participation is not only a concept, but also a commentary on the mistakes of doing one-sided research as a projection of one's own expectations. Therefore, the concept can still be used as an umbrella term for specific forms of participation – but never in a nonreflective way and without proper balancing (i.e., one should not forget that participation as a concept has a history and a much broader meaning). In this sense, dark participation is also an incomplete

concept by design. A more general approach to participation—neither naively positive nor fascinated by the dark—would be needed to fully achieve the goal of a balanced discussion:

> (...) media and communication research must be careful that it is not taking the exception as the rule. (...) A normalization of the debate and maturity beyond uni-polar depictions of the world is essential. (...) This would require the development of integrative theories on the conditions of participation that are neither driven by wishful thinking nor doom and gloom. (Quandt, 2018, pp. 44–45)

3 Reception and discussion of the concept

The original publication of "Dark Participation" stimulated a discussion on the concept and led to some "strong, and sometimes even quite emotional reactions" (Quandt, 2021, p. 84). This may be due to the dual message of the piece and its critical perspective on previous approaches to participation (including the work of the piece's author).

For example, Carpentier et al. (2019) criticized the concept of dark participation on the basis of a democratic theory perspective. The authors point out that dark participation and related concepts are rather "perversions of participation" (Carpentier et al., 2019, p. 25). From their (normative) perspective, participation is an essential component of democracy and, as such, an ethical idea by definition. Carpentier et al. argue that this understanding of participation as an ethical idea allows for a differentiation of participation intensity but makes concepts of bad or dark participation inherently contradictory. Instead, the authors propose a focus on differences in participation intensity. Furthermore, they distinguish between participation and the results of participation, and they emphasize that although participation is ethical in itself, the results of participation may not necessarily be ethical. Consequently, even if the results are negative, the process of participation is ethical in itself. From their perspective, the social practices covered by the concept of dark participation cannot be considered participatory, and they perceive them as antagonistic forms of violence (Carpentier et al., 2019).

In contrast to Carpentier et al., Kligler-Vilenchik (2018) does not rule out the possibility of dark participation. Rather, she calls for concurrent research on "good participation" (p. 111) and proposes focusing more scientific attention on the research

of participatory phenomena in everyday contexts, assuming it would align better with the positive view of participation. She further argues that one should not limit oneself to case studies with extreme examples (Kligler-Vilenchik, 2018).

Other authors did not necessarily criticize the concept and the differentiation between "dark" and other, more positive forms of participation. Instead, they asked for more details, expansions, or a different contextualization. For example, in his initial commentary on the original piece, Katz (2018) proposes an integration of the concept into a historical perspective. He identifies parallels between the current situation and the arrival of the telephone and considers the lack of making such connections in a more systematic way a "missed opportunity" (Katz, 2018, p. 104).

While some of these critical pieces (of which the ones mentioned above are just a selection) make some valuable points about dark participation as a theoretical concept, they partially miss its use as a means to elicit an "aha reaction" by the reader in the context of the original publication (as outlined above). Indeed, one may even argue that the article's somewhat uncommon "mirror trick" has been overlooked by some critics, and that their criticism therefore points in the wrong direction (since their position does not oppose the original piece's stated intent).

4 Extensions and transfer of the concept

The original publication not only elicited a critical reception, but also prompted follow-up works that expanded on its core ideas. In some ways, this is to be expected: as a universal concept, dark participation is deliberately open to further delineation and can function as a starting point for empirical research and theoretical extensions. In particular, the concept has been picked up by journalism research since it aligns with the long tradition of research on participation in that particular field.

For example, Nordheim and Kleinen-von Königslöw (2021) identify a growing infiltration of the journalistic system by antagonistic actors as concomitant with the process of digitalization due to a specific destructive potential inherent in participatory technology. They argue that this infiltration of the system intensifies journalism's already-existing crisis. To describe and classify these relationships and sample cases of antagonistic behavior, the authors expand the concept of dark participation by drawing on "The Parasite," a work

by French philosopher Michel Serres (2007). Building on a normative perspective (which interestingly very much contrasts the above-mentioned normative criticism by Carpentier et al., 2019), the authors take up the concept of dark participation and add the idea of certain actors being "parasitic," such as political-strategic actors and self-proclaimed "alternative media" of the alt-right, manipulators that use journalistic structures for disinformation campaigns, and even large platform providers. These "parasites" position themselves as intermediaries of the system's boundaries. As Nordheim and Kleinen-von Königslöw (2021) note, the parasites then function as a subsystem and inherent part of the journalistic system and act from within by utilizing journalistic resources while compromising the values on which the freedoms of a democratic public are based. Furthermore, parasitic disruption triggers differentiation and de-differentiation in the media system and initiates a re-definition of system boundaries. In such a dysfunctional process, the parasites destructively modify the system from within (as both part of it and as an antagonist force), ultimately threatening its integrity.

Based on the understanding of participation as "one of the guiding normative values of journalism in the digital sphere" (as proposed by Kreiss and Brennen, 2016), Anderson and Revers (2018) draw on the concept of dark participation and contribute to a deeper understanding of the evolution and transformation of participation by reconstructing the evolution of societal and journalistic meta-discourse about citizen participation in the news production process. In their socio-historic analysis (which potentially adds the missing historic perspective called for by Katz, 2018; see above), they also problematize participation as an underlying journalistic epistemology. As a form of journalistic knowledge, this "participatory epistemology" modifies professional expertise through public interaction—although not always with expected or desirable results, as they conclude, "Dismissing the interests of Trump supporters as false consciousness does not detract from the uncomfortable reality that the internet gave many people the opportunity to find and express their previously unheard voices and make them heard, including by reproducing and modifying racist memes" (Anderson & Revers 2018, p. 32). As they note, however, the roots of this may be found earlier and in an ideologically very different context, i.e., in the early left-activist Indymedia movement that "was one of the earliest progenitors of these developments,

promiscuously mixing participation, political identity, and agonistic politics, and deeply influencing journalism as a result" (p. 32).

While the above authors extended the theoretical base concept or contextualized it, others differentiated it by identifying factors that may influence the phenomenon or explain its current flourishing. Sjøvaag (2019) suggests a refinement of the concept by considering the economic interests of the media that may contribute to the persistence of dark participation. She argues that media deliberately opened spaces for participation—and thus opportunities for dark participation as well—for financial reasons. They promoted the production of user-generated content as a content strategy with a particularly low cost (Sjøvaag, 2019).

User-generated content as a target of dark participation has also been discussed by others. For example, Van Leuven et al. (2018) note that it is becoming increasingly difficult for journalists to identify the dissolving boundaries between elite and non-elite actors. For example, astroturfing campaigns or the manipulation of online discussions serve as means to maximize the public relations efforts of elite actors, and due to the strong presence of influencers, the boundaries between public relations material and user-generated content are also increasingly dissolving (Van Leuven et al., 2018). Essentially, this enables various options for manipulation and dark participation.

Finally, the concept of dark participation has also been transferred to contexts beyond journalism and social media. For example, Kowert (2020) analyzes the degradation of gamer cultures into toxic ones due to the prevalence of "toxic" gamer behavior characterized by exclusion and hostility. She draws on the concept of dark participation in order to categorize and analyze forms of toxicity in games. To do so, she develops a comprehensive catalogue of what can be defined as dark participation in games and classifies toxic behaviors based on characteristic features on a spectrum ranging from verbal to behavioral and transient to strategic (Kowert, 2020).

5 Dark participation as work in progress

When the dark participation concept was proposed just a few years ago (in 2018), it seemed to hit a nerve within the academic community of communication scholars. Not only did it trigger critical reflection in debate pieces (see above) and

serve as a reference point for empirical studies[1], it also led to several extensions and transfers beyond the intended application in social media and online forums of journalistic media. The concept is obviously universal enough to be applied to related areas, such as participation in digital games (see above). This universality is not necessarily surprising since the overview in the original article was "one that leaves the concept fully open for further delineation," as de Vreese (2021, p. 215) notes.

This openness was purposeful, as discussed above. The original publication worked on two levels: it introduced the concept of dark participation itself and outlined its potential variants in a general model. Furthermore, it used the development of this model as a persuasive device to later reveal to the reader that this model—when not being balanced against other forms of participation—may be as misleading and one-sided as earlier approaches to normatively positive participation. In that sense, both concept and model were meant to be incomplete, as they ignored certain aspects of participation by design (analogous to earlier approaches but with reversed intentions).

This form of self-awareness may be a benefit of the concept vis-à-vis other concepts that are currently discussed in relation to issues of online communication (such as online hate speech, incivility, mis- and disinformation etc.; see the chapters by Sponholz, Frischlich, Benesch, Bormann & Ziegele, and Udupa in this volume for a more comprehensive discussion). Dark participation—when used as intended—links to the previous rich discussion of participation in the field and does not negate earlier approaches, instead balancing them with an intentionally bleak mirror image (that is, indeed, a reflection in a dual meaning). This embedding in an ongoing debate on participation may be seen as a relevant advantage of the concept, especially when approaching it from a communication studies perspective: participation as a process has been at the heart of numerous works in political communication and journalism studies. These discuss the role of actively participating citizens in democratic processes or public communication, and dark participation builds on these rich foundations. Related to this, by pronouncing the role of the actors (as participants) in an inherently social process (i.e., participation), the approach is genuinely compatible with a *social*-scientific viewpoint,

1 The use in empirical studies was not the focus of this theory-oriented overview. For examples, see Bodrunova et al. (2021), Chang, Haider and Ferrara (2021), Frischlich, Boberg and Quandt (2019), and Wintterlin et al. (2020, 2021).

arguably more so than approaches that primarily link the issues to a specific type of content (such as hate speech, mis-/disinformation, etc.).

In addition, the weakly specified, rather universal model allows—and even calls—for extensions. Indeed, as noted above, several authors developed the concept further or took it as a starting point for their own deliberations. Some linked it to broader debates on the role of citizens and other participants in public (online) communication, while others added more depth to the categorization and specified various forms of dark participation. Admittedly, some of these works took the concept and model as their starting point "without the proper 'balancing' contextualization — maybe overlooking the mirror trick this article [the original publication] really is" (Quandt, 2019, p. 85). However, there were numerous thoughtful expansions that placed the piece in context, and even without contextualization, expansions may be very valuable as long as the warning of the original piece about a one-sided discussion of participation is not ignored in the field in general.

It needs to be noted, however, that the benefits of the concept may also be its greatest weaknesses: The concept is tied to actors and the process of participation in social contexts—and therefore, it is also open to other actors' (re)interpretation and multiple viewpoints. The perception of participation as "dark" is an external attribution; as noted above, a destructive and seemingly dysfunctional action (when judged against social norms) may be totally functional from the subjective viewpoint of the participators or supporting parties. Here, content-based concepts (such as hate speech) may be easier to discern since they may be linked to specific and measurable content features (such as negative sentiments, swear words, etc.), whereas the views of participators, the targets of dark participation, the various audiences, and the external scientific observer will most likely diverge. Indeed, this may lead to a discussion of values and norms and what type of (anti)social behavior is defined as "dark"—and by whom.

Furthermore, the universal approach of the model makes it largely unspecific. While the original publication included some cases that were used to illustrate variants of dark participation, it did not offer an exhaustive mapping of empirical cases on the dimensions outlined by the model (since this mapping was not within that piece's scope). Indeed, one may even argue that the model is so universal that it may be transferred to all kinds of participation, not just its "dark" form—potentially with the exception of the "authentic evil" reason subcategory (which could be re-labeled, in a more generic way, as spontaneous, transient, and affective; this

may also include "positive" forms of impulsive, emotion-driven behavior without tactical elements or strategic planning). As explained above, this openness for further delineation was done on purpose, but leaves the concept as a work in progress *by design*. In that sense, a more comprehensive discussion of dark participation would not only mean a differentiation, refinement, and expansion of the concept and model itself, but also a re-balancing and consolidation with all other potential forms of participation, in line with the original piece's intent.

As de Vreese states, this process could entail a re-calibration of communication studies in general and lead beyond "the 'doom and gloom' perspective" that "seems pervasive" these days:

> In the midst of worries about, and research into trolling, incivility, conspiracy, mis- and disinformation, automated pollution of the information environment, populism, and democratic backsliding, is there also space for optimism and a positive research agenda? (...) The bottom line is, that in the era of darkness, it will also be a task of scholars to provide guidance on the upsides. (de Vreese, 2021, p. 216)

In this sense, dark participation is not only a concept. As paradoxical as this may seem at first sight, it is also a call for action to research the concept and to understand participation in a much broader—and positive—way.

Thorsten Quandt is Professor of Online Communication at the University of Münster, Germany. https://orcid.org/0000-0003-1937-0282

Johanna Klapproth is a communication scientist at the University of Münster, Germany.

References

Anderson, A. A., Yeo, S. K., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2018). Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, *30*(1), 156–168. https://doi.org/10.1093/ijpor/edw022

Anderson, C. W., & Revers, M. (2018). From counter-power to counter-pepe: The vagaries of participatory epistemology in a digital age. *Media and Communication*, *6*(4), 24–25. https://doi.org/10.17645/mac.v6i4.1492

Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Planning Association*, *35*(4), 216–224.

Aro, J. (2016). The cyberspace war: Propaganda and trolling as warfare tools. *European View*, *15*(1), 121–132. https://doi.org/10.1007/s12290-016-0395-5

Bartlett, J., & Krasodomski-Jones, A. (2015). Counter-speech examining content that challenges extremism online. *Demos*. http://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bodrunova, S. S., Litvinenko, A., Blekanov, I., & Nepiyushchikh, D. (2021). Constructive aggression? Multiple roles of aggressive content in political discourse on Russian YouTube. *Media and Communication*, *9*, 181–194. http://dx.doi.org/10.17645/mac.v9i1.3469

Bowman, S., & Willis, C. (2003). *We media: How audiences are shaping the future of news and information*. The Media Center at the American Press Institute.

Bruns, A. (2008). *Blogs, Wikipedia, Second Life and beyond: From Production to Produsage*. Peter Lang.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Carpentier, N., Melo, A., & Ribeiro, F. (2019). Rescuing participation: A critique on the dark participation concept. *Communicação e Sociedade*, *36*, 17–35. https://doi.org/10.17231/comsoc.36(2019).2341

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here. *Proceedings of the ACM on Human-Computer Interaction* (1–22). https://doi.org/10.1145/3134666

Chang, H. C. H., Haider, S., & Ferrara, E. (2021). Digital civic participation and misinformation during the 2020 Taiwanese presidential election. *Media and Communication, 9*(1), 144–157. https://doi.org/10.17645/mac.v9i1.3405

Deuze, M., Bruns, A., & Neuberger, C. (2007). Preparing for an age of participatory news. *Journalism Practice*, *1*(3), 322–338. https://doi.org/10.1080/17512780701504864

de Vreese, C. (2021). Beyond the darkness: Research on participation in online media and discourse. *Media and Communication*, *9*(1), 215–216. https://doi.org/10.17645/mac.v9i1.3815

Domingo, D., Quandt, T., Heinonen, A., Paulussen, S., Singer, J., & Vujnovic, M. (2008). Participatory journalism practices in the media and beyond: An international comparative study of initiatives in online newspapers. *Journalism Practice*, *2*(3), 326–342. https://doi.org/10.1080/17512780802281065

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Frischlich, L., Boberg, S., & Quandt, T. (2019). Comment sections as targets of dark participation? Journalists' evaluation and moderation of deviant user comments. *Journalism Studies*, *20*(14), 2014–2033. https://doi.org/10.1080/1461670X.2018.1556320

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. *ArXiv*. https://arxiv.org/abs/2006.01974

Hanitzsch, T., & Quandt, T. (2012). Online journalism in Germany. In E. Siapera & A. Veglis (Eds.), *The handbook of global online journalism* (pp. 429–444). Wiley-Blackwell.

Hartley, J. (2000). Communicative democracy in a redactional society: The future of journalism studies. *Journalism*, *1*(1), 39–48.

Katz, J. E. (2018). Commentary on news and participation through and beyond proprietary platforms in an age of social media. *Media and Communication*, *6*(4), 103–106. https://doi.org/10.17645/mac.v6i4.1743

Kligler-Vilenchik, N. (2018). Why we should keep studying good (and everyday) participation: An analogy to political participation. *Media and Communication*, *6*(4), 111–114. https://doi.org/10.17645/mac.v6i4.1744

Kowert, R. (2020). Dark participation in games. *Frontiers in Psychology*, *11*, 598947. https://doi.org/10.3389/fpsyg.2020.598947

Kuhn, T.S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kunelius, R. (2001). Conversation: A metaphor and a method for better journalism? *Journalism Studies*, *2*(1), 31-54. https://doi.org/10.1080/14616700117091

Kreiss, D., & Brennen, J. S. (2016). Normative theories of digital journalism. In C. W. Anderson, D. Domingo, A. Hermida, & T. Witschge (Eds.), *Sage handbook of digital journalism studies*. Sage.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586–609. https://doi.org/10.1111/jcom.12312

Nordheim, G. V., & Kleinen-von Königslöw, K. (2021). Uninvited dinner guests: A theoretical perspective on the antagonists of journalism based on Serres' Parasite. *Media and Communication*, *9*(1), 88–98. https://doi.org/10.17645/mac.v9i1.3419

Peters, C., & Witschge, T. (2014). From grand narratives of democracy to small expectations of participation. *Journalism Practice*, *9*(1), 19–34. https://doi.org/10.1080/17512786.2014.928455

Quandt, T. (2018). Dark participation. *Media & Communication*, *6*(4), 36–48. https://doi.org/10.17645/mac.v6i4.1519

Quandt, T. (2021). Can we hide in shadows when the times are dark? *Media and Communciation*, *9*(1), 84–87. http://dx.doi.org/10.17645/mac.v9i1.4020

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication, 35*(3), 213–229. https://doi.org/10.1177/0267323120922066

Rosen, J. (2006, 30 June). The people formerly known as the audience. *HuffPost*. https://www.huffpost.com/entry/the-people-formerly-known_b_24113

Santana, A. D. (2016). Controlling the conversation: The availability of commenting forums in online newspapers. *Journalism Studies*, *17*(2), 141–158. https://doi.org/10.1080/1461670X.2014.972076

Serres, M. (2007). *The parasite*. University of Minnesota Press.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *arXiv*. https://arxiv.org/abs/1603.07709

Sjøvaag, H. (2019). *Journalism between the state and the market*. Routledge.

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: Motives and inhibitors to write and read. *Information, Communication & Society*, *18*(7), 798–815. https://doi.org/10.1080/1369118X.2014.997268

Usher, N., & Carlson, M. (2018). The midlife crisis of the network society. *Media & Communication*, *6*(4), 107–110. https://doi.org/10.17645/mac.v6i4.1751

Van Leuven, S., Kruikemeier, S., Lecheler, S., & Hermans, L. (2018). Online and newsworthy: Have online sources changed journalism? *Digital Journalism, 6*(7), 798–806. https://doi.org/10.1080/21670811.2018.1498747

Vujnovic, M., Singer, J. B., Paulussen, S., Heinonen, A., Reich, Z., Quandt, T., Hermida, A., & Domingo, D. (2010). Exploring the political-economic factors of participatory journalism: Views of online journalists in ten countries. *Journalism Practice*, *4*, 285–296. https://doi.org/10.1080/17512781003640588

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe Report, 27*, 1–107. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Wintterlin, F., Langmann, K., Boberg, S., Frischlich, L., Schatz-Eckrodt, T., & Quandt, T. (2021). Lost in the stream? Professional efficacy perceptions of journalists in the context of dark participation. *Journalism,* 1–18. https://doi.org/10.1177/14648849211016984

Wintterlin, F., Schatto-Eckrodt, T., Frischlich, L., Boberg, S., & Quandt, T. (2020). How to cope with dark participation: Moderation practices in German newsrooms. *Digital Journalism, 8*(7), 904–924. https://doi.org/10.1080/21670811.2020.1797519

Ziegele, M., Jost, P., Bormann, M., & Heinbach, D. (2018). Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *SCM Studies in Communication and Media*, *7*(4), 525–554. https://doi.org/10.5771/2192-4007-2018-4-525

Digital
Communication
Research.de

Recommended citation: Masullo, G. M. (2023). Future directions for online incivility research. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 273–286). Digital Communication Research. https://doi.org/10.48541/dcr.v12.16

Abstract: This chapter makes a normative argument that incivility scholars should shift directions in exploring aversive online communication. Specifically, it is vital for scholars to consider various subsets of incivility (e.g., profanity or hate speech), rather than treat incivility as a monolith and to acknowledge that different types are not equally damaging to democracy or interpersonal relations. Furthermore, this chapter calls for more attention to how incivility of all types hurts those from marginalized groups and how and why those with less societal power are more frequent targets of toxicity, as well as how to protect them. It also proposes that the role of online platforms, like Facebook, WeChat, and WhatsApp, be integorrated more fully in regard to incivility and that incivility be studied in concert with other types of problematic speech, such as misinformation and disinformation.

DOI 10.48541/dcr.v12.16

Gina M. Masullo

Future Directions for Online Incivility Research

1 Rethinking incivility research

Ask average people what they think of online comments on news websites or social media, and soon enough you are likely to hear the common adage: "Don't read the comments." The thinking is that online comments are vast cesspools of vitriol. The implication is that the incivility that peppers these comments renders them useless for political discourse. Scholars all too often have adopted a similar approach, highlighting a clear normative assumption that "incivility is bad and should be eliminated" (Chen, Muddimann, et al., 2019, p. 1).

This chapter puts forth a theoretical argument to challenge this notion. While incivility is certainly rife online (e.g., Chen, 2017; Coe et al., 2014), clear evidence exists that various types of incivility are not equally damaging to political discussions (Rossini, 2020) or perceived as such (Stryker et al., 2016). Indeed, when incivility is defined as impoliteness, research suggests it may actually jump-start political engagement at least in the short term (Chen, 2017; Chen & Lu, 2017) even as it foments negative emotions (e.g., Gervais, 2015; Rösner et al., 2016). Thus, understanding the role and influence of incivility in online discussions is decidedly complicated. As the study of incivility has progressed and our understanding has grown, we must shift to asking new questions, considering different outcome

variables, and abandoning old assumptions. I put forth two principles that undergird the approach that I am calling for: Specify what type of incivility you are considering, rather than treating incivility as a monolith, and realize that different types of incivility are not equally damaging to democracy or interpersonal relations (Stryker et al., 2016; Chen, Muddimann, et al., 2019).

Keeping these two principles in mind, I call for an expanded online incivility research agenda with a broader vantage point. Instead of asking—what is the influence of uncivil discourse? – we should ask more specific questions. How do different types of uncivil content, such as pejorative speech versus profanity, differ in their harm? Instead of assuming incivility will have negative effects, we should consider positive outcomes such as a boost in political engagement (e.g., Chen, 2017; Chen & Lu, 2017) and parse how they vary between different subsets of people or across platforms. Rather than focus our inquiry on the more common types of aversive speech, profanity and insults (Chen, 2017; Coe et al., 2014), we should delve more deeply into the effects of the less frequent but more antagonistic types of discourse, such as hateful or intolerant speech (Rossini, 2020). I argue that hate speech fits under the umbrella term of what scholars and the public label incivility (see Chen, 2017, for an overview of this arugment), although certainly some scholars see hate speech as conceptually distinct (see Paasch-Colberg et al., 2021, for an overview; see Sponholz and Frischlich in this volume, for a discussion of the hate speech concept).

Specifically, I urge that incivility research be expanded in three main areas, which I will discuss below. First, more attention should be paid to how incivility of all types hurts those from marginalized groups and how and why those with less societal power are more frequent targets of toxicity, as well as how to protect them. This approach puts more research emphasis on hate speech, arguably the most virulent type of incivility, rather than impoliteness, the least antagonistic type. Second, we should interrogate the role of online platforms, such as Facebook, Twitter, WeChat, and WhatsApp, in managing uncivil attacks. Finally, more research should probe how incivility intersects with and perhaps amplifies other problematic forms of online communication, such as misinformation and disinformation. While misinformation and disinformation are clearly problematic types of communication, I argue they are conceptually distinct from incivility because their most potent effect is in misleading the public, which is not part of incivility. Incivility and false information warrant study together because they are arguably the two major issues that trouble scholars and the public alike about online discussions.

2 Online incivility is not a monolith

First, I will unpack the two principles I have outlined that should be foundational to online incivility research going forward. Then I will examine the three areas for expanded research in greater detail. The first principle I put forth is that scholars should examine specific types of aversive online content, rather than treat incivility as a monolith. There is great debate in the literature over what constitutes uncivil communication although most definitions fall into two main camps (see Bormann & Ziegele in this volume, for an overview of different incivility conceptions). In the impoliteness camp, incivility is defined as profanity, name-calling, and insults (Chen, 2017; Muddiman, 2017; Rossini, 2020). This approach often relies on impoliteness or face theories, which argue that uncivil speech threatens people's constructed sense of self, called face, leading to emotional pain (Brown & Levinson, 1987; Goffman, 1959; Metts & Cupach, 2015). The other approach defines incivility more virulently as threats to democracy; stereotyping; or racists, sexist, xenophobic, and homophobic communication (Papacharissi, 2004), a subcategory that Rossini (2020) calls "intolerant discourse" (p. 2). This type of incivility is rooted in the theory of deliberative democracy, which relies on the normative ideal that discussions across differences should be rational, respectful exchanges that seek to reach consensus and are, therefore, valuable in a democracy (Fishkin, 1991; Gutmann & Thompson, 1996; Jacobs et al., 2009). These ways of considering incivility offer distinct theoretical and operational approaches that should not be conflated. Even more importantly, other types of content that might fit definitions of incivility—such as lying accusations (Kenski et al., 2018), hyperbole and distortion (Gervais, 2015; 2017), and lack of political compromise (Muddiman, 2017) – that have received scant study should receive more attention. Indeed, Stryker et al. (2016) considered 23 types of behavior or speech that might be considered uncivil, including vulgarity, refusing to listen to others, and shouting, and found that slurs and threatening or encouraging harm were perceived as most uncivil. Yet, study after study, including some of my own (e.g., Chen, 2017; Chen & Lu, 2017), focus on forms of rudeness (e.g., Lee et al., 2019; Rösner et al., 2016), the type of incivility that Stryker et al. (2016) found was perceived as less damaging to political discourse. Our efforts should move to a multi-dimensional approach to incivility when possible (e.g., Chen et al., 2020; Oz et al., 2018; Ozler et al., 2020; Ziegele et al., 2020; see also Bormann & Ziegele

in this volume). I would even go so far as to suggest we should stop calling everything incivility and, rather, use the specific terms (e.g., "hate speech" or "profanity") that more narrowly pertain to what we mean. Our research then should focus on these more extreme types of speech that offer more troubling implications (e.g., Chen, Fadnis, & Whipple, 2019; Murthy & Sharma, 2019; Paasch-Colberg et al., 2021). Of course, from a practical standpoint, we may need to retain the concept of incivility as an umbrella term for all these types of communication and as a theoretical perspective, but, when possible, we should use more specific terminology. For example, studies that examine specific types of speech, such as obscenity or politically motivated hate speech (e.g., Bodrunova et al., 2021), offer more knowledge than those that aim to tackle *incivility* in general.

In the earlier days of incivility research, methodological issues may have led to reductive operational definitions for incivility. For example, experiments require that researchers focus only on few forms of incivility because manipulating several types of incivility would require exhaustively large sample sizes. Technological limitations initially meant that efforts to automate detection of incivility using machine learning could only capture less-nuanced attributes of incivility, such as profanity (e.g., Lee et al., 2019). Human coders were often used to better detect subtle uncivil attributes (Guo et al., 2016). However, this was expensive and time-consuming and limited the amount of content that could be reasonably analyzed (Muddiman et al., 2019). Newer approaches, such as using manually validated organic dictionaries (Guo et al., 2016; Muddiman et al., 2019), and machine learning models have shown success in detecting multi-faceted forms of incivility across various domains, such as comments and tweets (Davidson et al., 2020; Ozler et al., 2020), although many still misclassify complex types of incivility (e.g., Stoll et al., 2020). Even when human coders are employed for smaller datasets (e.g., Chen et al., 2020; Oz et al., 2018) or in combination with automated coding (Kenski et al., 2018), efforts should be made to consider and compare different attributes of incivility. Experiments, of course, may still focus on only several types of incivility, and that is fine, as long as these types are identified and some experiments delve into the more virulent types of incivility.

These methodological issues suggest that we need to expand how we explore incivility, employing both quantitative and qualitiative approaches to provide a fuller understanding of how the public perceives incivility and the effects it has. In-depth interviews with the public or with content moderators can provide insights into

how incivility is perceived and identified, for example, in ways where quantitative coding may fall short. Theoretical work can forge connections between disparate quantitative or qualitative studies that is woefully lacking.

3 All incivility is not equally harmful

The second underlying principle is that different types of online incivility are not equally harmful. When communication is considered dichotomously—either uncivil or civil—there can be a tendency to view one as always good and one as always bad. This is a flawed assumption (Chen, Muddimann, et al., 2019; Rossini, 2020). We need more research that asks questions that tap into more subtle questions: Under what conditions is incivility harmful? What attributes of incivility are more harmful than others? Are there situations where civility might not be beneficial? We need more research that offers insight into how people experience different types of incivility and what it means for public discourse. What effect on political discourse or emotions does profanity have that differ from the effects of threats to democracy or stereotyping? Are there particular subgroups of the population that encounter greater or lesser effects? These questions need to be addressed. Rather than look at incivility as always harmful and civility as always righteous, we need to understand the overlap between good and bad. Deciding what is civil or uncivil, Herbst (2010) argues, is a "strategic tool or weapon in politics" (p. 6), such that those in power can squelch speech they disagree with by labeling it as uncivil. It is temporary and changeable and fluid across contexts, she posits (see Litvinenko in this volume, for a similar argument). If that is true, and I believe it is, clearly incivility is also malleable. Yet, the literature often assumes that online incivility by default is harmful. In addition, we need more research outside the United States and other western democracies. We have limited evidence that people perceive types of incivility differently across cultures and countries (Weber et al., 2020), but more cross-cultural work in this area is direly needed to have a fuller understand of online incivility's effects on discourse and society.

4 Marginalized groups

I have outlined the two main principles that should be kept in focus in incivility research going forward. Now I shift my attention to the three areas that I suggest need greater attention. The first is that we need more research about how online incivility of all types disproportionately targets those from marginalized groups, and, even more importantly, how people from these groups can be more protected online. For example, we know that women and people of color and other marginalized groups are frequent targets of incivility online (Chen, Fadnis, & Whipple, 2019; Chen et al., 2020; Edström, 2016; Ferrier, 2018; Pain & Chen, 2019; Searles et al., 2020; Sobieraj, 2020; Southern & Harmer, 2021), but we know little about what interventions are most effective to safeguard them. Journalists provide a cogent example. We know that female journalists face a barrage of online attacks (Chen et al., 2020; Searles et al., 2020) from threatening messages to the unauthorized release of private information, and that this sometimes compels them to change how they tell stories or even consider leaving the profession (Ferrier, 2018). Politicians offer another notable example. While all politicians run the risk of being attacked online, barbs against people identifying as female are more likely to be gendered or stereotypical assaults on their identity (Southern & Hamer, 2021). But it is not just journalists or politicians. All women who dare to participate in online public discourse, particularly about politics, face the threat of violent resistance: "The abuse targets their identities, pummeling them with rape threats, attacks on their appearance, and presumed sexual behavior, and a cacophony of misogynistic, racist, xenophobic, and homophobic stereotypes and epithets" (Sobieraj, 2020, p. 4). Expanding incivility research to more countries and cultures will help address this in some ways, but we also need more studies that specifically focus on marginalized groups, such as people of color, refugees, or LGBTIA+ people. We need more research into how the digital space can be changed or managed at a structural level to prevent this. What roles should newsrooms, platforms, or government play in solving these problems? Can existing laws be better employed or are new laws necessary? How can newsrooms help their employees feel safe? This is important because understanding how to protect the marginalized online will help achieve a more user-friendly digital space for everyone. The strategies that work for those with less power in society will improve discourse for all.

5 Role of platforms

The second line of research that I argue deserves more focused attention is what role platforms should play in managing uncivil attacks online. Currently, a patchwork of efforts aim to ensure a productive commenting space on social media and news websites. Moderators police content, and these efforts improve discussions (Ksiazek, 2018; Masullo et al., 2020; Stroud et al., 2015), but the task is emotionally exhausting (Riedl et al., 2020). Users flag unseemly content (Naab et al., 2018) or even dialog with commenters in hopes of improving discussions (Ziegele et al., 2020). Platforms and news organizations are performative by telling users in advance through terms of service or online posts what type of behavior and content will be permitted (Gillespie, 2018). Yet, despite all these efforts, calls are frequent that more should be done (see Sobieraj et al., 2020, for related arguments). Are platforms or newsrooms responsible for ensuring a robust democratic discourse can take place on their sites? Is it right or ethical for privately owned companies to take on this role? Should governments regulate platforms as public utilities to ensure they do this task? Will that improve discussions? Does that force platforms into a role they shouldn't have? How would that even work, considering many platforms cross national boundaries? These questions need research-based answers. This is particularly true at this current moment as some social media platforms took the unprecedented step of banning former U.S. President Donald J. Trump because his combative posts culminated with a mob of his supporters violently attempting to prevent Congress from certifying Joseph Biden as the victor over Trump (Denham, 2021). Regardless of how scholars feel about this particular banning, the banning raised urgent questions about internet governance and the role and power of social media platforms in contemporary lives and highlights the need for more study in this area. It leads to important questions, such as: Who really controls speech? What entities should have the power to govern online discourse? What are the ramifications of banning politicians, or others, from engaging online? What are the ethical and legal questions surrounding such bans? All are ripe area for more inquiry.

6 Online incivility and other problematic discourse

The third and final avenue for research on online incivility that I propose is understanding the intersection between incivility and false information online. These two concepts should be studied together because they are arguably the two major issues that scholars and the public raise about online discourse, and we know that in general incivility can undermine the persuasiveness of communication (Chen & Ng, 2016; Jenkins & Dragojevic, 2013). Yet, for the most part incivility and false information are treated as separate research streams. The problem of misinformation, the unintentional spread of false information, and disinformation, the purposeful spread of inaccurate communication (Tandoc et al., 2018), are focal concerns in the public consciousness. We know correcting misperceptions from faulty information, whether purposeful or not, is challenging (Nyhan & Reifler, 2010) but possible (e.g., Bode & Vraga, 2015), but we know less about how incivility may influence how people process false information or corrections of that information. This is important to understand because, given the free-wheeling discourse online, it seems logical the people may come across uncivil misinformation or disinformation or acerbic corrections to this false information. Bode et al. (2020) have shown in an experiment that corrections to misleading information about the safety of raw milk on Facebook are effective regardless of whether the tone of the correction is uncivil, affirmational, or neutral, suggesting tone is not the driving force in whether people embrace or reject a correction to misinformation. But more research in this area is warranted. Kim and Chen (2021) demonstrated that angry emoticons on social media comments that attempt to correct misinformation altered how those messages are perceived. Yet, so many questions remain unanswered. Do people discount misinformation that is uncivil or is it more arousing and, therefore, more powerful? Do people reject or embrace uncivil disinformation that outrages them morally, such as accusations that politicians are not telling the truth? Does how people respond to these messages depend on whether an out-group or in-groups is spreading the falsehood or whether a person realizes the message is not true? Does this vary based on what type of false information is considered? Given the monumental concern that false information and incivility present online, it is vital to understand more about how these two concepts intersect.

7 Going forward

In summary, there is a fruitful path ahead for the study of online incivility. While the topic has received a plethora of study, there are holes in the literature that scholars should fill. We have a firm foundation of incivility research at this point, but research going forward should focus more specifically on differences between various types of incivility, rather than treat it as a monolith. Also, the days of seeing incivility as always normatively bad and civility as always normatively good should be over, I urge. We need to consider different types of harm for different types of incivility, and also leave open the idea that incivility may have benefits even if they are unintended. Research going forward should look to finding solutions, not just illuminating problems. In particular, we need more work on how the digital space can be improved so that it is safer for those from marginalized groups. We need more study into how platforms can and should manage incivility and what ramifications their actions have on the larger society. Finally, we need to consider incivility in concert with other forms of aversive online communication, such as misinformation and disinformation.

These trajectories of research will offer many benefits. First, they will bring a richer, more nuanced understanding of how online incivility operates and its effects. It will help us theorize more about the role of incivility in society, and it will help us solve problems related to acerbic online communication more broadly. I cannot imagine a day anytime soon when online communication will disappear. If anything, we likely will be communicating more and more online than we do today. That means online incivility is with us in the future, so the need to bring incivility research to a new level is particularly important.

Gina M. Masullo is Associate Professor in the School of Journalism and Media and Associate Director of the Center for Media Engagement, both at The University of Texas at Austin, USA. https://orcid.org/0000-0002-4909-2116

References

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, *65*(4), 619–638. https://doi.org/10.1111/jcom.12166

Bode, L., Vraga, E. K., & Tully, M. (2020). Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-026

Bodrunova, S. S., Litvinenko, A., Blekanov, I., & Nepiyushchikh, D. (2021). Constructive aggression? Multiple roles of aggressive content in political discourse on Russian YouTube. *Media and Communication*, *9*(1), 181–194. https://doi.org/10.17645/mac.v9i1.3469

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Palgrave Macmillan.

Chen, G. M., Fadnis, D., & Whipple, K. (2019). Can we talk about race? Exploring online comments about race-related shootings. *Howard Journal of Communications*, *31*(1), 35–49. https://doi.org/10.1080/10646175.2019.1590256

Chen, G. M., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media, 61*(1), 108–125. https://doi.org/10.1080/08838151.2016.1273922

Chen, G. M., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, *5*(3). https://doi.org/10.1177/2056305119862641

Chen, G. M., & Ng, Y. M. M. (2016). Third-person perception of online comments: Civil ones persuade you more than me. *Computers in Human Behavior, 55*, *Part B*, 736–742. https://doi.org/10.1016/j.chb.2015.10.014

Chen, G. M., Pain, P., Chen, V., Mekelburg, M., & Springer, N. (2020). 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, *21*(7), 877–895. https://doi.org/10.1177/1464884918768500

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a new classifier for automated identification of incivility in social media. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 95–101. Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.alw-1.12

Denham, H. (2021, January 14). These are the platforms that have banned Trump and his allies. *The Washington Post*. https://www.washingtonpost.com/technology/2021/01/11/trump-banned-social-media/

Edström, M. (2016). The trolls disappear in the light: Swedish experiences of mediated sexualised hate speech in the aftermath of Behring Breivik. *International Journal for Crime, Justice and Social Democracy*, *5*(2), 96–106. https://doi.org/10.5204/ijcjsd.v5i2.314

Ferrier, M. (2018). Attacks and harassment: The impact on female journalists and their reporting. International Women's Media Foundation. https://www.iwmf.org/attacks-and-harassment/

Fishkin, J. S. (1991). *Democracy and deliberation: New directions for democratic reform*. Yale University Press.

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, *12*(2), 167–185. https://doi.org/10.1080/19331681.2014.997416

Gervais, B. T. (2017). More than mimicry? The role of anger in uncivil reactions to elite political incivility. *International Journal of Public Opinion Research*, *29*(3), 384–405. https://doi.org/10.1093/ijpor/edw010

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

Goffman, E. (1959). *Presentation of self in everyday life*. Doubleday.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359. https://doi.org/10.1177/1077699016639231

Gutmann, A., & Thompson, D. (1996). *Democracy and disagreement.* Harvard University Press.

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.

Jacobs, L. R., Cook, F. L., & Delli Caprini, M. X. (2009). *Talking together: Public deliberation and political participation in America*. The University of Chicago Press.

Jenkins, M., & Dragojevic, M. (2013). Explaining the process of resistance to persuasion: a politeness-theory based approach. *Communication Research, 40*(4), 559–590. https://doi.org/10.1177/0093650211420136

Kenski, K., Filer, C. R., & Conway-Silva, B. A. (2018). Lying, Liars, and Lies: Incivility in 2016 presidential candidate and campaign tweets during the invisible primary. *American Behavioral Scientist*, *62*(3), 286–299. https://doi.org/10.1177/0002764217724840

Kim, J. W., & Chen, G. M. (2021). Exploring the influence of comment tone and content in response to misinformation in social media news. *Journalism Practice*, *15*(4), 456–470. https://doi.org/10.1080/17512786.2020.1739550

Ksiazek T. B. (2018). Commenting on the news. *Journalism Studies, 19*(5), 650–673. https://doi.org/10.1080/1461670X.2016.1209977

Lee, F. L. F., Liang, H., & Tang, G. K. Y. (2019). Online incivility, cyberbalkanization, and the dynamics of opinion polarization during and after a mass protest event. *International Journal of Communication*, *13*, 4940–4959.

Masullo, G. M., Riedl, M. J., & Huang, Q. E. (2020). Engagement moderation: What journalists should say to improve online discussions. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2020.1808858

Metts, S., & Cupach, W. R. (2015). Face theory: Goffman's dramatist approach to interpersonal interaction. In D. O. Braithwaite & P. Schrodt (Eds.), *Engaging theories in interpersonal communication* (pp. 203–214). Sage.

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*, 3182–3202.

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (Re)Claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, *36*(2), 214–226. https://doi.org/10.1080/10584609.2018.1517843

Murthy, D., & Sharma, S. (2019). Visualizing YouTube's comment space: Online hostility as a networked phenomena. *New Media & Society*, *21*(1), 191–213. https://doi.org/10.1177/1461444818792393

Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*(2), 777–795. https://doi.org/10.1177/1461444816670923

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society, 20*(9), 3400–3419. https://doi.org/10.1177/1461444817749516

Ozler, K. B., Kenski, K., Rains, S., Shmargad, Y., Coe, K., & Bethard, S. (2020). Fine-tuning for multi-domain and multi-label uncivil language detection. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 28–33. Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.alw-1.4

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180. https://doi.org/10.17645/mac.v9i1.3399

Pain, P., & Chen, V. (2019). This reporter is so ugly, how can she appear on TV? *Journalism Practice*, *13*(2), 140–158. https://doi.org/10.1080/17512786.2017.1423236

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Riedl, M. J., Masullo, G. M., & Whipple, K. N. (2020). The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior*, *107*, Article 106262. https://doi.org/10.1016/j.chb.2020.106262

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior*, *58*, 461–470. https://doi.org/10.1016/j.chb.2016.01.022

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*. Advance online publication. https://doi.org/10.1177/0093650220921314

Searles, K., Spencer, S., & Duru, A. (2020). Don't read the comments: The effects of abusive comments on perceptions of women authors' credibility. *Information, Communication & Society*, *23*(7), 947–962. https://doi.org/10.1080/1369118X.2018.1534985

Sobieraj, S. (2020). *Credible threat: Attacks against women online and the future of democracy*. Oxford University Press.

Sobieraj, S., Masullo, G. M., Cohen, P. N., Gillespie, T., & Jackson, S. J. (2020). Politicians, social media, and digital publics: Old rights, new terrain. *American Behavioral Scientist*, *64*(11), 1646–1669. https://doi.org/10.1177/0002764220945357

Southern, R., & Harmer, E. (2021). Twitter, incivility and "everyday" gendered othering: An analysis of tweets sent to UK members of parliament. *Social Science Computer Review, 39*(2), 259–275. https://doi.org/10.1177/0894439319865519

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research, 2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-Mediated Communication*, *20*(2), 188–203. https://doi.org/10.1111/jcc4.12104

Stryker, R., Conway, B. A., & Danielson, J. T. (2016). What is political incivility? *Communication Monographs*, *83*(4), 535–556. https://doi.org/10.1080/03637751.2016.1201207

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news." *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Weber, I., Laban, A., Masullo, G. M., Gonçalves, J., Torres da Silva, M., & Hofhuis, J. (2020). International perspectives on what's considered hateful or profane online. Center for Media Engagement. https://mediaengagement.org/research/perspectives-on-online-profanity

Ziegele, M., Naab, T. K., & Jost, P. (2020). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, *22*(5), 731–751. https://doi.org/10.1177/1461444819870130

III. METHODOLOGICAL PERSPECTIVES: OPERATIONALIZATION, AUTOMATION AND DATA

Digital
Communication
Research.de

Recommended citation: Bahador, B. (2023). Monitoring hate speech and the limits of current definition. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 291–298). Digital Communication Research. https://doi.org/10.48541/dcr.v12.17

Abstract: Current definitions of hate speech are inadequate as the basis for monitoring hate speech targeted at groups. First, they do not capture escalating group-targeted negative speech which can be a precursor to more extreme forms of hate speech such as dehumanization, demonization, and incitement to violence. While not hate speech, such negative speech is an early warning that could be helpful for a hate speech monitoring system to track, as responses and interventions, especially to the offline harms of hate speech, can take time to operationalize. Second, current definitions of hate speech do not capture hateful rhetoric aimed at groups not traditionally included in hate speech definitions (those without immutable qualities), such as groups targeted for hate based on profession-based identity like journalists. This chapter presents some suggestions for addressing these issues, including a hate speech intensity scale.

DOI 10.48541/dcr.v12.17

Babak Bahador

Monitoring Hate Speech and the Limits of Current Definition

1 Introduction

Monitoring hate speech in order to prevent offline harms is a laudable goal that has proven largely elusive to date. This is due to a number of factors including the limits of current definitions, knowing exactly when such speech triggers offline harms, tracking hate speech in real time and creating and implementing effective interventions. This article is primarily focused on the first issue regarding how current definitions of hate speech can limit effective hate speech monitoring. The article begins by examining how hate speech is typically currently defined and some limitations that this poses for monitoring due to the restricted scope of the groups and language included. The article argues that these limits mean that some non-traditonal types of groups targeted with hate are excluded, even though they could also become victims of hate-based violence. Furthermore, it argues that hate speech monitoring should include language escalating to traditional hate speech content so early warning signs can be detected and action taken earlier. Once more extreme hate speech takes hold, it could also be a sign that it is too late to implement more peaceful preventative actions. Finally, the article introduces an hate speech intensity scale that includes early warning categories for hate speech monitoring.

2 The limits of defining hate speech

In its most blantant manifestations, hate speech is communication aimed at groups of people to dehumanize, demonize or incite violence against them. If hate against a particular out-group (the group targeted for hate speech) is successfully sold to an in-group (the group the hate speech attempts to persuade), then all members of the out-group are viewed as a negative stereotype, losing their individuality and humanity. In such scenarios, which are often driven by falsehoods and exaggerated fears about the out-group, retribution against all members of the out-group is justified as they all represent the same threat to the in-group (Bahador, 2012).

Most definitions of hate speech limit its targets to groups that hold immutable qualities such as a particular race, nationality, religion, ethnicity, gender, age bracket or sexual orientation (see Sponholz in this volume). However, in research that measured hate aimed at groups more broadly, findings showed that groups with immutable qualities were less frequently targeted versus other types of groups not usually included in hate speech definitions (Bahador, Kerchner et al., 2019). The first of these types of groups can be classified as professions and industries[1], with journalists and the media sector a primary and leading example. While professions and employment in particular industries is by choice (so not immutable), they nonetheless are groups that are distinguishable and a growing target of hate speech. The concern here is less about harassment of journalists for particular content they produce, but about attacks based on group identity, which makes it similar to other groups with immutable qualities.

Hate speech against journalists has grown notably over recent years, not least because journalists act as a check on authoritarian power, which has been a growing trend worldwide (Sulzberger, 2019). This is particularly the case for female journalists, who are under unprecedented levels of attack online and are targeted both for their gender and profession (The Guardian, 2021).

Another notable group typology frequently targeted for hateful rhetoric is foreign countries. While this is related to the traditional immutable category of nationality (e.g., Chinese), there are often negative references to the country

1 This definition excludes professions and industries that engage in violent or other malicious behavior.

itself (e.g., China) that can increase negative public sentiment towards the country (Brewer et al., 2003), and most importantly, its people and those associated with them. For example, references to China as the actor responsible for the Covid-19 global pandemic by certain political leaders is considered to be a key factor that led to a spike in hate crimes against Asians in the United States including American citizens from Asian descent (BBC News, 2021). Those political leaders did not mention anything about Chinese people or Americans of Asian descent. They mentioned China, and this is excluded from traditional definitions of hate speech (as it is a country, not race or nationality), showing that the current definitions are inadequate for addressing a serious problem.

While it is certainly appropriate to criticize foreign governments, those in influential positions, such as journalists and politicians, often inadvertently refer to the country without sufficiently delineating that their critique is of the government and its actions and policies, not the people. To avoid such conflation and its potential negative effects, those in power need to be precise and only refer to the foreign governments, government institutions and agencies or political leaders and not the country as a whole. When criticising states such as the United States or Russia, the criticism should be against the government specifically, for example, "the Russian government attacked," not "Russia attacked"; or "the U.S. military bombed this," not "the U.S. bombed this." The latter cases build hate towards people associated with the country; the former offer legitimate criticism of the government which should be subject to scrutiny.

In this research, four different hate-speech group typologies are distinguished: 1) immutable qualities (traditional hate speech groups), 2) professions and industries, 3) countries, and 4) "other," which captures other groups of people who are targeted for hate speech but otherwise excluded from the other three categories. These include groups such as "the elite," which are generally excluded in traditional definitions.[2] If one wants to capture the full breadth of hate-speech group targets, this more extensive approach will capture more hate speech.

2 Terms such as the "the elite" often refer to a variety of other groups but may mean different groups to different audiences. For example, in a survey of Americans, conservatives often considered the elite to be cultural, political and academic elite such as actors, politicians and professors, while liberal Americans thought elites were economic, industrial and financial elites (Bahador, Entman & Knüpfer, 2019).

3 Early warning to hate speech

When considering what type messaging should be considered hate speech, most definitions include language that attemts to demonize, dehumanize or incite violence against groups. Dehumanization is a tactic that depicts groups as less than human and usually involves associating them with sub-human creatures such as rats and cockroaches or non-human forms such as garbage or dirt. Alternatively, groups can be demonized, in which they become threatening super-human creatures such as monsters and demons, or equated to fatal threats such as cancer or a virus. When presented this way, the elimination of such groups becomes beneficial and desirable, as removing them takes away a perceived threat to the in-group (Dower, 1986; Keen, 1991; Carruthers, 2011; Bahador 2015). Calls to attack, harm or kill groups—often the same ones that were dehumanized and demonized—is incitement, which is another central type of hate speech. Incitement is often the most extreme type of hate speech content. Even in the United States, in which hate speech is generally protected on free speech grounds under the First Amendment of the Constitution, it is still a crime to incite "immiment lawless action" if likely to occur within a short period of time (Tucker, 2015).

As with the previous concern over limiting hate speech groups to only ones with immutable qualities (and missing other groups that are also subject to hate), it is also problematic to restrict hate speech definitions to only the most severe types (dehumanization, demonization and incitement) if other speech builds up hate and disdain towards groups. Hate in the context of hate speech, after all, can be defined as a human emotion that is triggered through exposure to a particular type of information. When it emerges, this emotion involves an enduring dislike for a group, a loss of empathy for them, and a desire for harm against them (Waltman & Mattheis, 2017). However, there is no reason to assume that other types of negative speech against groups also do not create hate. At its root, hate against groups begins when an us-them dynamic is created and a different group is differentiated from your own and negative actions and characteristics are allocated to them, coming over time to define the group and all its members as a negative stereotype. But even if negative words towards groups such as insults do not constitute hate speech, it is an early warning that should be addressed before it becomes acceptable and builds tolerance for more extreme forms of speech. In any hate speech monitoring system, it is important to have early warning and

not just activate the system when it has become dangerous. As such, incorporating language that can be considered as early warning on the road to hate speech should be an important consideration of any monitoring system.

4 Hate speech intensity scale

To monitor hate speech in a way that incorporates both a broader set of group categories and an early warning element, a hate speech intensity scale is proposed, demonstrating the escalating nature of hate speech content. To make it easy to follow, colors, numbers, titles, descriptions and examples are provided in Table 1. Furthermore, a distinction is made between how groups are characterized (referred to as "rhetoric") and "responses" or actions the in-group is recommended to take against the out-group.

The hate speech intensity scale has six categories. Categories 1 to 3 are referred to as "early warning." Category 4 is dehumanization/demonization, and categories 5 and 6 are associations with or calls for violence (#5) and death (#6). The following section goes through the 6 categories in more detail.

In this scale, the first early warning category involves disagreement with groups. While there is nothing wrong with disagreement in principle, and it can be argued that it is essential to democracy, the exercise does involve the creation of an us versus them dynamic, with "them" being viewed collectively in a negative light. Also, there is likely some misinformation involved in such a claim against a group, as rarely will an entire group hold similar views and beliefs. Collectivizing their views or beliefs, therefore, is likely to miss important differences amongst group members. By itself, such rhetoric is likely not hateful, and thus, the green color designation indicating it is safe to proceed (as per a traffic light). However, it is something to start monitoring for the purposes of a monitoring system.

The second early warning category involves language that blames a group for particular negative actions, often carried out by one or a few members. However, there is a tendency to blame the entire out-group in such scenarios for the negative actions of a few. This category includes non-violent negative actions, such as claims that the group stole or withdrew from a positive event. When such alleged actions are ambiguous on the use of violence (e.g., they stopped them) or use of non-violent

negative metaphors, they fit in this category (if unambiguous on violence, it's classified a 5). Responses involve non-violent actions the in-group should do towards the out-group, such as voting them out or protesting against them.

Table 1: Hate speech intensity scale

	Title	Description	Examples
	6. Death	Rhetoric includes literal killing by group. Responses include the literal death/elimination of a group.	Killed, annihilate, destroy
	5. Violence	Rhetoric includes infliction of physical harm or metaphoric/ aspirational physical harm or death. Responses include calls for literal violence or metaphoric/aspirational physical harm or death.	Punched, raped, starved, torturing, mugging
	4. Demonizing and Dehumanizing	Rhetoric includes sub-human and superhuman characteristics. There are no responses for #4.	Rat, monkey, Nazi, demon, cancer, monster
	3. Negative character	Rhetoric includes non-violent characterizations and insults. There are no responses for #3.	Stupid, thief, aggressor, fake, crazy
	2. Negative actions	Rhetoric includes negative non-violent actions associated with the group. Responses include non-violent actions including metaphors.	Threatened, stole, outrageous act, poor treatment, alienate
	1. Disagreement	Rhetoric includes disagreeing at the idea/belief level. Responses include challenging claims, ideas, beliefs, or trying to change their view.	False, incorrect, wrong, challenge, persuade, change minds

The third early warning typology includes negative characterizations or insults. This is worse than just negative non-violent actions, as it makes an intrinsic claim about the group as opposed to a one-off action claim. As this category is not action oriented (unlike #1, 2, 5 and 6), there are no responses. The fourth category

is also the second typology and can be considered an extreme form of negative characterization involving dehumanization and/or demonization. Like the third category, there are no responses in this category.

The fifth and sixth categories are part of the third and most intense typology, involving violent actions and death. The fifth category refers to literal violence allocated to the out-group either in their past, present or future. This also includes metaphorical or aspirational violence that is either nonlethal or lethal. Responses call for literal non-lethal violence towards the out-group such as assaulting them. The sixth category involves rhetorically referring to the out-group as killers (past, present and future). Responses call for our side to kill the out-group.

To monitor hate speech effectively, it is important to not miss any notable groups targeted for hate (such as journalists), even if they don't fit into traditional hate speech definitions. Furthermore, it is critical to see how hate speech builds up before it starts to be more harmful with stronger rhetoric. To this end, the hate speech intensity scale offers a tool that could be operationalized to monitor hate speech. In early experimentation using this tool to monitor hate speech from leading media personalities in the U.S., we found that about half of all hate speech is against journalists and the media (Bahador, Kerchner et al., 2019). We also found few examples of more extreme speech (#4, 5 and 6 on the scale) representing less than 5% of all cases using this scale. However, #2 and 3 (negative actions and characteristics allocated to groups) were prevalent, accounting the vast majority of cases.

Babak Bahador is Research Professor at the School of Media and Public Affairs (SMPA) at George Washington University, USA, and a Senior Fellow at the University of Canterbury in New Zealand. https://orcid.org/0000-0001-7872-9764

References

Bahador, B. (2012). Rehumanizing enemy images: Media framing from war to peace. In K. V. Korostelina (Ed.), *Forming a culture of peace: Reframing narratives of intergroup relations, equity and justice* (pp. 195–211). Palgrave Macmillan.

Bahador, B. (2015). The media and deconstruction of the enemy image. In V. Hawkins & L. Hoffmann (Eds.), *Communication and peace: Mapping an emerging field* (pp. 120–132). Routledge.

Bahador, B., Kerchner, D., Bacon, L., & Menas, A. (2019). Monitoring hate speech in the US Media. Working Paper. Media and Peacebuilding Project. George Washington University. https://cpb-us-e1.wpmucdn.com/blogs.gwu.edu/dist/8/846/files/2019/03/Monitoring-Hate-Speech-in-the-US-Media-3_22-z0h5kk.pdf

Bahador, B., Entman, R., & Knüpfer C. (2019). Who's elite and how the answer matters to politics. *Political Communication, 36*(1), 195–202. https://doi.org/10.1080/10584609.2018.1548412

BBC News (2021, May 21). Covid 'hate crimes' against Asian Americans on rise. https://www.bbc.com/news/world-us-canada-56218684

Brewer, P. R., Joseph, G., & Willnat, L. (2003). Priming or framing: Media influence on attitudes toward foreign countries. *International Communication Gazette, 65*(6), 493–508. https://doi.org/10.1177/0016549203065006005

Carruthers, S. (2011). *The media and war*. Palgrave MacMillan.

Dower, J. W. (1986). *War without mercy: Race and power in the Pacific War*. Pantheon Books.

Keen, S. (1991). *Faces of the enemy: Reflections on the hostile imagination*. Harper & Row.

Sulzberger, A. G. (2019, September 23). The growing threat to journalism around the world. *New York Times*. https://www.nytimes.com/2019/09/23/opinion/press-freedom-arthur-sulzberger.html

The Guardian (2021, May 9). The Guardian view on online abuse of female journalists: a problem for all. *Editorial*. https://www.theguardian.com/commentisfree/2021/may/09/the-guardian-view-on-online-abuse-of-female-journalists-a-problem-for-all

Tucker, E. (2015, December 31). How federal law draws a line between freedom of speech and hate crimes. *PBS News Hour*. https://www.pbs.org/newshour/nation/how-federal-law-draws-a-line-between-free-speech-and-hate-crimes

Waltman, M. S., & Mattheis, A. (2017). Understanding hate speech. In *Oxford encyclopedia of communication*. https://doi.org/10.1093/acrefore/9780190228613.013.422

Digital
Communication
Research.de

Recommended citation: Laaksonen, S.-M. (2023). The datafication of hate speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 301–317). Digital Communication Research. https://doi.org/10.48541/dcr.v12.18

Abstract: Hate speech has been identified as a pressing problem in society, and several automated approaches have been designed to detect and prevent it. This chapter reflects on the operationalizations, transformations, and reductions required by the datafication of hate to build such an automated system. The observations are based on an action research setting during a hate speech monitoring project conducted in a multi-organizational collaboration during the Finnish municipal elections in 2017. The project developed an adequately well-working algorithmic solution using supervised machine learning. However, the automated approach requires heavy simplification, such as using rudimentary scales for classifying hate speech and relying on word-based approaches, while in reality hate speech is a nuanced linguistic and social phenomenon with various tones and forms. The chapter concludes by suggesting some practical implications for developing hate speech recognition systems.

DOI 10.48541/dcr.v12.18

Salla-Maaria Laaksonen

The Datafication of Hate Speech[1]

1 Tempting but difficult automated detection of hate speech

Hateful speech online, often targeting and discriminating specific ethnic or religious groups and minorities, has become a pressing problem in societies and an intriguing problem for social, political, and computational research (e.g., Matamoros-Fernández & Farkas, 2021; Gagliardone et al., 2015; Baider et al., 2017). What is challenging is that hate speech as a term refers to a variety of discriminating or otherwise disturbing speech acts online (e.g., Baider et al., 2017). Further, while hate speech can be detected in public text-based social media discussions, it also takes more subtle forms through memes, targeted propaganda, hate groups, and hate sites (e.g., Brown, 2018; Farkas & Neumeyer, 2018; Roversi, 2008).

Despite the ambiguity of and political debates surrounding the term itself, hate speech is frequently framed as a technological problem: on the one hand, it is a problem because social media platforms and their algorithms help generate and circulate hateful and intolerant content in society (e.g., Udupa & Pohjonen, 2019; Matamoros-Férnandez, 2017); on the other hand, machine learning developers and researchers try to tackle the challenge of identifying and monitoring hateful online content (e.g., Burnap & Williams, 2015, 2016; Davidson et al., 2017).

1 This chapter is based on a previous article (Laaksonen et al., 2020).

Algorithmic solutions for hate speech recognition and prevention are being developed by platform companies and academic research projects. In public discussions, such endeavors are often presented as triumphs of technology: "Facebook pulls 22.5 million hate speech posts in quarter" (Wagner, 2020), or "YouTube removes more than 100,000 videos for violating its hate speech policy" (Binder, 2019).

What actually happens behind these big numbers and success-reporting headlines, however, is rarely disclosed. As users of commercial platforms, we live only with the deliverables and decisions produced by these systems (e.g., Brown, 2018). To build an automated system to identify hate speech, hate needs to be datafied; that is, it needs to be transformed into something that is identifiable, quantifiable, and countable—in essence, understandable for the machine (see van Dijck, 2014; Beer, 2019). Hate speech detection systems, particularly the ones in industrial use, have been criticized for their inadequacy and inconsistency (e.g., Sankin, 2019; Makuch & Lamoureux, 2019), and it is easy to find examples of content that has gone undetected and yet clearly—when interpreted by a human expert—should be removed according to existing content policies.

This chapter discusses the underlying, often hidden, choices related to datafication and the operationalization of hate speech when building technological systems to combat it. The chapter builds on first-hand experiences during action research in a collaborative project in which a hate speech detection system was developed and implemented to monitor the social media activity of political candidates during municipal elections in Finland 2017 (see Laaksonen et al., 2020; Haapoja et al., 2020). The monitoring project involved two NGOs: the Non-Discrimination Ombudsman (NDO, a governmental body to prevent and monitor discrimination), a software company, and researchers from two universities. For one month, the public social media messages of all candidates were collected from social media platform APIs, classified using a machine learning system created for the project, and sent to the NDO for manual checking and potential follow-up procedures. New, manually assigned scores were used to retrain the algorithmic model during the project. The project's aims reached beyond technical solutions: the main goal was to promote campaigning without denigration and hate. Therefore, all political parties were informed about the monitoring.

This chapter discusses and critically reflects on the process of operationalization and datafication of hate speech from contested definitions to quantified algorithmic probabilities. The process of datafying hate speech for computational

purposes emerges as a series of transformations in which the phenomenon that is known to be broad, contextual, and complex essentially becomes reduced to a simple number. Therefore, it becomes an affective object that is measurable, commensurable, and thus seemingly controllable for society that increasingly strives for rationality and technological control.

2 Difficult definitions

To identify an entity, it first needs to be defined. In the case of hate speech, this is a daunting task (see also Sponholz and Frischlich in this volume). In the European context, the debate over hate speech in the past few decades has revolved around questions of ethnicity, religion, multiculturalism, and nationalism (e.g., Berry et al., 2015; Baider et al., 2017), which also makes it a contested topic. The most severe forms of hate speech have been defined in international treaties, the most important of which is the Universal Declaration of Human Rights (UDHR, 1948). Despite the ongoing heated public debate, legislation in most European countries, including Finland, does not contain a definition for any criminal act termed hate speech. The Finnish criminal law code defines various offenses that potentially involve hate speech, such as incitement to hatred (Rikoslaki/Criminal Code 11§10), defamation (Criminal Code 24§9), or illegal assault (Criminal Code 25§7).

Due to the lack of a legal basis, many projects that engage in hate speech recognition use definitions available in various treaties, recommendations, and statements (e.g., European Commission, 2018, 2016; OHCHR, 2013). One frequently used source is the Council of Europe's Committee of Ministers Recommendation 97(20) on hate speech, which defines hate speech as covering "all forms of expression which spread, incite, promote, or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance" (Council of Europe's Committee of Ministers, 1997, p. 107).

The situation is further complicated by the colloquial use of the term (Udupa & Pohjonen, 2019; Brown, 2017a). Hate speech now refers to a variety of speech acts and other ill behavior, both offline and online, ranging from the penal criminal acts discussed above to speech and behavior that is uncivil and disturbing, yet tolerated. This complicates the everyday understanding of, or chance to reach, a consensus on exactly what constitutes hate speech. In its most colloquial and broad-based

definition, hate speech can refer to, for example, verbal discrimination or attacks against various non-ethnic minorities, political hate speech, misogyny, violent pornography, online bullying and harassment, trolling, or doxing—and it has also been referred to as, for example, cyberhate (Brown, 2018), cyber violence (UN Broadband Commission, 2015), or toxic speech (e.g., Perspective API).

Some researchers have suggested separating cases of *hard* or overt hate speech from *soft* or covert hate speech (e.g., Baider et al.,2017). Soft forms of hate speech are not illegal but still raise concerns regarding discrimination. Indeed, one ongoing debate has to do with what can potentially be regarded as a speech act severe enough to constitute illegal hate speech, which groups should be protected from hate speech, and whether the harms caused by hate speech should be considered actual and direct or societal and indirect (e.g., Article 19, 2015; Calvert, 1997; Udupa & Pohjonen, 2019). These debates are reflected in the theoretical discussion on hate speech as discourse, a form of othering that does not necessitate that actual or overt hatred be expressed in words—a speech act or discourse can contain a covert expression of hatred embedded in the context of the speech act (e.g., Brown, 2017a, 2017b; Baider, 2019, 2020). Such discourses do not necessarily have concrete, real-life consequences; instead, they contribute to the overall atmosphere regarding, for example, minorities.

In our project, we chose to build on a broader definition of hate speech than the one allowed for by Finnish legislation and aimed to cover the forms of speech that can be considered either illegal or "legal" hate speech while leaving the final judgment to the NDO lawyers. We grounded our definition in the Council of Europe's Committee of Ministers Recommendation (1997). Further, we used the materials compiled by the NGO Article 19 (2015) for their six-part test for hate speech identification as well as materials produced by the Ethical Journalism Network for journalists to identify hate speech (EJN, n.d). As a result, we generated a list of more fine-grained features of a message to be categorized as hate speech. Such a message contains any of the following: 1) a call to violent action; 2) a call to discriminate or to promote discrimination; 3) an attempt to degrade human dignity based on their characteristics; 4) a threat of violence or the promotion of violent action; or 5) contempt, solicitation, name calling, or slandering. Obviously, the presence of these features in a given message might still be a question of interpretation, and there might be messages in which the feature is indirectly present. However, a formal definition was required to initiate our automated recognition project.

3 Beyond words

Hate speech or online hate is considered a complicated set of practices that are not easily reduced to mere content features of the speech act itself (Brown, 2017a; 2017b; Baider et al., 2017). Materials that instruct humans to identify hateful speech—such as the Article19 test we used—often advise taking wide considerations into account, such as issues related to the context of the speech act, the position of the speaker, and the possible reach of the post. Most automated text mining methods, however, typically start with the words only. They often rely on word lists, bag-of-word approaches, or ngrams (e.g., Greevy & Smeaton, 2004; Pendar, 2007; Dadvar et al., 2013; Munger, 2016). Some more recent detectors utilize bag-of-word vectors combined with word dependencies to identify syntactic grammatical relationships in a sentence (Burnap & Williams, 2015), semantic word embeddings (Badjatiya et al., 2017), or neural networks (Al-Makhadmeh & Tolba, 2019; Relia et al., 2019).

Many of the studies referenced above highlight the difficulties inherent in hate speech detection, particularly the problem of separating hate speech from other types of offensive language (e.g., Davidson et al., 2017). Hate speech cannot be reduced to words or lists of words, even though they can be indicative of hate (Burnap & Williams, 2015; Udupa & Pohjonen, 2019). The actual sentiment or affective tone of a particular message relies immensely on the final form of the expression. Therefore, context-aware systems, such as word embeddings, should enable a fine-grained understanding of word contexts and semantics. Consider, for example, the sentence "Send them all back home" in the context of immigration discussion. It indicates a covert form of hate speech: none of the words as such are indicative of hate, but the combination of words generates a call to action, and the context specifies the meaning of the word "home." To identify this dependency, we need to know what "them all" refers to in the sentence.

The word-centered approach becomes even more problematic when working with social media data, which is quite specific by nature. It is characterized by vernacular expressions and contains mundane words and grammatical variance—which is particularly the case with the Finnish language, where the spoken and written language differ considerably. Further, many forms of social media increasingly support audiovisual forms of communication. Not only are several platforms built around images and videos, but also the use of visual elements, such as emojis

and gifs, is becoming more common on every platform. When treating social media data as text, these visual messages merely appear as empty spaces. Taking the visual forms of communication adequately into account would require more sophisticated data collection methods and, in practice, separate algorithms to identify any content from the visual messages. Furthermore, identifying the sentiments underlying images or multimodal data is a task that is far more difficult than text-based sentiment analysis (e.g., Soleymania et al., 2017).

Some contextual elements are easier to consider; for example, in our project, the larger context was marked by elections. All monitored accounts were political candidates speaking from a political position legitimized by the party, which indicated a clear status. The questions related to the thematic context of a specific utterance and the potential harm incited by it are much more complex. We considered different ways to examine the context of the messages, including, for example, downloading the message thread in which the original message was posted to identify the topic, or running some analyses on the posters' accounts, as suggested by ElSherief et al. (2018). Existing technologies also make it possible to, for example, extract numerical data on a message's reach to evaluate its popularity and visibility, which are considered to affect the potential dangerousness of a post (e.g., EJN, n.d). However, implementation of such methods would have introduced new considerations to our work: expanding the data collection to include messages from non-political actors, such as ordinary citizens, always requires solid justifications—particularly if done by a project that includes a governmental actor.

4 Datafication starts with training data

Automated models to identify hate speech depend on training data annotated by humans. This means human annotators first need to agree on the criteria of hate speech, and then produce a training data set of preferably thousands of messages. This data is then fed to the model as the datafied definition of hate speech. The production of training data is a laborious process that involves potential biases.

It is well known that the quality and content of training data highly affects the performance of machine learning algorithms (e.g., Friedman et al., 2001; Mackenzie, 2017). When choosing the dataset, we provide additional cues to the machine

learning model regarding the kind of content we are looking for. These cues are dependent on, first, the availability of data and, second, on our ability to select reliable, representative data. The biases potentially caused by the training data are sometimes rather obvious in existing systems. For example, Google Jigsaw Conversation AI, a state-of-the-art model for toxic language detection, has been accused of giving higher toxicity scores to sentences that include female/women than male/men (Jigsaw, 2018). Such differences are due to the over-representation of certain classes in the training data that the system is built on. Unless carefully balanced, any collected real-life dataset contains more toxic comments concerning women, so the evaluation of toxicity becomes attached to those specific words that should only be the "neutral context." It is important to note that such biases are difficult to anticipate before being exposed by audits or scandals.

Being aware of these issues, in our project, we tried to create a training dataset that was as balanced as possible. We used a combination of data collected for another racism-related research project and from a large open Finnish language social media dataset containing more general discussions (Lagus et al., 2016). Using keyword searches of common target groups for hate speech as well as common slur words, we aimed to build a dataset that covered hate speech targeted at ethnic and religious minorities. After running some first tests with this dataset, it seemed that our model emphasized words that describe certain minorities. However, a hate speech recognition system that identifies, for example, all Muslim-related speech as potential hate speech is biased and hardly useful. Therefore, we decided to expand our training dataset by including data that targeted other minorities, such as the disabled or the Swedish-speaking minority, or representatives of certain political parties.

Unfortunately, all of these efforts have little temporal persistence. Another known issue of context associated with machine learning models is that developed models rarely perform well if used in another, even slightly different, setting (Yu et al., 2008; Grimmer & Stewart, 2013). Thus, as language develops, politicians change, and new political issues emerge, the detection algorithms trained with old training data might not be accurate anymore. In our project, another disadvantage was that the training dataset did not consist of messages written by politicians but those written by regular citizens, which might imply a different language style altogether. However, datafication also forces us to work with

those data that are available; collecting enough (thousands of) genuine examples of hate speech made by politicians in Finnish is probably not even possible.

As noted, hate speech is an evolving linguistic phenomenon, and its characteristics follow discussions and trends in a given cultural and technological context. Users on social media platforms are aware of the quantification and monitoring of specific keywords used by social media platforms (e.g., Gerrard, 2018). Hence, they constantly develop new ways of expressing emotions such as hate and intolerance more covertly by, for example, misspelling words on purpose or generating new pejoratives or creative metaphors (Brown, 2018; Baider et al., 2017). Think of, for instance, a rather offensive but cunningly masked statement made by a Finnish politician: "*An immigrant is a blemish on the street.*" No training dataset could have enough reminiscent messages to grasp the connotation of the immigrant-blemish metaphor.

5 The hidden interpretation

The training data must be annotated by humans, which brings a component of interpretation into the detection process. As in any content analysis, the reliability and stability of the analysis are controlled by generating shared guidelines for coding and calculating inter-coder reliability. In our project, we used a scale from 0 to 3 to annotate the severity of hate speech (with 3 clearly indicating hate speech, 2 indicating disturbing angry speech, 1 indicating normal discussion with a critical tone, and 0 being neutral). Working with the spreadsheets of data with this scale was a blunt moment of quantification—turning message content and meaning into a single digit, a figure of anticipation (Mackenzie, 2013) – which strips off all the nuances in the verbal expression.

Indeed, annotating the training data taught us that identifying hate speech is not unambiguous, even for humans. We were forced to revisit the definitions and refine the codebook several times before reaching a common understanding. With four coders, we spent almost six hours coding subsets of 100 messages before reaching an acceptable level of agreement, as measured by Cohen's kappa (>.70), while discussing our classification principles after each failed round. It became clear that the coder's own knowledge of the issue and related expressions affected their judgements. For example, a person easily recognizes only slur

words familiar to them. During our classification, we discussed, for instance, the expression "Tim of the night" ("yön Timo" in Finnish), a pejorative expression used to refer to colored people in Finnish. One of our annotators had never encountered the term before, and the hateful content of the message was not obvious without this prior knowledge. The message seemed to be about a specific person instead of referring to an entire group of people with a group noun.

Thus, the labeling is dependent on the previous knowledge of the annotators, both concerning hate speech definitions and national discourses and online culture. In this vein, Waseem and Hovy (2016) showed that amateur annotators are more eager to label messages as hate speech than trained experts. Similarly, Davidson et al. (2017) highlighted the underlying cultural connotations, finding that messages with racist or homophobic content were more likely to be classified as hate speech than sexist messages, which were generally classified only as offensive. Trained annotators thus need good knowledge of both the phenomenon being classified and related cultural connotations; it is essential to be aware of local slur words and other expressions, as well as any juridical definitions that the system may be based on. This means that crowdsourced annotations should be used only with great caution.

After the level of agreement was met, the rest of the training set was coded by the researchers individually, accompanied by a nagging feeling that the variety of messages was so broad that we could probably still find messages on which we disagreed in our individual slots. While categorizing any linguistic phenomenon is a process of reduction, datafication forces us to think even further about probabilities and live with uncertainties.

6 Binary commensurable hate

While some recent advances in natural language processing might help overcome translation issues (e.g., BERT, Devlin et al., 2019; Waseem et al., 2018), hate speech recognition models are language and context sensitive due to their reliance on the training data. Therefore, in our own project, we could not use any existing industry solutions or open libraries typically built and trained for English-language data. Instead, we had to develop a custom text classification model from concept definition to training data composition and model selection. Using standard libraries,

we tested several combinations of feature extraction and machine learning models to identify those that would perform best with our training data. Thus, the model was trained using 90% of the data, and its performance was then tested with the remaining 10%. Based on the performance metrics (Laaksonen et al., 2020), we chose to use a combination of a bag-of-words feature extraction model and support vector machines. Thus, while we acknowledged that hate speech is more than words, the standard machine learning evaluation procedures led us to pick a combination of algorithms that, ironically, emphasized words.

To train the model, our original four-level scale was reduced to a binary classification of clearly denoted hate speech versus other types of speech. This was done because it was a simpler task for the algorithms and because we did not have enough data for each of the categories for the model to perform reliably. The dataset was skewed even with the four-level scale, with non-hate speech dominating the dataset. Here, our somewhat forced numeric evaluation of hate was thus reduced to a binary variable, which was further simplified for the datafied process.

After it was run, our machine learning system assigned a probability score for each message. These scores were then used to sort messages based on how likely they were to contain hate speech. Hence, by following the necessities of the selected approach, the textual training data were quantified and abstracted to a format that allowed for the transformation of hate into *probabilities* (Mackenzie, 2013). This transformation makes hate commensurable, an element that can be measured against a standard, and allows manual or automated ordering of the messages being investigated.

In the training phase, the system reached a precision of 0.79 and recall of 0.98, and thus indeed was able to identify hateful messages to some extent if we accepted our training data as the standard. However, during the actual project, when compared with the manual screening performed by the NDO representatives, it became clear that the model was too sensitive. In the end, only 205 out of a total dataset of 26,618 posts were classified as hate speech by the machine learning system, and after manual screening, only five posts were determined to contain illegal hate speech.

7 Lessons learned

As highlighted in hate speech literature (e.g., Brown, 2017a, 2018b; Baider et al., 2017; Udupa & Pohjonen, 2019), hate speech is a concept with varying definitions, juridical interpretations, and cultural connotations, which makes the automated recognition of it a challenging technical endeavor—but precisely because of that, it represents a type of societal issue many actors are hoping to solve with technology. As discussed in this chapter, these solutions require the datafication and quantification of emotions and affective language, which is not straightforward.

While an adequately well-working machine-learning solution was developed in our project, the automated approach required heavy simplification, such as using rudimentary scales for classifying hate speech, which in reality has several tones and varieties. The main goal of the project essentially turned out to be the quantification of hate. This occurred, first, when classifying the training data, and second, when vectorizing the textual data for the machine learning method (Mackenzie, 2017). In the process of conducting quantification and vectorization, we inevitably flattened the data and lost some of the variety in expressions. This, however, is precisely what makes algorithms powerful through their ability to perform abstraction (Pasquinelli, 2015, cited in Mackenzie, 2017, p. 9).

Experiences in our project showed that recognizing hate speech is not an unambiguous task, even for humans, which makes it a complicated task for machines that rely on specific, quantified features. It is a task that can be achieved in the sense that probabilities are produced, but their validity should be critically evaluated by a human. Algorithmic systems rarely perform their tasks perfectly when dealing with complex language data (Grimmer & Stewart, 2013).

Based on our monitoring project, a system that works to monitor hate speech or other forms of toxic language online should be a long-term, constant project with an *iterative and context-aware approach* to its development. This requires first, reliably annotated training data and a continuous flow of updated, human-annotated data for retraining the model. Such an implementation would, for example, better account for the shifting nuances in the forms of soft hate speech and the periphrases and euphemisms being used. The retraining loop in our system showed that the prediction scores became more accurate during the one-month project period.

Second, hate speech recognition models should not focus only on the content of the message but should also consider the contextual factors related to hate speech, as emphasized by various studies, recommendations, and definitions (e.g., Gagliardone et al., 2015; Article19, 2015; OHCHR, 2013). These aspects include the broader discussion context of the message, the status and position of the poster of the message, and an evaluation of the publicity attracted by the message (see Rabat Action Plan, OHCHR, 2013, section 29). With the current experiences from both research projects and platform actions, it seems unlikely that such systems could be fully automated in the near future.

Salla-Maaria Laaksonen is a researcher in the Centre for Consumer Society Research at the University of Helsinki, Finland. https://orcid.org/0000-0003-3532-2387

References

Al-Makhadmeh, Z., & Tolba, A. (2019). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*, *102*, 501–522. https://doi.org/10.1007/s00607-019-00745-0

Article 19. (2015). Hate speech explained. A toolkit. https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-%282015-Edition%29.pdf

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759–760). ACM Press. https://doi.org/10.1145/3041021.3054223

Baider, F. H. (2020). Pragmatics lost? Overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society*, *11*(2), 196–218. https://doi.org/10.1075/ps.20004.bai

Baider, F. H. (2019). Le discours de haine dissimulée: le mépris pour humilier. *Déviance et Société*, *43*(3), 359–387. https://doi.org/10.3917/ds.433.0359

Baider, F. H., Assimakopoulos, S., & Millar, S. L. (2017). Introduction and background. In S. Assimakopoulos, F. H. Baider, & S. Millar (Eds.), *Online hate speech in the European Union: A discourse-analytic perspective* (pp. 1–16). Springer.

Beer, D. (2019). *The data gaze: Capitalism, power and perception*. SAGE.

Berry, M., Garcia-Blanco, I., & Moore, K. (2015). Press coverage of the refugee and migrant crisis in the EU: A content analysis of five European countries. Report for the UNHCR. Available at: http://www.unhcr.org/protection/operations/56bb369c9/press-coverage-refugee-migrantcrisis-eu-content-analysis-five-european.html

Binder, M. (2019). YouTube removes more than 100,000 videos for violating its hate speech policy. Mashable Tech, September 3, 2019. https://mashable.com/article/youtube-hate-speech-policy-removals/?europe=true

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326. https://doi.org/10.1177/1468796817709846

Brown A. (2017a). What is hate speech? Part 1: The myth of hate. *Law and Philosophy 36*(5), 419–468. https://doi.org/10.1007/s10982-017-9297-1

Brown A. (2017b). What is hate speech? Part 2: Family resemblances. *Law and Philosophy 36*(5), 561–613. https://doi.org/10.1007/s10982-017-9300-x

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science, 5*(1), 11. https://doi.org/10.1140/epjds/s13688-016-0072-6

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet, 7*(2), 223–242. https://doi.org/10.1002/poi3.85

Calvert, C. (1997). Hate speech and its harms: A communication theory perspective. *Journal of Communication, 17*(1), 4–16. https://doi.org/10.1111/j.1460-2466.1997.tb02690.x

Council of Europe's Committee of Ministers (1997) Recommendation 97(20) of the Committee of Ministers to Member States on "hate speech". https://rm.coe.int/1680505d5b

Criminal Code 19.12.1889/39. Rikoslaki (Finnish Criminal Code). https://www.finlex.fi/fi/laki/ajantasa/1889/18890039001

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). *Improving cyberbullying detection with user context. In Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR'13)*, 693–696. https://doi.org/10.1007/978-3-642-36973-5_62

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Eleventh International AAAI Conference on Web and Social Media. https://arxiv.org/abs/1703.04009

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. http://arxiv.org/abs/1810.04805

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: hate speech instigators and their targets. In *Proceedings of Twelfth International AAAI Conference on Web and Social Media,* June 25-28, 2018, Palo Alto, CA. https://doi.org/10.48550/arXiv.1804.04649

EJN—Ethical Journalism Network. (n.d.). Hate-speech: A five-point test for journalists. Available: https://ethicaljournalismnetwork.org/hate-speech-a-5-point-test-for-journalists

European Commission. (2018). Commission recommendation of 1.3.2018 on measures to effectively tackle illegal content online (C(2018) 1177 final). https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online

European Commission. (2016). Code of conduct on countering illegal hate speech online. Announced together with the IT companies. http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf

Farkas, J., & Neumayer, C. (2018). Disguised propaganda from digital to social media. In J. Hunsinger, L. Klastrup, & M. M. Allen (Eds.), *Second international handbook of internet research* (pp. 1–17). Springer Netherlands.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. UNESCO series on internet freedom. https://unesdoc.unesco.org/ark:/48223/pf0000233231

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media and Society, 20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Greevy, E., & Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*, 468–469. https://doi.org/10.1145/1008992.1009074

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Haapoja, J., Laaksonen, S. M., & Lampinen, A. (2020). Gaming algorithmic hate-speech detection: Stakes, parties, and moves. *Social Media and Society*, *6*(2). https://doi.org/10.1177/2056305120924778

Jigsaw (2018). Unintended bias and names of frequently targeted groups. Blog post on the False Positive/Medium on March 9, 2018. https://medium.com/the-false-positive/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23

Laaksonen S-M., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri R. (2020). The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data, 3(3).* https://doi.org/10.3389/fdata.2020.00003

Lagus, K., Pantzar, M., Ruckenstein, M., & Ylisiurua, M. (2016). *SUOMI24 – Muodonantoa aineistolle*. Kuluttajatutkimuskeskus, Valtiotieteellisen tiedekunnan julkaisuja 2016:10. University of Helsinki.

Mackenzie, A. (2017). *Machine learners: Archaeology of data practice*. MIT Press.

Mackenzie, A. (2013). Programming subjects in the regime of anticipation: Software studies and subjectivity. *Subjectivity, 6*(4), 391–405. https://doi.org/10.1057/sub.2013.12

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930–946. http://doi.org/10.1080/1369118X.2017.1293130

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–24. https://doi.org/10.1177/1527476420982230

Makuch, B., & Lamoureux, M. (2019). Web hosting companies shut down a series of Neo-Nazi websites. Vice Motherboard 29.3.2019. https://www.vice.com/en_us/article/7xnn8b/web-hosting-companies-shut-down-a-series-of-neo-nazi-websites

Munger, K. (2016). This researcher programmed bots to fight racism on Twitter. It worked. *Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/17/this-researcher-programmed-bots-to-fight-racism-on-twitter-it-worked/

OHCHR. (2013). Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred (A/HRC/22/17/Add.4). https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf

Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. In *Proceedings of the First IEEE International Conference on Semantic Computing*, 235–241. https://doi.org/10.1109/ICSC.2007.32

Relia, K., Li, Z., Cook, S. H., & Chunara, R. (2019). Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 U.S. cities. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 417–427). https://doi.org/10.48550/arXiv.1902.00119

Roversi, A. (2008). *Hate on the net. Extremist sites, neo-fascism on-line, electronic jihad*. Aldershot Hampshire: Ashgate.

Sankin, A. (2019). YouTube said it was getting serious about hate speech. Why is it still full of extremists? *Gizmodo*. https://gizmodo.com/youtube-said-it-was-getting-serious-about-hate-speech-1836596239

UDHR—Universal Declaration of Human Rights. (1948). United Nations General Assembly resolution 217A. https://www.un.org/en/about-us/universal-declaration-of-human-rights

UN Broadband Commission. (2015). Cyber violence against women and girls. A worldwide wake-up call. A discussion paper by the UN Broadband Commission for Digital Development Working Group on Broadband and Gender. https://www.broadbandcommission.org//wp-content/uploads/2021/02/WGGender_Executivesummary2015.pdf

Udupa, S., & Pohjonen, M. (2019). Extreme speech and global digital cultures. *International Journal of Communication, 13*, 3049–3067.

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance and Society, 12*(2), 197–208. https://doi.org/10.24908/ss.v12i2.4776

Wagner, K. (2020). Facebook pulls 22.5 million hate speech posts in quarter. Bloomberg August 11, 2020. https://www.bloomberg.com/news/articles/2020-08-11/facebook-pulls-22-5-million-hate-speech-posts-in-second-quarter

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science* (pp. 138–142). http://doi.org/10.18653/v1/N16-2013

Waseem, Z., Thorne, J., & Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In J. Goldbeck (Ed.), *Online Harassment* (pp. 29–55). Springer.

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, *5*(1), 33–48. https://doi.org/10.1080/19331680802149608

Digital
Communication
Research.de

Recommended citation: Baden, C. (2023). Evasive offenses: Linguistic limits to the detection of hate speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 319–332). Digital Communication Research. https://doi.org/10.48541/dcr.v12.19

Abstract: As long as we have attempted to sanction untoward speech, others have devised strategies for expressing themselves while dodging such sanctions. In this intervention, I review the arms race between technological filters designed to curb hate speech, and evasive language practices designed to avoid detection by these filters. I argue that, following important advances in the detection of relatively overt uses of hate speech, further advances will need to address hate speech that relies on culturally or situationally available context knowledge and linguistic ambiguities to convey its intended offenses. Resolving such forms of hate speech not only poses increasingly unreasonable demands on available data and technologies, but does so for limited, uncertain gains, as many evasive uses of language effectively defy unique valid classification.

DOI 10.48541/dcr.v12.19

Christian Baden

Evasive Offenses

Linguistic limits to the detection of hate speech

1 Introduction

In an arms race, the offender is typically one step ahead: As defensive technologies are largely designed to fend off known threats, new offensive strategies continue to challenge the development of ever more sophisticated responses. Some threats remain durably beyond the reach of an effective defense, either because they are too unpredictable, or because suitable defenses would infringe in unjustifiable ways upon the liberties of those that they purport to defend, and it is preferable to tolerate the remaining risk. In this intervention, I will argue that this is true not only in security, in cybersecurity, and many other domains, but also in the detection of hate speech.

In the following, I will sketch a rough, but I hope informative caricature of the arms race that has unfolded over the past decades between hate speech and opposing efforts at maintaining civil discourse in online environments. Specifically, I will point out major advances in available technology, as well as specific evasive strategies adopted by users of hate speech (or other sanctioned language uses) in an effort to elude these technological filters. As I will show, many earlier technological advances have successively improved our capacity to detect hate

speech, but have focused on its comparatively plain variants—notably, misspellings, neologisms, and polysemic expressions. With the progressing deployment of context-enriched, AI-based filtering (Kumaresan & Vidanage, 2019), those uses of hate speech that continue to evade unique classification increasingly rely on cultural and situational context knowledge as well as linguistic ambiguity to convey intended offenses. Resolving such uses of evasive language not only poses demands on available data and language processing technologies that quickly exceed defensible dimensions; in many cases, it may even prove impossible to obtain a unique, valid classification. To the extent that further gains are increasingly unlikely, incurring sensitive biases and raising serious ethical objections, we might as well acknowledge that hate speech ultimately constitutes a *social* problem—one that may well be contained, but cannot be resolved, by technical means.

2 The evasive nature of language

As a starting point, we need to acknowledge that language lives, in a sense that does not stretch the metaphor very far (Mufwene, 2001): Words and meanings evolve to match new realities and address new purposes, and language uses respond to the socio-cultural and socio-technical environments that they inhabit. Where it is challenged, language adapts and finds new ways to meet its purpose—for it is the purposes, not the words, that ultimately govern how language is used. Accordingly, any effort to sanction specific uses of language provokes opposing efforts to achieve the same objective while circumventing the sanction (e.g., Gerrard, 2018).

While this is true generally for how language is used, it is particularly true for what linguists call speech acts (Searle, 1969), that is, the use of language not merely to describe, express or otherwise inform, but to elicit certain social effects. Given that this 'pragmatic' use of language for managing social relations is inherently controversial, all languages have developed manifold strategies for committing the same speech act, using different words and expressions depending on its sanctioning in a social context. In circumstances where we don't (have to) fear sanctioning, we may say *in plain words* what we mean ("What you say is absurd"), but for each use, there is typically a whole bouquet of expressions that convey the same meaning in ways but are more likely to pass as acceptable in

situations governed by more restrictive behavioral rules (e.g., "I seriously doubt that," "Oh please, let's not go there again," "Right" [sarcastically]; Bavelas et al., 1990). Using various forms of evasive language uses, we can criticize our partner's cooking (e.g., "Interesting..."), express our disdain for our boss (e.g., "Our wonderful leader"), offer a bribe ("I am sure we can find an agreeable solution...") or inquire whether someone might be interested in sexual relations (e.g., "Want to come up for one more drink?") – all the while maintaining a plausible pretense that this was not our intended meaning, should the response be adverse (Gruber, 1993; Obeng, 1997). As long as we could get ourselves into trouble by what we say, evasive language uses have been there for us to dodge expected sanctions. Accordingly, when algorithmic sanctioning entered the stage of digital communication, language was ready for it.

3 The words that weren't so

As more or less anything in natural language processing technologies, also the sanctioning of inappropriate speech started as a list of keywords—typically, of more or less openly derogatory labels or references to racist, anti-Semitic, misogynist or otherwise hostile discourses (Zelenkauskaite et al., 2021). Noticing that certain terms were suppressed, users of early chat rooms and forums quickly learned to use creative spellings, truncated words (a particularly interesting case is "f***"/"f-ing," where written—and to some extent even spoken—language use redacts itself in anticipation of being redacted, thereby evading redaction while simultaneously marking the sanctioning of the expressed meaning; Fairman, 2006), and acronyms. Leet (the replacement of certain letters by numbers) was one outcome, and many para-linguistic symbols (e.g., the "(((They)))" meme; Tuters & Hagen, 2020) and neologisms (e.g., "cuck," "libtard"; Hodge & Hallgrimsdottir, 2020) were born to outsmart the filtering algorithm. Keyword lists evolved and grew in pace, trying to catch any known and increasingly conventionalized spellings, and fuzzy matches increasingly enabled algorithms to also catch simple variations, such as (accidental or deliberate) misspellings and leet.

At the same time, the redaction of any expressions used as swearwords, inappropriate comments or hate speech rapidly revealed an important limitation of such keyword-based strategies, which chiefly derived from two main problems.

On the one hand, redacting or posts containing certain words effectively disabled also discussions wherein offensive terms were used not to trade insults, but to negotiate communication norms and their policing (e.g., debates about real or hypothetical uses of offensive terms in other communication environments). On the other hand, problems arose when words were used to convey offensive meaning that also had other uses (Magu et al., 2017).

4 The words that weren't that

In response to the possibility to use potentially offensive terms in non-offensive ways, one key strategy was to augment existing keyword lists with additional disambiguation criteria. For instance, algorithms might distinguish whether a term was used as part of a quote, thus enabling users to quote and criticize the others' words or negotiate communication norms without triggering a sanction. Longer expressions could be considered to distinguish between "white trash," "white trash can," and "this white trash can lick my..." (Warner & Hirschberg, 2012). Algorithms could be taught to distinguish uses of "swine" within and outside an agricultural context, or recognize the token "Fucking" as a reference to the so-named town in Austria (recently renamed Fugging).[1] Of course, any such rule-based filters could easily be gamed, as users figured out which combinations the algorithm might catch or tolerate, generating new expressions and linguistic obfuscations that were plain to the reader, but unclear to the machine. Still, context-based disambiguation constituted an important advance in the detection of hateful speech.

That said, disambiguation needs by far outstretched the capacity of text-based algorithms. One problem arises from the use of terms that are mostly used in benign ways (e.g., "chocolate," "snowflake," "Skype"; Magu et al., 2017) but can be also used to express contempt and hatred (e.g., as racial slur). As the specific meaning of such terms often arises from the wider context of a statement, valid disambiguation rules are near-impossible to define. Moreover, especially group labels such as "gay" or "Jew" can be used in both offensive and benign ways in more or less identical linguistic contexts (e.g., "seems everyone is gay there"),

1 https://www.politico.eu/article/austrian-village-of-f-king-to-be-renamed-fugging/

while the meaning depends on who is saying these words, and to whom: their derogatory potential rest half in the inaccuracy of their use (e.g., calling a man "little girl"; Schmidt & Wiegand, 2017), and half in the subcultural valuation of their denoted meaning (e.g., among anti-Semites, homophobes; Hodge & Hallgrimmsdottir, 2020). Not only do subcultures develop their own, idiosyncratic vocabularies and expressions to express hostility in oblique, identity-coded ways, multiplying the range of indicators and rules require consideration (e.g., in German youth culture, "victim" can denote a contemptible weakling and fool; in the misogynic Incel [involuntary celibate] movement, "Stacy" constitutes a sexually objectifying, resentful reference to a pretty woman; Jaki et al., 2019); but the very same expression can often be read to convey or not convey an insult, depending on the reader's habitual language use and awareness of communication contexts (see Litvinenko in this volume, for the various layers of such contexts).

Moreover, ethical issues arise from defining membership categories such as "Jew," "gay," "feminist" or "black" as potentially offensive terms, and any mistaken suppression of such references may justly raise public outcry.

5 The words that weren't needed

With the advance of machine learning based natural language processing, filters once again appeared to catch up with the manifold variations in language use in context. Relying on an appraisal of entire textual contributions and large databases of reference cases to distinguish textually similar, but semantically or pragmatically different language uses, supervised algorithms are capable of flagging problematic uses with much improved nuance and accuracy (Schmidt & Wiegand, 2017). Still, blind spots exist wherever relevant terms are absent in the reference corpus, or if there are too few reference cases to draw confident inferences. While the problem arises primarily for rare expression and can be mitigated by more inclusive training samples, this strategy quickly becomes unwieldy for highly heterogeneous communication contexts, where very many different uses may require consideration. Especially considering the reliance of machine learning algorithms on past language use, the constant evolution of digital discourse continuously weakens the predictive power of past reference cases. New events and situations enable new variations in the use of suspect words that

the algorithm could impossibly predict, and language users continually develop new ways to express their contempt. Moreover, machine learning strategies are particularly slow to adapt to new or rare language uses, as they depend on a sufficient number of cases to be manually rated, included in the training data, and accumulated to sustain confident algorithmic disambiguation.

Beyond those challenges raised by terms that may or may not express contempt, yet greater challenges arise from the expression of contempt without resort to potentially offensive terms. As machine learning tools tend to err in the direction of terms' more common usages, they are unlikely to recognize hate speech conveyed by means of entirely innocuous words (e.g., "back in the day, we would have put them on a train to the East," here conveying a veiled holocaust reference). Based on a recent project that I conducted together with Tzlil Sharon, which aimed to identify references to conspiracy theories in online text, such veiled references appear to be surprisingly pervasive (Baden & Sharon, 2021). Chiefly, there appear to be four main variants: First, allusions point at intended meanings without specifying them (e.g., "They sure got paid many Shekels for this," suggesting some anti-Semitic conspiracy theory; "I have a rope and a cozy spot…" an oblique reference to lynching), leaving it to the reader to complete the interpretation (Obeng, 1997; Wilson & Sperber, 2012). Second, language use often reaches beyond the present text into co-present contributions, using anaphora (e.g., "this," "she") to import additional meaning (e.g., "They're going to kill all the pigs in the region [to prevent the swine flu from spreading]" – "That is bad news for [German Chancellor] Merkel!"; see also Halliday & Hasan, 1976). Very similarly, multimodal communication reaches beyond the present text into co-present visual information to create additional meanings (Ben-David & Matamoros-Fernández, 2016). Third, intertextuality does the same, but reaches out to absent, supposedly familiar texts (e.g., "Die Fahnen hoch…" quoting the first words of the Horst Wessel song, anthem of the National Socialist German Workers' Party; see also Kristeva, 1981). Fourth, speakers can avoid specifying offensive meanings, using exophoric references to events, actors, or other objects in the world that are presumed to be known to other readers (e.g., posting on the day of the Christchurch terror attack: "What a great day, hopefully also soon here;" "Time to go ER," ER being the initials of Elliot Rodger, the early leader of the Incel movement and perpetrator of the 2014 Isla Vista attacks; Jaki et al., 2019).

While it is theoretically possible to algorithmically model those disambiguations needed to handle such oblique forms of hate speech (Kumaresan & Vidanage, 2019), in practice, such an enterprise quickly approaches its limits. For instance, adjacent interactive speech content can be included in the training data—but it is often unclear what nearby information an anaphora refers to (e.g., "That's way too nice" might refer to the preceding comment, the hierarchically superior comment, or the original post, each time inviting different readings; Halliday & Hasan, 1976). Likewise, it would be invalid to assume that any included anaphora refers to adjacent texts, as the same words can be used also exophorically to refer to salient present situations outside the text. Many allusions, and most intertextual references might be disambiguated by contextualizing present posts against relevant reference corpora, such as natural discourse samples on related matters, encyclopedic knowledge, or the day's news (Baden, 2018). Alas, knowing just what relevant reference corpus might be required more often than not requires that one is already familiar with a wide variety of related language uses, contextual knowledge, and recent news. Moreover, even if relevant reference corpora can be identified in an inductive fashion (e.g., by online search), machine classification still requires relevant reference materials to be labeled (Warner & Hirschberg, 2012). Clearly, continually annotating and adding any potentially relevant text to an ever-growing reference corpus, just in case that any of it might be needed to disambiguate potential hate speech, is not a viable strategy.

Even if it were possible to enable classification by considering such encompassing reference corpora, any expansion of context data shifts the detection of hateful content further away from binary, rule-based decisions toward probabilistic judgments, where both 1 (*certainly objectionable*) and 0 (*certainly harmless*) are rare occurrences. The larger the reference data, the more likely will instances be matched by pure coincidence, inflating false positive ratios (Kumaresan & Vindage, 2019). The same is true for every expansion of the textual context considered toward classification. In addition, increased reliance on reference corpora shifts the responsibility for detection away from software-controlled rulesets toward a reliance on third party algorithms (e.g., google) and patterns that emerge inductively from the reference data. Given the sensitivity of falsely redacting legitimate contents, and the consequent need for rather high classification thresholds, context-augmented machine classification is likely to achieve

at most modest improvements in detection, at considerable cost in terms of algorithmic complexity, data and labor demands, and justification.

6 The words that weren't enough

Recognizing these limitations, much current practice in content moderation continues to rely on human judgment—often following a flagging procedure that relies on any of the algorithms sketched above (Kalsnes & Ihlebæk, 2021). While much less systematic in their appraisal of available information and context, human judges should generally outperform algorithms in their capacity to detect veiled offensive content—simply because such oblique expressions are designed to be understood by humans, and missed by computers. Picking up on suspicious word choices and omissions, human judges can disambiguate allusions, intertextuality and references to adjacent texts or present situations by comprehending the context wherein a user comment was made.

And yet, even humans are often unable to decide the status of a comment—not because they cannot extrapolate those meanings expressed by the text, but because the same text supports more than one possible meaning (Boxman-Shabtai & Shifman, 2014; Warner & Hirschberg, 2012). Beyond the use of language to convey unambiguous meaning in oblique ways, the same strategies also permit the construction of properly ambiguous messages. For instance, does the comment "Someday my friends and I will come visit" convey a friend's announcement, a fan's admiration, or a veiled threat? Unless we know more about the relation between the commenter and the addressee, the statement defies disambiguation. "Why don't you go home, leave us in peace!" is ambiguous (personal/collective "you," which may/may not be a racial reference, home as home/home country, us as particular group/nationalist reference, etc.) even if we know their relation not to be close. A particularly important genre of ambiguity concerns apparent irony or humor, wherein it remains unclear whether denoted meanings are endorsed or rejected (e.g., Boxman-Shabtai & Shifman, 2014; Hodge & Hallgrimsdottir, 2020).

Using ambiguity, authors can express even meanings that are heavily sanctioned—e.g., violent threats, calls to violence, and other criminal offenses—while maintaining plausible deniability and (likely) avoiding algorithmic redaction.

Contrary to intuition,[2] such ambiguity is actually quite common in contentious discourse. For instance, in our study of conspiracist discourse, less than a tenth of all references to conspiracy theories were entirely unambiguous (Baden & Sharon, 2021). Some statements cued conspiracy theories, but left a backdoor open for benign readings (e.g., "[US Senator] Bernie [Sanders] is controlled opposition"). Others were fully ambiguous: "Nobody sued the media for creating an atmosphere like this." Of course, conspiracist discourse is known for its evasive style, as proponents of conspiracy theories have long faced social sanctions; however, the same should be true for hate speech.

One drawback of ambiguity is, of course, that the speaker's intentions may be misunderstood—a problem solved in conspiracist discourse by primarily addressing fellow believers whose predilection for certain interpretations can be safely predicted. The same logic enables ambiguous hate speech to the extent that it is intended primarily to be understood by fellow haters (Magu et al., 2017). However, to ensure that also addressed outsiders catch the intended drift, authors need to either decrease ambiguity (increasing the risk of redaction and other sanctions), or demonstratively emphasize the ambiguity, so as to alert readers to the availability of additional, hostile meanings (e.g., by adding "..." or "😉"). Unable to conclude confidently that available benign meanings were intended, the addressee is thus forced to construct and consider also the offensive interpretation.

Inversely, many cases of ambiguous statements are arguably harmless and arise accidentally when people choose their words carelessly and fail to exclude alternative, hostile meanings (e.g., when US Senate minority leader Schumer said that two Trump appointees to the Supreme Court would "pay the price" for a vote against abortion rights).[3] Consequently, flagging any speech that potentially supports offensive meanings inevitably captures numerous harmless or unintended instances, while excusing any that support harmless meanings likely misses some of the most hostile, but deliberately cloaked attacks. Especially for statements

2 When confronted with ambiguous statements, readers typically decide intuitively on one preferred reading and ignore other available interpretations, raising the illusion that most language is unambiguous. However, when prompted to make no assumptions but systematically evaluate those meanings enabled by a statement, many more statements turn out to be ambiguous (Eco, 1979)

3 https://www.washingtonpost.com/nation/2020/03/05/schumer-trump-supreme-court/

which support multiple equally plausible interpretations, there cannot be a consistent policy even for human judgment, as coders are forced to choose between redacting content that can plausibly be defended as harmless, or permitting content that can reasonably be understood as hate speech.

7 The words that were read

One final approach, accordingly, that has been widely adopted for the moderation of digital content, relies on audiences' subjective interpretations to flag offensive content. Contents get redacted, or submitted for review, if a certain number or proportion of readers regards them as offensive and flags them as such (Kalsnes & Ihlebæk, 2021). In this way, moderators can exploit the vastly superior capacity of diverse audiences to recognize oblique meanings—although at the cost of inevitably moderating post hoc, with considerable delay. However, also this strategy comes with important limitations.

To begin, especially where hateful comments are apparent only to members of extremist communities, most readers are likely to miss offensive meanings, while those who "get" the expressed hostility are likely to agree and thus unlikely to report the statement (Jaki et al., 2019). The more hate speech relies on context-based disambiguation and ambiguity, the more its detection depends on individuals who are literate in extremist discourses but in disagreement with their underlying values (see Becker & Troschke in this volume).

Furthermore, user complaint-based moderation is always vulnerable to targeted campaigning, as has been recently made salient by the rise of "cancel culture," predominantly in US-based communication forums (Ng, 2020). Given the fundamental ambiguity of language as well as the wealth of available contexts, it is very often possible to construct a statement as offensive—even if it was neither so intended nor widely understood as such. Activist users can thus use the flagging option to strategically suppress unwelcome voices wherever these miss possible ambiguities in their statements—a threat that is particular salient in the context of satire, which frequently relies on ambiguous language to confront contentious issues.

8 Conclusion

Over the course of the past two to three decades, there have been several important advances in our capacity to algorithmically detect and redact hate speech. At the same time, every advance has also revealed new limitations and contingencies in the classification of potentially offensive meanings, and provoked further adaptations in the use of evasive language suitable to express hostility in ways that are unlikely to be detected.

As I have attempted to show in this chapter, many important limitations in our capacity to detect hate speech do not primarily reflect inadequacies in those tools and algorithms employed to classify natural language, but derive from the evasive use and ambiguity of language itself (Bavelas et al., 1990). While available algorithms are increasingly capable of resolving ambiguities that exist *within* the classified text (e.g., misspellings, polysemy or different pragmatic uses; Schmidt & Wiegand, 2017), most of the remaining ambiguities reach *beyond* the text itself into intertextual context, the identities of involved actors, and the embedding social situation and communication culture (Wilson & Sperber, 2012). Of course, it is in principle possible to include ever wider context data, consider metadata information, or utilize reader reactions and talkbacks to augment classification (Baden, 2018); alas, given the vast range of potentially relevant contextual information (e.g., concurrent news, subcultural discourses, historical reference material, popular culture), including sensitive personal data (e.g., if accurate classification requires the knowledge that an addressee is gay, female, or from New York; Kumaresan & Vidanage, 2019), such an endeavor appears neither particularly practicable nor ethically defensible. Additional issues arise where detection relies on users' subjective judgments and third party-controlled data sets, and where binary decisions to permit or censor content are based on probabilistic, error-prone classifications (see Laaksonen in this volume, for a further discussion of these issues). Even if all these issues could be solved, further disambiguation is unlikely to push back the frontier by much: As in any arms race, hostile users of digital communication technologies are likely to respond to such advances by retreating deeper into the realm of ambiguous language, for which there logically cannot be an algorithmic disambiguation.

Moreover, any attempt to classify and sanction ambiguous speech is bound to raise intense contestation and public backlash (Shen & Rosé, 2019). Beyond the

inevitable rise in misclassifications, redacting comments that can be plausibly construed as harmless not only invites the justifiable indignation of the sanctioned authors, but it also sets a precedent for preemptively suppressing potentially offensive content. Accused of either censoring free speech or permitting contents that can be interpreted as offensive, neither ambiguous language, nor the black-boxed probabilistic classification can offer much grounds for justification, and even human judgment remains subjective and contestable. In light of the considerable demand on data and algorithms, the limited scope of likely improvements in detection, and the substantial damage for democratic public debates that may arise from an ill-justified suppression of ambiguous statements, attempting to pursue hate speech into the realms of evasive and ambiguous language may well do more harm than good. In terms of the arms race metaphor introduced above, an effective defense against heavily context-sensitive, evasive forms hate speech most likely requires unjustifiable infringements upon people's privacy and freedom—and where hateful communication is clad in fully ambiguous uses of language, there can be no effective defense.

Christian Baden is Associate Professor at the Department of Communication and Journalism at the Hebrew University of Jerusalem, Israel. https://orcid.org/0000-0002-3771-3413

References

Baden, C. (2018). Reconstructing frames from intertextual news discourse: A semantic network approach to news framing analysis. In P. D'Angelo (Ed.), *Doing news framing analysis II: Empirical and theoretical perspectives* (pp. 3–26). Routledge.

Baden, C., & Sharon, T. (2021). Blinded by the lies? Toward an integrated definition of conspiracy theories. *Communication Theory, 31*(1), 82–106. https://doi.org/10.1093/ct/qtaa023

Bavelas, J. B., Black, A., Chovil, N., & Mullet, J. (1990). *Equivocal communication.* Sage.

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication, 10*, 1167–1193.

Boxman-Shabtai, L., & Shifman, L. (2014). Evasive targets: Deciphering polysemy in mediated humor. *Journal of Communication, 64*(5), 977–998. https://doi.org/10.1111/jcom.12116

Eco, U. (1979). *The role of the reader: Explorations in the semiotics of texts.* Indiana University Press.

Fairman, C. M. (2006). Fuck. *Public Law and Legal Theory Working Paper Series, 59*. Ohio State University.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society, 20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Gruber, H. (1993). Political language and textual vagueness. *Pragmatics, 3*(1), 1–28. https://doi.org/10.1075/prag.3.1.01gru

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* Routledge.

Hodge, E., & Hallgrimsdottir, H. (2020). Networks of hate: The alt-right, "troll culture", and the cultural geography of social movement spaces online. *Journal of Borderland Studies, 35*(4), 563–580. https://doi.org/10.1080/08865655.2019.1571935

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict, 7*(2), 240–268. https://doi.org/10.1075/jlac.00026.jak

Kalsnes, B., & Ihlebæk, K. A. (2021). Hiding hate speech: Political moderation on Facebook. *Media, Culture & Society, 43*(2), 326–342. https://doi.org/10.1177/0163443720957562

Kristeva, J. (1981). *Language and desire: A semiotic approach to literature and art.* Columbia University Press.

Kumaresan, K., & Vidanage, K. (2019). HateSense: Tackling ambiguity in hate speech detection. *Proceedings of the IEEE 2019 National Information Technology Conference,* 20-26. https://doi.org/10.1109/NITC48475.2019.9114528

Magu, R., Joshi, K., & Luo, J. (2017). Detecting the hate code on social media. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1). https://ojs.aaai.org/index.php/ICWSM/article/view/14921

Mufwene, S. S. (2001). *The ecology of language evolution.* Cambridge University Press.

Ng, E. (2020). No grand pronouncements here...: Reflections on cancel culture and digital media participation. *Television & New Media, 21*(6), 621–627. https://doi.org/10.1177/1527476420918828

Obeng, S. G. (1997). Language and politics: Indirectness in political discourse. *Discourse & Society, 8*(1), 49–83. https://doi.org/10.1177/0957926597008001004

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/W17-1101

Searle, J. R. (1969). *Speech acts.* Cambridge University Press

Shen, Q., & Rosé, C. P. (2019). The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. *Proceedings of the Third Workshop on Abusive Language Online*, 58–69. https://doi.org/10.18653/v1/W19-3507

Tuters, M., & Hagen, S. (2020). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*, *22*(12), 2218–2237. https://doi.org/10.1177/1461444819888746

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Proceedings of the 2012 Workshop on Language in Social Media,* 19–26.

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Zelenkauskaite, A., Toivanen, P., Huhtamäki, J., & Valaskivi, K. (2021). Shades of hatred online: 4chan memetic duplicate circulation surge during hybrid media events. *First Monday, 26*(1). https://doi.org/10.5210/fm.v26i1.11075

Digital
Communication
Research.de

Recommended citation: Becker, M. J., & Troschke, H. (2023). Decoding implicit hate speech: The example of antisemitism. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 335–352). Digital Communication Research. https://doi.org/10.48541/dcr.v12.20

Abstract: This article deals with the problem and the different levels of implicity in the context of antisemitic hate speech. By means of authentic examples stemming from social media debates, we show how alluded antisemitic concepts can be inferred on the basis of conservative interpretation. We point out the role of different sources of knowledge in reconstructing the ideas implied in a statement as well as potential sources of error in the interpretation process. The article thus focuses on the methods and findings of the interdisciplinary research project "Decoding Antisemitism" conducted at the Center for Research on Antisemitism (ZfA) at TU Berlin. By doing so, it explains the procedure of qualitative content analysis as a fruitful approach to understand the actual repertoire of antisemitic hate speech online.

DOI 10.48541/dcr.v12.20

Matthias J. Becker & Hagen Troschke

Decoding Implicit Hate Speech

The example of antisemitism

1 Why focus on implicity?

Anyone examining forms of hate speech, either on- or offline, will soon come into contact with utterances that communicate hateful ideas, but without clearly assignable word material. In this article, we will demonstrate how implicit hate speech functions using the example of antisemitism.[1] Writers do not necessarily have to refer to Jews or Israel, nor do they have to reproduce antisemitic stereotypes, such as greed or infanticide, or express invective, threats, or death wishes—all of which should be classed as explicit antisemitism. Writers communicate antisemitic ideas through an enormous spectrum of language use patterns both at the word and sentence levels.[2] This spectrum of linguistic variations of meaning—among them the

1 The prerequisite here is a viable definition of antisemitism that takes into account both historical and contemporary manifestations of hostility toward Jews. As a basis for our work, we use the internationally recognised IHRA definition (IHRA, 2016). It nevertheless had to be scientifically specified by assigning the various antisemitic tropes to the enumeration of aspects of antisemitism given with the definition.

2 On reasons for communicating antisemitism in implicit forms, see Troschke and Becker (2019, pp. 152–154) and Becker (2020).

dominant field of implicit patterns (i.e., semantically ambiguous or underspecified)—has to be taken into consideration in order to be able to make reliable statements about the actual presence of antisemitism online.

Two recent studies underscore the importance of taking implicit statements into account when measuring antisemitism online. A corpus analysis of the comments sections of the news websites *Zeit Online* (German) and *The Guardian* (British), focusing on the use of Nazi comparisons in Middle East discourse (with such comparisons understood as a form of current antisemitism), showed that only one out of 304 Nazi comparisons in reader comments was *explicit* (i.e., equating Israel and Nazi Germany based on the classic pattern *X is like Y*; Becker, 2021). All other Nazi comparisons were characterized by varying degrees of implicity, either through incomplete comparisons, innovative metaphors, or onomastic or open allusions, which require world knowledge to extrapolate them.[3]

A similar picture emerged in the context of the Berlin-based research project "Decoding Antisemitism."[4] The project's first "Discourse Report," which primarily refers to British news websites, shows that the commenting readership tended to communicate antisemitic stereotypes implicitly (Becker, Troschke et al., 2021).[5] The examination of approximately 1,200 comments from web debates on the Jewish billionaire and philanthropist George Soros found that—provided that there is a qualitative approach—roughly 15 per cent of the examined comments contained antisemitic statements. Prior to the qualitative analysis, we conducted searches with relevant words that represent antisemitic concepts. On this basis

3 Other corpus analyses of the German-speaking internet confirm the finding that the linguistic reproductions of antisemitic attributions are largely implicit (Schwarz-Friesel, 2020). With regard to implicit language usage patterns outside of the internet, see, for example, Schwarz-Friesel and Reinharz (2017).

4 From 2020, the three-year pilot project "Decoding Antisemitism: An AI-Driven Study on Hate Speech and Imagery Online" is being carried out at the Center for Research on Antisemitism at TU Berlin in cooperation with King's College London. In this project, which is funded by the Alfred Landecker Foundation, antisemitism in comments sections on British, French, and German mainstream news websites and social media platforms are qualitatively and quantitatively analyzed (see https://www.alfredlandecker.org/en/article/decoding-antisemitism).

5 The report focuses on web debates triggered by the media, *The Guardian*, *The Independent*, and the *Daily Mail*, about Jewish billionaire and philanthropist George Soros, the EHRC report on antisemitism in the British Labour Party, and the exclusion of Jeremy Corbyn.

alone, only a small fraction of the antisemitic statements were identified. Thus, only by cataloguing current language usage patterns based on qualitative analysis can we refine quantitative studies and increase the degree to which they represent a research object.[6]

This article explores how the problem of implicity can be grasped by research practice, using findings from the Decoding Antisemitism project's analysis of implicitly produced antisemitic meanings in social media comments. It presents various forms of implicity and discusses how these can be included in the analysis. The article then lays out the role played by different sources of knowledge for understanding or even inferring the subject matter of a comment and refers to sources of error for the interpretation process. Finally, we illustrate our interpretive approach using examples from social media. In this way, we also show where the limits of the interpretation of implicit statements may lie.

2 Knowledge areas for extrapolating the implicit

How do we deal with implicit, ciphered statements? To fully extrapolate the meanings conveyed by implicit statements or statements containing implicity, we need to distinguish *three areas of knowledge*: First, the interpreters require the necessary *language knowledge* to recognize and understand even the most delicate nuances in a statement. Second, the *context* of a statement (e.g., within a thread) and its potential impact must be taken into account. Third, relevant *world knowledge* of the broader context is required, including general knowledge about the cultural space (including society, politics, history), discourses, and conventions, as well as specific knowledge of the subject whose implicit mediation is to be investigated.[7] In our case, this is the indispensable knowledge of historical and contemporary antisemitism in all its manifestations.

6 See also the project's second "Discourse Report," which compares results from qualitative and quantitative analyses of commentary sections discussing the escalation phase of the Middle East conflict in May 2021 (Becker, Allington et al., 2021).

7 For *world knowledge*, see, for example, Plümacher (2006) and Schwarz-Friesel (2013, pp. 37–41).

The necessary *language knowledge* is available to all members of a language community with sufficient linguistic competence. However, there are differences between language producers and recipients in terms of the precision with which the structure of meanings can be understood.

Context knowledge results from a reception of the context that is equally accessible to all. In our case, it includes the trigger for the comments, whether it is media articles, posts, videos, initial comments, or associated comments. Here, too, despite the same contextual information, all recipients differ in the extent to which they are able to fully integrate this knowledge into the process of forming conclusions.

In terms of *world knowledge*, major differences can exist between recipients regarding a particular subject. These differences are greater the further the object is away from the center of general cultural knowledge. The more extensive the object-specific world knowledge, the quicker meanings based on implicit acceptances or assumptions can be fully identified. This is therefore an important prerequisite for the analysis. In this particular case, this knowledge is constituted by our research on historical and contemporary forms of antisemitism and antisemitic discourses and our insights stemming from past analyses. This knowledge is required to assign meaning to the most linguistically explicit antisemitic statements and even more so in the case of implicit comments. The latter sometimes only refer to fragments of an antisemitic concept. However, the whole of these fragments must be known to the interpreters in order for them to extrapolate the respective concept from the reference to a part of it.

3 Securing interpretations

The application of the knowledge from these three areas determines the extent to which all meanings and nuances are identified when categorizing texts. Insufficient knowledge as well as incomplete or faulty reasoning processes at the linguistic and conceptual levels can lead to antisemitic meanings not being recognized or being interpreted in a text without sufficient evidence. The effect of over-interpretation can, however, also be the result of over-sensitivity caused by priming: The continuous examination of the subject matter in the coding process can lead to false presumptions of the (not reliably verifiable)

presence of the subject in the texts. Distortions are possible in two directions: overlooking antisemitic meanings or false positives.

This means it is crucial that statements should be categorized conservatively. Conservative attribution means deciding in favor of the more likely meaning in situations where there are at least two possible interpretations of an ambiguous utterance in order to prevent false positives. Conflicting valid interpretations are set out, the probabilities of the correctness of different interpretations are weighed against each other, and the use of world knowledge to enrich each interpretation is undertaken with caution. On the one hand, the clearly defined interpretation scheme (compiled in a guidebook[8]) allows one to arrive at meaningful interpretations; on the other hand, it avoids over-interpreting a statement (false positives).

When the same text corpus is categorized by several coders looking for certain content, their respective levels of topic-related world knowledge are the element most likely to produce differences in their conclusions. The extent of world knowledge thus has a decisive impact on the comparability of reasoning and coding processes between coders and the resulting categorizations of texts. Therefore, the level of this world knowledge has to be raised collectively in advance.

In order to minimize deviations in interpretation—and thereby categorization—as much as possible, it is important to define the procedure with a comprehensive (and continuously refined) guidebook containing coding instructions. This allows for the orderly presentation of all conceptual and linguistic-semiotic phenomena (along with a listing of numerous representative and distinguishing examples) for the benefit of facilitating the general understanding of the phenomenon in question. This guidebook is constructed using both definitions drawn from existing research literature and those developed inductively in the course of engagement with the empirical material itself. It serves as a common knowledge base, both with regard to the object of investigation and all the areas where implicity is found.

In our research project, we created distinctions between antisemitic and non-antisemitic attributions as follows: All anti-Jewish stereotypes used explicitly against Soros, for example, were coded as antisemitic, since it can be assumed

8 The guidebook contains the key elements of the resources used by human coders to analyze comment threads: stereotypes and linguistic and image-analytical categories defined and substantiated with explicit and implicit examples.

that Soros' Jewish identity is widely known. Even if a particular writer unintentionally expressed himself or herself in this way, there is high potential that a negative attribution will be associated with Soros' Jewishness by a large portion of the web comment's readership. If, on the other hand, Soros' actions as an investment banker are demonized, this, as well as any form of criticism of Soros and his practices, however harsh, is not regarded as antisemitic, even if the underlying worldview, which we cannot infer directly from the comment, may be antisemitic. Our understanding of antisemitism comprises antisemitic concepts, insults (whether specifically antisemitic or aimed at Jewish identity), and various speech acts that express the wish to harm Jews or Israelis based on their Jewish or Israeli identity. For example, "Soros is the evil of the world/the evilest person" is an antisemitic utterance. "Soros is an evil banker," however, is an example of a strongly negative, but non-antisemitic, evaluation, explicitly linked to his professional background as a banker.

In the first phase of the coding process, consensual validation across coders—reaching intersubjective agreement via discussion about categorizations—should take place in a succession of small steps. Once a common understanding has been established, checks on intercoder agreement can take place at longer intervals. Furthermore, regularly collating intercoder reliability—and resolving disagreement and, if necessary, henceforth adjusting guidebook instructions—assures the quality of the coding.

On the basis of the guidebook, we have developed an extensive code system using the analytical tool MAXQDA that includes antisemitic concepts (at the content level, e.g., stereotypes) as well as phenomena at the linguistic and semiotic levels through which the concepts are communicated. In this way, an utterance can be coded with regard to its conceptual as well as structural particularities. We can then record the linguistic—and possibly semiotic or semiotically accompanied—expressions of content and examine how they are combined.[9]

9 For methodological literature on qualitative content analysis, see, for example, Mayring (2015) and Kuckartz (2018).

4 Interpretative approach

In this section, we show how an antisemitic concept can be communicated via different types of utterance. We will then demonstrate how implicitness can be realized at the different levels and how these levels can interact. Subsequently, we will move on to authentic corpus examples from our research and, based on the interpretation of these, we will reconstruct our interpretative journey through the three aforementioned areas of knowledge, providing an insight into our conservative approach and the use of our guidebook.

In order to extrapolate its meaning, we break down a comment into units of meaning, determine which propositions are present, track how it is embedded in the context, and identify linguistic forms of implicity, as well as any informational gaps that have to be filled by conclusions. We then summarize the conclusions regarding the individual components of the comment and how they relate to each other.

Language and world knowledge are relevant for every interpretation. An interpretation without world knowledge is impossible for the identification of antisemitism. Certainly, an explicit attribution can be understood from a linguistic point of view (and without being augmented by world knowledge); however, the utterance still needs to be situated within the ideology of antisemitism. Context knowledge is very often a relevant source for interpretation, but there are statements that can be comprehensively interpreted without contextual information because their meaning is independent from the context and would be the same in other contexts.

The possibilities for disguising a concept within subtle words, that is to communicate it implicitly, are numerous and can be found on several levels that are becoming increasingly complex. They begin at the word level when a writer uses *acronyms* or *puns*, for example. Changes to the surface of the word add another unit of meaning without having to do so explicitly. Another possibility involves using *allusions*, where the surface of the lexeme used remains intact, but—due to the apparent conflict between the meaning of the allusion and the utterance into which it has been transplanted—an indirectly communicated meaning is constituted. At the word group or sentence level, implicit antisemitism can be communicated by means of so-called *indirect speech acts*, in which what is meant results from the combined evaluation of all the communicated

units. These phenomena represent only a small part of how implicitness is constituted. In what follows, we will give examples of the various levels.

4.1 Word level: Using synonyms

Let us begin with a simple concept: The word *money* is the linguistic expression of the concept of a payment tool and means of storing value. However, when analyzing authentic language data, it quickly becomes clear that concepts are reproduced linguistically with a high degree of variation and that the standardized expression is only used sometimes. This makes the linguistic variance for antisemitic concepts all the greater since, for the most part, they do not have standardized linguistic expressions in the language community. Speakers who refer to the concept MONEY[10] can alternatively use words from other concept areas, such as *dough*, *bread*, *loot*, *wedge*, *moolah*, and *lolly*, as synonyms. Therefore, various signifiers exist for the concept being conveyed. When extrapolating the intended meaning, the reader or listener can use the mental lexicon entry for MONEY: Which synonyms exist within the language community, and is there a match here? Alternatively, the context helps to make the extrapolation of the intended object more precise. This already shows how interactions between different knowledge bases have to be reconciled by the recipient faced with a situational speech act.

The actions of "asking for money" or "demanding money" can also be paraphrased in English and French, as in German, with the metaphorical phrase, "holding out one's hand." The connection is underpinned by a reasoning process based on language knowledge through a conventional metaphor (Skirl, 2009). Metaphors can exist at the word or sentence level and serve to concretize abstract facts. This means that they have a function that promotes knowledge. At the same time, however, they can also manipulate by conveying a controversial thought in an indirect and partly elaborated way, thus giving it the status of being sayable.

The conceptual connection between Jews and AVARICE has a prominent position in antisemitic thinking. Writers with this mindset might speak directly of

10 Since stereotypes are phenomena that exist on the conceptual (i.e., mental) level and can be reproduced using language, stereotypes are given in small caps on the following pages in accordance with the conventions of cognitive linguistics.

"greedy and money-grabbing Jews"—but they could also conceptualize Jews using the metaphorical phrase mentioned above and thus involve themselves in the discourse in a more subtle way. Due to the conventional status of the metaphor in combination with the social condemnation of antisemitism in post-war Germany, the statement "Jews are always holding out their hands" would certainly also face sanction in numerous contexts of expression.

4.2 Sentence level: Indirect speech acts, irony, and rhetorical questions

A further increase in implicity is achieved through changes at the sentence level. Speech act theory deals with speech acts of a direct and indirect nature (Searle, 1969, 1975). Indirect speech acts are statements that, word-for-word, do not express what is meant. In such cases, there are two levels: one part that is expressed literally (the literal or secondary act), and another part that is intended (the primary act). Irony is an example of an indirect speech act. The association of Jews with money and greed can be expressed in the context of an ironic statement, such as: "Yes, yes, I know, Jews have never made much of money ..." To decipher the irony, knowledge shared with the writer must be used. In this case, it is initially language knowledge: The emphasized (exaggerated) affirmation that opens the statement, its reinforcement by the generic statement, and the omission points that leave the issue open can serve as indicators that in this statement the assertion of the direct speech act is negated by an indirect one.

However, irony can also be deciphered through world knowledge, and it can be concluded in this case too that by using such information, which includes knowledge of the corresponding antisemitic stereotype, the stereotype is indirectly affirmed. The corresponding knowledge base and its contrast with the information in the statement indicate that there is an implicature to be drawn here (Levinson, 1983). The implicature is a conclusion relating to the actual meaning of the sentence: what is intended. It is up to the recipient to determine here that in the statement the assertion of the direct speech act is negated by an indirect one and then to infer what is actually being meant. The use of irony enables the writer to present the insinuation of AVARICE in a persuasive way. In addition, he or she can avoid being pinned down to the meaning when threatened by sanctions.

Parts of the meaning of the statement can thereby be deleted in terms of information—they can be withdrawn or denied.

A rhetorical question is another indirect speech act. If someone writes, "Do Jews always hold out their hands?", he or she can withdraw from the threat of sanctions by claiming to have innocently asked a genuine question. In addition to this slightly encrypting function, rhetorical questions serve to determine and thus emphasize in an elaborated form what is being asked (Lee-Goldman, 2006).

The presence of irony as well as rhetorical questions and the need for corresponding conclusions can be extrapolated from the language, context, and world knowledge. If someone wants to increase the degree of implicity in statements linking Jews to money by, for example, placing a question pronoun in the subject position, as in "Who is (once again) holding out their hand?", context knowledge can be a prerequisite, in this case, of an anaphoric reference, for understanding who is being referred to here (in a roundabout way through an indirect speech act).

We may encounter rhetorical questions like these in German contexts of expression when talking about the Nazi past and the culture of remembrance, for example, in comments sections on the internet that refer to these subjects. Simply drawing on language and context knowledge in the inference process may potentially leave the recipient unsatisfied about the meaning of the statement because even in light of the comment's trigger—the article itself—there may be several people or groups or no named person who could fit the role of demanding money. In these cases, it is only possible to identify who is meant by augmenting the context information and interpreting it using world knowledge.

4.3 Changing stereotypes

After the end of National Socialism, antisemitic thinking did not simply disappear in Germany. Instead, previous stereotypes were updated. The stereotype avarice, at least in its explicit form of presentation, was no longer widespread in the public communication space (Bergmann & Erb, 1986). It was not only the form of expression that changed but also the concept of the stereotype. The insinuation of general avarice was joined by that of the instrumentalization of the holocaust (and of antisemitism in general), the allegation that Jews would capitalize on particular issues and subsequently use the Holocaust to make themselves

rich. It is this expression of specifically German secondary antisemitism that the abovementioned rhetorical question refers to and that allows the implicature to be drawn that Jews should be used as the subject of the sentence.[11] The presupposition in brackets "once again" also makes the assumption that this is not a one-off incident but a routine relationship between both sides, the victims or their descendants and the (later) society of the perpetrators, in which the former are said to harass and take advantage of the latter.

5 Corpus examples

Based on this chain of examples, it is clear how the three areas of knowledge (partly connected with each other) are used to (gradually or even fully) extrapolate the levels of meaning behind language usage patterns in terms of their complexity and ambiguity. We will now present a range of web comments from our recent corpora to illustrate how, in our research, we draw on the aforementioned areas of knowledge to show how chains of inference are constructed.

Below the line of a BBC documentary uploaded to YouTube that critiques the usage of antisemitic conspiracy theories in discussions about George Soros, a commenter reproduces an antisemitic concept by using, among other things, a semiotic marker:

"Evil $oros hands."

The finding, resulting from language knowledge, that Soros' name was not spelled correctly and that a dollar sign was inserted instead, leads to the question as to why this was done. Here, by drawing an implicature, an informational gap has to be filled: The signifiers are merged into a compound. It is also a pun since changes are simultaneously made to the surface of the name Soros. The compound brings together the concept areas (the individual Soros with that of

11 Experience abroad shows that the manifestation of a secondary guilt-deflecting or exonerating antisemitism is largely unknown. There, the statements discussed here could not be fully interpreted, which again illustrates how access to cultural and milieu-specific world knowledge is fundamental for the extrapolation of language and thinking patterns.

money). This leads to the conclusion that Soros is associated with money and that an affinity for or the pursuit of money is alleged to be a fixed trait of his character. In this way, he is assigned the stereotype GREED—in addition to the explicit attribution of EVIL.

In a *Daily Mail* article, Soros is described as a philanthropist. In a comment that refers to this, the user's view is that this attribution should be corrected to say:

> "Philanthrocapitalist not philanthropist. Huge difference."

The play on words, "philanthrocapitalist," represents the (linguistically unsuccessful) attempt to portray Soros as a special "friend of capital" or "capitalism" and evokes the stereotype GREED by insinuating that Soros prioritizes the opportunities of profiteering that capitalism offers above caring about the well-being of people. Language knowledge is sufficient for decoding the compound word. However, to be able to understand that the attribution is directed at Soros, context knowledge in the form of the anaphorical connection between "philanthropist" and Soros has to be included.

Context knowledge can be linked to a comment using, among other things, anaphors. Due to their obviousness, in many cases, anaphoric connections apparently do not need to be mentioned. However, given the fact that in many current research projects on hate speech the testing of automatic recognition takes place on the basis of machine learning, the proportion of anaphoric connections in the construction of meaning of a hate comment becomes more important: The more the antisemitic meaning is based on this contextual information, the more difficult it is for algorithms to correctly categorize the text if they are unable to take that information into account.

In the context of an announcement by Soros that he will fund universities, one comment reads:

> "He will finance the Far Left Globalism Marxist Indoctrination and students brainwash ...!"

The pronoun "he" refers to Soros. Without this knowledge, the attribution made in the comment could not be assigned the intended Jewish object, and the specifically antisemitic nature of this post would not be recognized. At the same

time, the topos of the comment would also independently reflect an antisemitic perspective, namely the insinuation that there is a system of global reach and absolute power said to be based on left-wing political ideologies. The connection to universities is established through "students." The accompanying attributes convey the image of a university teaching in the service of an ideology supported by Soros. This draws on the antisemitic topos of POWER and CONSPIRACY, which claims that Jews are behind ideologies or social developments. This topos can also be expressed in insinuations of influence on media, politics, and socialization factors, resulting in developments that are beneficial to them.

Context knowledge may also be required for a multiple-step interpretation. Referring to warnings made by Soros in an interview about the situation in the European Union (EU), a user remarks, "Its days [that of the EU] were numbered regardless of COVID-19" and thus predicts its imminent end. Another user replies with "Hopefully so are his!" and thus takes up the prediction, supplies it with a wish, and turns it against Soros with a change of object—namely wishing him an early death. In order to infer the antisemitic death wish, the anaphoric connections from "are" to "were" in the reference comment and from "his" to "Soros" in the article must be identified and linked.

The antisemitic content of a text can also be extrapolated without any context knowledge, as in the following comment also posted in response to the above-mentioned documentary by the BBC on antisemitic conspiracy theories:

"WWG1WGA IGWT !"

Clearly world knowledge is required to decode this. The first acronym stands for the emblematic slogan, "Where we go one we go all," of the adherents of the meta-conspiracy theory QAnon, which is linked to a number of antisemitic ideas and acts as an allusion to this. The slogan is intended to create a sense of community and strengthen the solidarity of its supporters. This function is supported by the subsequent acronym, which stands for "In God we trust" and provides the aspect of confidence in and the support of a higher power for one's own cause—and thus its legitimacy. The exclamation mark emphasizes that the slogan is to be understood here as an appeal or a commitment to QAnon. The affirmation of antisemitic conspiracy theories is an act of antisemitic communication. However, this affirmation can be extrapolated in another way by adding context knowledge. Since the

comment is a reaction to the repudiation of conspiracy theories, the implicature can be drawn that the allusion must be understood as an opposing position that serves to demarcate a common bogeyman—in this case, Soros.[12]

In reference to an article that discusses criticism of Soros on Facebook, a user states:

> "Soros, Zuckerberg, all good hearted Christian names. They want you to believe that the meek shall inherit the earth"

Since the user patently falsely identifies prominent Jewish people as Christians, the reader can (assuming his or her world knowledge allows him or her to recognize this fact) thus draw the implicature that, in the user's opinion, the named people are also not "good hearted" (a trait here assigned to Christians) but rather are to be identified as wicked. It is an indirect speech act in the form of irony. Subsequently, it is claimed that it is in their interests to lull people into false belief: From our world knowledge about Christian theology and the relationship of this assumption and hope for the "meek" to actual history, we know that this is not true nor will be true. Therefore, here too, the opposite must be concluded in order to understand the attribution. The conclusion here is thus that both individuals are eager to make people believe in an error, thereby putting them (or keeping them) in a state of defenselessness, which makes them incapable of acting in the face of the world's challenges; this will then allow the former to achieve their true goals more easily. We know from the implicature in the first sentence that they are said to pursue these goals with bad intentions. The reverse implies that Soros and Zuckerberg, rather than striving for the (Christian) ideal, actually want a world order of hardness and ruthlessness in which people are subjugated to their power. Correspondingly, the stereotypes of DECEIT, HYPOCRISY, and GREED FOR POWER are found here.

In response to the BBC documentary, an accusation that regularly crops up that its content or even the BBC itself was influenced by Soros (or Jews in general).

12 Since the web comment refers to the context of a BBC documentary on antisemitism, it can be assumed that precisely those aspects of the QAnon conspiracy theories are activated that are clearly antisemitic – and not the aspects that are not inherently antisemitic.

In the following example, the user rejects the connection between antisemitism and conspiracy theories in relation to Soros and indirectly claims that the BBC is spreading falsehoods for money:

> "This is nothing to do with anti-Semitism. I suggest you (The BBC film makers) look into this more carefully; those who still know what the truth is and haven't taken your 30 pieces of silver."

The user formulates a construction of opposites between "those who still know what the truth is and haven't taken [any money]" and those who are allegedly bribed. From the proposition, "taken your 30 pieces of silver," a form of financial influence can be generally inferred. Adding world knowledge, on the other hand, enables it to be classified into an antisemitic world view since, according to the Gospels, Judas betrayed Jesus and obtained this sum of money. In this respect, it is an allusion to a core concept of anti-Judaism: the betrayal of God's son by a Jew, which all Jews have been accused of since then. The rejection of antisemitism goes hand in hand with the invocation of one of the oldest antisemitic attributions. At the same time, the INFLUENCE ON THE MEDIA stereotype is activated.

It is not always possible to assign a valid interpretation to a text. It is more often the case that, though conclusive interpretations are possible, the probability of their correctness appears to be too low to be able to categorize them on this basis. This includes the following reaction to the abovementioned BBC documentary:

> "TRUMP-PENCE 20-24 America.Freedom.Constitution."

Due to the world knowledge that Trump supports conspiracy theories, the implicature can be drawn that this sudden reference to a second term in office propagated by the user is based on the idea or wish that Trump and Pence will fight the imagined conspiracies. Alternative interpretations here would be that the user, inspired by the trigger, changes the topic and does not want to refer to the context or that it is simply a multiple post that does not establish a connection to the context.

One of the posts about Soros's support for universities reads:

> "He needs to go"

It remains unclear whether this refers to wishing that Soros should withdraw from the public eye or that he should die. For this reason and in contrast to utterances like "The end is near …" or "Soros should be neutralized," we do not assign this text to the death wish category.

6 Conclusion

These examples demonstrate the ways in which antisemitic stereotypes can appear in a variety of forms and show how categorizing them using only a few indicators or merely quantitatively on the basis of, for example, key words, collocations, or n-grams, is only able to partially capture (antisemitic) hate speech. The differentiated code system we work with opens up the possibility of, but also demonstrates the need for, determining the existence of antisemitic concepts more precisely than would be the case, for example, with a code system that only differentiates antisemitic and non-antisemitic texts or the different forms of antisemitism. At the same time, it makes coding easier as individual concepts are constantly in view instead of having to be repeatedly recalled in relation to a general category. The advantages such an approach provides for qualitative analysis are also applicable to the subsequent step of machine learning. The data provided might theoretically allow algorithms not only to recognize antisemitic concepts in the course of the learning process in accordance with the categorization of the concepts in the training data, but even to learn to differentiate them (to a certain extent).

The linguistic level in the code system supports the visualization process when making interpretations, as this has to be substantiated specifically on the basis of the language usage. Both the conceptual and the linguistic levels play a double role: They support the interpretation process using the existing categories in a close examination of the indicators, and they enable more detailed analysis results.

Matthias J. Becker is project lead of the "Decoding Antisemitism" project and a postdoctoral researcher at the Center for Research on Antisemitism (ZfA) at TU Berlin, Germany. https://orcid.org/0000-0003-2847-4542

Hagen Troschke is a researcher in the "Decoding Antisemitism" project. https://orcid.org/0000-0003-4098-4115

References

Becker, M. J. (2020). Antisemitism on the Internet: An underestimated challenge requiring research-based action. *Justice*, (64), 32–40. https://www.ijl.org/justicem/no64/index.html#32

Becker, M. J. (2021). *Antisemitism in reader comments: Analogies for reckoning with the past*. Palgrave Macmillan.

Becker, M. J., Allington, D., Ascone, L., Bolton, M., Chapelan, A., Krasni, J., Placzynta, K., Scheiber, M., Troschke, H., & Vincent, C. (2021). *Decoding antisemitism: An AI-driven study on hate speech and imagery online*. Discourse Report 2. Technische Universität Berlin: Zentrum für Antisemitismusforschung. https://decoding-antisemitism.eu/publications/second-discourse-report/

Becker, M. J., Troschke, H., & Allington, D. (2021). *Decoding antisemitism: An AI-driven study on hate speech and imagery online.* First Discourse Report. Technische Universität Berlin: Zentrum für Antisemitismusforschung. https://decoding-antisemitism.eu/publications/first-discourse-report/

Bergmann, W., & Erb, R. (1986). Kommunikationslatenz, Moral und öffentliche Meinung. Theoretische Überlegungen zum Antisemitismus in der Bundesrepublik Deutschland [Communication latency, morality and public opinion. Theoretical reflections on antisemitism in the Federal Republic of Germany]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *38*, 223–246.

IHRA – International Holocaust Remembrance Alliance. (2016). *Working definition of antisemitism*. https://www.holocaustremembrance.com/resources/working-definitions-charters/working-definition-antisemitism

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung (Grundlagentexte Methoden)* [Qualitative content analysis. Methods, practice, computer support (Basic texts on methods)]. Beltz.

Lee-Goldman, R. (2006). *A typology of rhetorical questions. Syntax and semantics circle*. UCB.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* [Qualitative content analysis. Basics and techniques] (12th ed.). Beltz.

Plümacher, M. (2006). "Weltwissen." Ein sprachwissenschaftlicher Terminus phänomenologisch betrachtet ["World knowledge." A linguistic term phenomenologically considered]. In D. Lohmar & D. Fonfara (Eds.), *Interdisziplinäre Perspektiven der Phänomenologie. Neue Felder der Kooperation: Cognitive Science, Neurowissenschaften, Psychologie, Soziologie, Politikwissenschaft und Religionswissenschaft* [Interdisciplinary perspectives on phenomenology. New fields of cooperation: Cognitive science, neuroscience, psychology, sociology, political science and religious studies] (pp. 247–261). Springer.

Schwarz-Friesel, M. (2013). *Sprache und Emotion* [Language and emotion] (2nd ed.). UTB.

Schwarz-Friesel, M. (2020). "Antisemitism 2.0" – The spreading of Jew-hatred on the World Wide Web. In A. Lange, K. Mayerhofer, D. Porat, & L. H. Schiffman (Eds.), *Comprehending and confronting antisemitism. A multi-faceted approach* (pp. 311–338). De Gruyter.

Schwarz-Friesel, M., & Reinharz, J. (2017). *Inside the antisemitic mind: The language of Jew-hatred in contemporary Germany*. Brandeis University Press.

Searle, J. R. (1969). *Speech acts. An essay in the philosophy of language*. Cambridge University Press.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics. Vol. 3, Speech acts* (pp. 59–82). Academic Press.

Skirl, H. (2009). *Emergenz als Phänomen der Semantik am Beispiel des Metaphernverstehens. Emergente konzeptuelle Merkmale an der Schnittstelle von Semantik und Pragmatik* [Emergence as a phenomenon of semantics exemplified by metaphor comprehension. Emergent conceptual features at the interface of semantics and pragmatics]. Narr.

Troschke, H., & Becker, M. J. (2019). Antisemitismus im Internet. Erscheinungsformen, Spezifika, Bekämpfung [Antisemitism on the Internet. Manifestations, specifics, combating]. In G. Jikeli & O. Glöckner (Eds.), *Das neue Unbehagen. Antisemitismus in Deutschland und Europa heute* [The new unease. Antisemitism in contemporary Germany and Europe] (pp. 151–172). Olms.

Digital
Communication
Research.de

Recommended citation: Kim, J. Y. (2023). Machines do not decide hate speech: Machine learning, power, and the intersectional approach. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 355–369). Digital Communication Research. https://doi.org/10.48541/dcr.v12.21

Abstract: The advent of social media has increased digital content—and, with it, hate speech. Advancements in machine learning help detect online hate speech at scale, but scale is only one part of the problem related to moderating it. Machines do not decide what comprises hate speech, which is part of a societal norm. Power relations establish such norms and, thus, determine who can say what comprises hate speech. Without considering this data-generation process, a fair automated hate speech detection system cannot be built. This chapter first examines the relationship between power, hate speech, and machine learning. Then, it examines how the intersectional lens—focusing on power dynamics between and within social groups—helps identify bias in the data sets used to build automated hate speech detection systems.

DOI 10.48541/dcr.v12.21

Jae Yeon Kim

Machines Do Not Decide Hate Speech

Machine learning, power, and the intersectional approach[1]

1 Introduction

The advent of social media platforms—such as Twitter, Facebook, and YouTube—has increased digital content. Alongside this change, *hate speech*—defined as highly negative and often violent speech that targets historically disadvantaged groups (Walker, 1994; Jacobs & Potter, 1998; see also Sponholz in this volume) – has also increased. In response, social media platforms have leveraged machine learning to scale up their efforts to detect and moderate users' content (Gitari et al., 2015; Agrawal & Awekar, 2018; Watanabe et al., 2018; Koushik et al., 2019; see also Ahmad in this volume). Developing a system that relies less on human inspection and validation is desirable for these firms because this system's efficiency gains would allow them to grow further and increase profits.

Unfortunately, scale is only part of the problem related to hate speech detection and moderation. Marginalized groups and individuals (e.g., ethnic and racial minorities, women, lesbian, gay, bisexual, transgender, and queer

1 I thank Thomas R. Davidson, Renata Barreto, two anonymous reviewers, and the editors of this volume for their constructive comments on an earlier draft of this chapter.

[LGBTQ] people, immigrants, and people with disabilities) are major targets of hate speech, which is one reason why many social media platforms cite potential harm against marginalized people as the main reason to target hate speech (Twitter, 2021; Facebook, 2021; YouTube, 2021). A difficulty arises, however, in that these historically disadvantaged groups' speech is more likely than others' to be labeled as *hate speech* (Sap et al., 2019). Ideally, advancements in machine learning should have solved this problem by developing efficient, fair automated hate speech detection systems. For instance, the probability of labeling speech as *hate speech* should not depend on whether the speaker is a member of a marginalized group. Unfortunately, however, many scholars have found that these systems are vulnerable to racial, gender, and intersectional biases (Waseem & Hovy, 2016; Waseem, 2016; Tatman, 2017; Waseem et al., 2017; Davidson et al., 2019; Davidson & Bhattacharya, 2020; Kim et al., 2020; Zhou et al., 2021). So, why does this paradox persist despite so many technological innovations?

A closer inspection reveals that this vulnerability is not ironic. The meaning of *hate speech* changes over time and across places (Walker, 1994; Gelber, 2002; Bleich, 2011; see also Litvinenko in this volume) because shifts in power relations determine who can say what comprises hate speech (Binns et al., 2017; Geva et al., 2019; Al Kuwatly et al., 2020). If hate speech is a social construct, so is hate speech annotation. Automated hate speech detection relies on human-annotated data, and it faces a challenge in that hate speech annotation concerns deciding whether particular speech violates social norms. However, most norms have boundaries that can vary, depending on context. For instance, White people's use of the "n-word" to describe Black people is likely a racial slur, but the same word used among African Americans is unlikely to be offensive. These subtleties should be acknowledged in building an automated hate speech detection system. Otherwise, hate speech algorithms will be more likely to label African Americans' speech as offensive than White peoples' (see Sap et al., 2019). This "label bias" (Hinnefeld et al., 2018; Jiang & Nachum, 2020), defined as the misannotation of training data, is a fundamental challenge in building a fair artificial intelligence system. Practitioners and scholars define *machine learning performance* based on its prediction accuracy, but if the ground truth that an algorithm predicts is invalid, whether its prediction is effective becomes a secondary question.

Without considering this data-generation process, a fair and automated hate speech detection system cannot be built. Focusing on the data-generation process

requires thinking about power because certain individuals and groups set boundaries around hate speech, and these norms influence hate speech annotation. In this vein, this chapter first discusses how power, hate speech, and automated hate speech detection systems are deeply interconnected. It then examines how the intersectional lens (i.e., a focus on power dynamics between and within social groups) helps identify bias in the data sets used to build automated hate speech detection systems. The chapter enriches the discussion of the obstacles to building a fair automated hate speech detection system and how to overcome them.

2 Bias in machine learning and hate speech detection

Bias in a machine learning application is usually defined as a residual category of fairness (for a review of the various definitions of *fairness* in machine learning applications, see: Gajane & Pechenizkiy, 2017; Corbett-Davies & Goel, 2018; Mitchell et al., 2021). A machine learning model is biased if it performs unevenly across subgroups, based on their protected features, such as race, ethnicity, and sexual orientation. Because a model's uneven performance can be defined in many ways, many definitions of *fairness* exist. For instance, if one's definition is *demographic parity* (Dwork et al., 2012; Feldman et al., 2015), in the ideal world, a predictive model should demonstrate an equally positive rate across demographic groups. Another influential metric is *equality of opportunity*. Under this definition, a machine learning model is *fair*, in a binary classification case, if its predicted outcome has equal true positive rates across demographic groups when $y = 1$ and equal false-positive rates when $y = 0$ (Hardt et al., 2016, pp. 1–2). From this conceptual perspective, a bias exists in an automated hate speech detection system if a certain racial group's speech is labeled *hate speech* more often than others'.

These mathematical definitions are convenient tools to assess how predicted outcomes may influence the welfare (allocation harms) and representation (representational harms) of a particular group compared to other groups' (Crawford, 2017; Barocas et al., 2020). Nevertheless, these "outcome-focused" indicators are limited because they do not inform researchers of how these outcomes were generated.

The concern regarding the data-generation process is particularly critical to understanding the elements missing from the current discussion on bias and fairness in machine learning. In empirical social science research, if an outcome

is biased racially (e.g., racial disparity in income and poverty), an attribute of race influenced the outcome (Holland, 1986; Greiner & Rubin, 2011; Sen & Wasow, 2016). For example, Bertrand and Mullainathan (2004) measured racial discrimination in the labor market by sending fictitious resumes in which job applicants' names varied. The field-experiment results indicated that candidates with White-sounding names (e.g., Emily and Greg) received more callbacks for interviews than Black-sounding names (e.g., Lakisha and Jamal). In this example, researchers were interested in estimating the effect of name attributes on resumes that may cause people to perceive a job applicant's race differently. In contrast, if machine learning applications' outcomes are biased racially, their models perform poorly for one racial group compared to others. Unlike in the earlier social science example, in this case, the machine learning literature does not focus on what caused the disparity in model performance across demographic groups that could exacerbate existing socioeconomic inequities (Kasy & Abebe, 2020).

To understand bias in machine learning applications and its origins, scholars and practitioners must understand that machine learning applications are embedded in society (Martin, 2019). Machine learning models depend on data for their performance, and a particular algorithm may outperform others, depending on the characteristics of the data sets it uses and the tasks it performs. Humans are involved in both generating training data and defining these tasks, and these decisions are susceptible to long-standing explicit and implicit human biases. Therefore, bias in machine learning applications, including automated hate-detection systems, encompasses a wide spectrum of societal and historical biases (Garg et al., 2018; Jo & Gebru, 2020). No panacea can solve this problem, and only a careful investigation of underlying causes can yield promising solutions.

Unfortunately, most solutions that have been presented focus on fixing the most immediate issue. For example, IBM released the "Diversity in Faces" data set in 2019 in response to criticisms of bias in the commercial use of computer vision algorithms because these algorithms discriminate against Black women (Buolamwini & Gebru, 2018). This effort is laudable, but an exclusive focus on training data sets insufficiently addressed the bias issue fully because representation bias is only one element among a broad set of societal and historical biases (Mehrabi et al., 2021). The more fundamental issue is not that training data sets lack sufficient amounts of Black women's faces but, rather, why this practice was accepted and not questioned in the first place. The main concern in this regard is power—not

the extent of observations related to different racial groups. Therefore, the larger social environment that defines what kinds of training data and labeling processes are acceptable must be investigated. This investigation is particularly necessary when training data are generated through normative judgments, which are highly susceptible to bias (see Bocian et al., 2020, for a recent review on this subject in social psychology).

Measurement is an issue related to this label bias, and it is also fundamental to making automated hate speech systems fair. *Objective ground truth* in hate speech data sets is difficult to define. If the goal of building a predictive model is differentiating between cats and dogs, a consensus could easily be reached on essential features that help make sound predictions (Deng et al., 2009). However, hate speech is part of a societal norm. What comprises hate speech varies across groups and over time, and its definition has become a contested political issue (Walker, 1994; Gelber, 2002; Bleich, 2011). Power relations establish such norms and, thus, determine who can say what comprises hate speech. Political stakes are involved in deciding that some speech is acceptable and other speech is not. If my group's speech is labeled hate speech and other groups' is not, the odds of my speech eliciting a political and legal toll are higher than others'. In a hierarchical society, power relations are unequal, and these relations determine who shapes rules and norms (Lukes, 1974). Therefore, a subordinate group's speech is more likely to be labeled *hate speech* than a dominant group's (Maass, 1999; Collins, 2002; Campbell-Kibler, 2009).

In the machine learning literature, researchers have circumvented this measurement problem by assuming either that well-defined ground truth exists or that the best approximation is available through social consensus (Dwork et al., 2012, p. 214). This assumption is convenient for building a compact theory, but it presents an important obstacle to be acknowledged in practice. To acknowledge the relationship between power, normative judgments, and hate speech labeling, researchers should recognize how the definition of *hate speech* is established socially. If ground truth is generated through an unequal social process, then making predictive models' performance similar across demographic groups is insufficient to make an automated hate speech detection system fair (Blodgett et al., 2020).

In principle, fairness in hate speech detection systems can be accomplished by promoting greater transparency and inclusion in building such detection systems.

2.1 Transparency

Researchers should acknowledge that hate speech is a contested concept and that some individuals and groups have more power to define *hate speech* than others. They should also provide a position statement in their research that describes why they define *hate speech* in a particular way within the context of their research background. For instance, researchers can construct *hate speech* as a categorical or continuous variable, or as a single-dimensional or multi-dimensional concept. They should explain why their definition is more appropriate to their research than other definitions. Model cards—a documentation tool for fair machine learning—helpfully illustrate this approach (Mitchell et al., 2019).

2.2 Inclusion

Furthermore, researchers should include people most likely to be harmed from automating hate speech detection in the development process so that they can provide critical feedback on data-collection and -annotation procedures (Frey et al., 2020; Halfaker & Geiger, 2020; Katell et al., 2020; Patton et al., 2020). This participatory approach is essential from ethical and scientific perspectives. These individuals possess deep knowledge of what speech targets them and what part of their speech practices could be mislabeled as *hate speech*.

Practicing these principles requires considering power in two steps: first, how does the dominant group define *societal norms* (between-group power relations), and second, how do these societal norms marginalize particular segments of subordinate groups (within-group power relations)? Clarifying these points helps identify which of researchers' assumptions should be transparent and which members of marginalized groups should be invited as research partners. In the next section, I discuss how the intersectional approach helps raise this type of awareness.

3 Why use an intersectional approach?

The intersectional approach helps explain what shapes people's perception of hate speech and how this bias is baked into data sets used to train hate speech

detection algorithms. This approach to hate speech differs from the approach that focuses on hate speech analysis at the content level, according to which *hate speech* can be intersubjectively defined based on certain characteristics (for a discussion of distinctions between these approaches, see, e.g., Sellars, 2016, pp. 14–18). It also differs from a similar approach that focuses on social identity theory (Tajfel, 1970; Tajfel & Turner, 1979; Brown et al., 1980; Perreault & Bourhis, 1999). According to the social identity approach, people are easily motivated to define group boundaries, favor their in-group, and denigrate their out-group; as a result, annotators are more likely to label speech by their out-group's members more negatively than speech by their in-group's members (Binns et al., 2017; Geva et al., 2019; Al Kuwatly et al., 2020). The solution to reducing label bias in this context is to recruit members of different groups as annotators (e.g., White and Black people, men and women). Then, when aggregated, the biases of annotators from different backgrounds would cancel each other out.

However, this approach raises another question: How should we define *diversity*? The above approach works only if the members of a particular group have strongly homogeneous opinions on hate speech. In practice, homogeneity means that if a researcher is investigating racial bias, then they should assume that other forms of bias—such as gender bias—do not exist. This assumption is unwarranted if different axes of discrimination (e.g., race, class, and gender) intersect and make a segment of a subordinate group more marginalized than other segments.

For instance, Cohen (1999) demonstrated how the intersection of race and sexuality explains African American communities' unwillingness to mobilize against the acquired immunodeficiency syndrome (AIDS) epidemic despite these communities' long history of involvement in racial justice movements. Racial elites (e.g., Black pastors) intentionally avoided including the AIDS epidemic in their political agendas because they did not want their groups' moral reputations tainted by the stigma attached to Black LGBTQ communities and their presumed relationship with the AIDS epidemic.

Although Cohen's research does not speak to hate speech analysis directly, its main insight—marginalization within a marginalized group—is relevant. Suppose a hate speech data set contains a significant volume of hate speech that targets members of the Black LGBTQ community in the United States and researchers recruit racially diverse annotators to build an automated hate speech detection system. Such an initiative fails to consider the gender dimension of the potential

bias issues within a racially marginalized group. Consequently, such an automated hate speech detection system remains vulnerable to societal and historical biases because hate speech targeting Black LGBTQ communities is highly likely not to be labeled *hate speech*. Even Black annotators might avoid labeling such attacks as *hate speech* because their community leaders had not addressed this problem publicly. These annotators might not recognize how problematic this form of speech can be.

In this case, the key to understanding the data-generation process is to think about power relations in the contexts of between- and within-group power relations (Crenshaw, 1990; Blodgett et al., 2020; Kasy & Abebe, 2020). A dominant group creates prevailing societal norms that condone certain sexual relations but not others. These norms define which thoughts, speech, and behaviors are acceptable within subordinate communities if they want to maintain their (moral) reputations in society at large. Depending on how this boundary is constructed and reproduced, some aspects of marginalization may be acknowledged more publicly than others.

4 Concluding remarks

Making fair automated hate speech detection systems requires a deeper understanding of who decides what comprises hate speech. Machine learning algorithms are powerful tools for detecting hate speech at scale, but an oversight remains. These models are trained with labeled data that are susceptible to historical and societal biases—a particularly acute problem in hate speech analysis because labeling hate speech means deciding what speech violates social norms. But who decides what comprises hate speech? If practitioners and scholars do not understand how people perceive hate speech, some groups' speech will be more protected than others.

To tackle this problem, I propose two principles. First, the transparency principle emphasizes acknowledging *hate speech* as a contested concept and understanding that some people have more power over its definition than others. Providing a position statement that describes why one *hate speech* definition is preferred over others is important to increase the transparency of the model-building process. Second, the inclusion principle underscores that including people who are most likely to be

harmed by hate speech in the creation of automated hate speech detection is crucial so that they can influence data-collection and -annotation procedures (Frey et al., 2020; Halfaker & Geiger, 2020; Katell et al., 2020; Patton et al., 2020). This participatory approach not only improves hate speech detection systems' accuracy but also makes the whole model-building process more democratic.

In practice, taking an intersectional approach (i.e., focusing on power dynamics between and within social groups) is essential to understanding how people's perceptions of hate speech influence their data annotation. For practical and research purposes, assuming that only one form of bias (e.g., racial bias) exists, while other forms (e.g., gender bias) do not exist, might be convenient. However, in reality, these various bias axes intersect, causing one segment of a historically disadvantaged group to suffer from marginalization more than other group members. For this reason, understanding hate speech requires understanding marginalization in both between- and within-group contexts (Kim et al., 2020).

Jae Yeon Kim is Assistant Professor of Data Science at the KDI School of Public Policy and Management, South Korea, and an affiliated researcher of the SNF Agora Institute at Johns Hopkins University, USA. https://orcid.org/0000-0002-6533-7910

References

Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in Information Retrieval* (pp. 141–153). Springer International Publishing.

Al Kuwatly, H., Wich, M., & Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 184–190. https://doi.org/10.18653/v1/2020.alw-1.21

Barocas, S., Biega, A. J., Fish, B., Niklas, J., & Stark, L. (2020). When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 695. https://doi.org/10.1145/3351095.3375691

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics*, 405–415. https://doi.org/10.1007/978-3-319-67256-4_32

Bleich, E. (2011). The rise of hate speech and hate crime laws in liberal democracies. *Journal of Ethnic and Migration Studies, 37*(6), 917–934. https://doi.org/10.1080/1369183X.2011.576195

Blodgett, S. L., Barocas, S., Daumé, III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. https://arxiv.org/abs/2005.14050

Bocian, K., Baryla, W., & Wojciszke, B. (2020). Egocentrism shapes moral judgements. *Social and Personality Psychology Compass, 14*(12), 1–14. https://doi.org/10.1111/spc3.12572

Brown, R. J., Tajfel, H., & Turner, J. C. (1980). Minimal group situations and intergroup discrimination: Comments on the paper by Aschenbrenner and Schaefer. *European Journal of Social Psychology, 10*(4), 399–414. https://doi.org/10.1002/ejsp.2420100407

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77–91). Association for Computing Machinery.

Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Language Variation and Change, 21*(1), 135–156. https://doi.org/10.1017/S0954394509000052

Cohen, C. J. (1999). *The boundaries of blackness: AIDS and the breakdown of Black politics.* University of Chicago Press.

Collins, P. H. (2002). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment.* Routledge.

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning.* https://arxiv.org/abs/1808.00023

Crawford, K. (2017). *The trouble with bias (NIPS 2017 keynote).* YouTube. https://www.youtube.com/watch?v=ggzWIipKraM

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43*(6), 1241–1299.

Davidson, T., & Bhattacharya, D. (2020). Examining racial bias in an online abuse corpus with structural topic modeling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, Data Challenge Workshop.* https://arxiv.org/abs/2005.13041

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online* (pp. 25–35). https://doi.org/10.18653/v1/W19-3504

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). Association for Computational Linguistics. https://doi.org/10.1109/CVPR.2009.5206848

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255

Facebook. (2021). *Community standards: Hate speech.* https://www.facebook.com/communitystandards/hate_speech/

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). Association for Computing Machinery. https://doi.org/10.1145/2783258.2783311

Frey, W. R., Patton, D. U., Gaskell, M. B., & McGregor, K. A. (2020). Artificial intelligence and inclusion: Formerly gang-involved youth as domain experts for analyzing unstructured Twitter data. *Social Science Computer Review, 38*(1), 42–56. https://doi.org/10.1177/0894439318788314

Gajane, P., & Pechenizkiy, M. (2017). *On formalizing fairness in prediction with machine learning.* arXiv. https://arxiv.org/abs/1710.03184

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gelber, K. (2002). *Speaking back: The free speech versus hate speech debate.* John Benjamins Publishing Company.

Geva, M., Goldberg, Y., & Berant, J. (2019). *Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets*. arXiv. https://arxiv.org/abs/1908.07898

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*(4), 215–230.

Greiner, D. J., & Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics, 93*(3), 775–785. https://doi.org/10.1162/REST_a_00110

Halfaker, A., & Geiger, R. S. (2020). Ores: Lowering barriers with participatory machine learning in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1–37. https://doi.org/10.1145/3415219

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3315–3323).

Hinnefeld, J. H., Cooman, P., Mammo, N., & Deese, R. (2018). *Evaluating fairness metrics in the presence of dataset bias*. arXiv. https://arxiv.org/abs/1809.09245

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, i*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Jacobs, J. B., & Potter, K. (1998). *Hate crimes: Criminal law & identity politics.* Oxford University Press.

Jiang, H., & Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics* (pp. 702–712).

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 306–316). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372829

Kasy, M., & Abebe, R. (2020). Fairness, equality, and power in algorithmic decision making. In *Proceedings of International Conference on Web-based Learning Workshop on Participatory Approaches to Machine Learning* (pp. 576–586). Association for the Advancement of Artificial Intelligence. https://doi.org/10.1145/3442188.3445919

Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., & Krafft, P. (2020). Toward situated interventions for algorithmic equity: Lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 45–55). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372874

Kim, J. Y., Ortiz, C., Nam, S., Santiago, S., & Datta, V. (2020). Intersectional bias in hate speech and abusive language datasets. In *Proceedings of the 14th AAAI International Conference on Web and Social Media, Data Challenge Workshop*. Association for the Advancement of Artificial Intelligence. https://arxiv.org/abs/2005.05921

Koushik, G., Rajeswari, K., & Muthusamy, S. K. (2019). Automated hate speech detection on Twitter. In *Proceedings of the 5th International Conference on Computing, Communication, Control and Automation* (pp. 1–4). IEEE. https://doi.org/10.1109/ICCUBEA47591.2019.9128428

Lukes, S. (1974). *Power: A radical view.* Macmillan.

Maass, A. (1999). Linguistic intergroup bias: Stereotype perpetuation through language. *Advances in Experimental Social Psychology, 31*, 79–121. https://doi.org/10.1016/S0065-2601(08)60272-5

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Patton, D. U., Frey, W. R., McGregor, K. A., Lee, F.-T., McKeown, K., & Moss, E. (2020). Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 337–342). Association for the Advancement of Artificial Intelligence. https://doi.org/10.1145/3375627.3375841

Perreault, S., & Bourhis, R. Y. (1999). Ethnocentrism, social identification, and discrimination. *Personality and Social Psychology Bulletin, 25*(1), 92–103. https://doi.org/10.1177%2F0146167299025001008

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

Sellars, A. (2016). Defining hate speech. *Berkman Klein Center Research Publication, 2016(20)*, 16–48. https://doi.org/10.2139/ssrn.2882244

Sen, M., & Wasow, O. (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science, 19*, 499–522. https://doi.org/10.1146/annurev-polisci-032015-010015

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American, 223*(5), 96–103.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–37). Brooks/Cole.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First Association for Computational Linguistics Workshop on Ethics in Natural Language Processing* (pp. 53–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1606

Twitter. (2021). *Hateful conduct policy.* https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Walker, S. (1994). *Hate speech: The history of an American controversy.* University of Nebraska Press.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science* (pp. 138–142).

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Student Research Workshop* (pp. 88–93). Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 78–84). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3012

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access, 6*, 13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394

YouTube. (2021). *Hate speech policy.* https://support.google.com/youtube/answer/2801939?hl=en

Zhou, X., Sap, M., Swayamdipta, S., Smith, N. A., & Choi, Y. (2021). *Challenges in automated debiasing for toxic language detection*. arXiv. https://arxiv.org/abs/2102.00086

Digital
Communication
Research.de

Recommended citation: Stoll, A. (2023). The accuracy trap or How to build a phony classifier. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 371–381). Digital Communication Research. https://doi.org/10.48541/dcr.v12.22

Abstract: This guide explains, in four steps, how to build a phony text classifier using *supervised machine learning*—a classifier that is absolutely unreliable but looks outwardly sophisticated and attractive. You might enjoy this text if one or more of the following statements apply to you: You are interested in the automated identification of hate speech or related content in online discussions, as long as it looks good; you want to do something with machine learning to impress your peer group, but you do not have the nerve to dig deep into this field as well; you are either a somewhat sneaky or a humorous person. Of course, however, if you are a good and decent researcher, you might also take hints from this text on how *not* to step into the accuracy trap and how not to fall for the tricks of phony classification.

DOI 10.48541/dcr.v12.22

Anke Stoll

The Accuracy Trap or How to Build a Phony Classifier

1 What is this about?

The approach of *Supervised Machine Learning* (SML) raises high expectations for the automated identification of hate speech and all related concepts. These expectations are not surprising, because, after all, *machine learning* is *artificial intelligence*, right? And artificial intelligence is supposed to be groundbreaking. This is already suggested by its very name. And why else would it be so hyped? So, if we do machine learning, build a classifier, train a model, can we not expect groundbreaking results? Of course, as always, it's not that simple. But at least, there is a safe way to build classifiers that *seem* to meet all these expectations—that achieve apparently outstanding results in detecting even the most complex concepts from only text information, based on only a few training instances, and they perform at least as well as the human annotators whose struggle is mirrored in poor values for intercoder agreement.

To build such a phony, hypocritical classification model, we must follow only four simple steps. First, we need a complex and rather elusive concept to identify from text, such as *hate speech*. In the second step, we have to ensure that our sample includes only a few relevant cases. Third, we must stick to the basic estimation functions of SML, including logistic regression or support vector machines.

And, in the fourth step, we must finally use only the standard metric *accuracy* to evaluate the model's performance and avoid any further in-depth performance evaluation if possible.

Following these four steps, we can be quite sure that our classifier will learn nothing about detecting hate speech at all and that it will be absolutely unreliable. But, without a closer look, it will not be noticed since our phony model meets all the expectations at first glance. The following chapters will show you in more detail how to build such a phony classifier that looks nice from the outside but is absolutely useless.

2 Step 1: Choose a contentious concept to classify

Depending on the research area, the SML approach with text data is applied to different issues of interest, including the all-time classics *spam* versus *ham* in emails and *sentiment* in product reviews. The issue of hate speech identification particularly concerns many areas of research and practice and is meanwhile an established research issue in many different fields. In communication studies (and related social sciences), the automated identification of content is not exactly a classical research question (yet), but it is rather relevant from a pragmatic point of view: as an approach of *automated content analysis,* SML is supposed to support the costly and elaborate manual measurement of content for huge amounts of text (e.g., Boumans & Trilling, 2016; Wilkerson & Casas, 2017; Scharkow, 2017; Sommer et al., 2014; Scharkow, 2013). Given its fancy name, the SML method itself is quite a bummer since it is actually just predictive modeling. *Predictive modeling* (and SML, too) means that a dependent variable is supposed to be predicted by a set of independent variables. In SML, the independent variables are called *features*. If the dependent variable is categorical, it is called a *class* or *category*—and the approach is called *classification*. In *text classification* (or *document classification*), the dependent variable is the content or the meaning of a text—for instance, whether a given text includes hate speech. (For a practical introduction to SML, see Müller & Guido, 2017; Géron, 2017).

To identify a text's content or meaning automatically, text characteristics are used as features of prediction. The most essential text characteristics are the words that a text does or does not include. This might sound trivial at first, but it is

hard to deny that sentiments, topics, or hate speech are expressed to a significant amount through the use of words. It is therefore reasonable to assume that words can cover some variance of content or meaning. Nevertheless, one can also quickly think of examples where this assumption falls short, where a significant part of the meaning is captured—for example, in the context of the situation or within the person who reads and processes a text. It shows that, for hate speech and its related concepts, this situation often is the case (e.g., Ross et al., 2017; Waseem, 2016).

Overall, a classifier's performance will depend on the modeled statistical relationship between the text features and the text category. When we stick to basic SML (instead of deep learning, for example), that is the statistical relationship between words and meaning, in many cases. But if words have different meanings in different contexts and important variables are missing, that approach starts to weaken. Fortunately, a phony classifier does not have to learn an actual, robust statistical relationship. Ironically, the exact opposite is the case (see also steps 2 and 3 in this guide). For the *hate speech* concept, therefore, one important condition for a phony classifier is fulfilled: an elusive concept supposed to be predicted based on words, even if other important information is probably needed to determine whether a text is considered *hate speech* or not. Luckily, this does not only apply to hate speech but to many concepts that interest communication scholars since they are seemingly impossible to determine without pages of instructions and hours of training (Krippendorf, 2018; Früh, 2015).

3 Step 2: Draw an imbalanced sample with few relevant instances

If we were interested in decent text classification, likely, we would fall over one of the many obstacles that prevent our finding the hoped-for relationship between features (here, words) and meaning (such as hate speech). Many of these obstacles come with the sample itself. But to build a phony classifier, we need not deal with any of these. The only crucial condition is that the sample for training includes just a small portion of relevant cases. Again, hate speech detection is a suitable use case for phony classification since hate speech does not appear in a great number of instances in a random sample of user comments, Tweets, or related text types (e.g., Papacharissi, 2004; Davidson et al., 2017; Coe et al., 2014; Zampieri et al., 2019; Risch et al., 2021; Friess et al., 2021). Since any pattern is, obviously, hard to learn from

only a few examples, such *rare event*s are quite unpopular in the whole research field of machine learning (see Haixiang et al., 2017, for an overview). Text data, however, are particularly painful since natural language is a very heterogeneous data type, so text data require huge sample sizes, and it takes forever until patterns form and can be tracked down by any statistical model (Mandelbrot, 1961; Jurafsky & Martin, 2009; Schütze et al., 2008). If we choose a concept such as *hate speech*, however, the heterogeneity problem is somewhat unavoidable since words can express hate in so many ways. Plus, many of the words in hate speech comments are not unambiguously hateful, meaning that they are also used in comments that are not hate speech at all. (See Baden in this volume for an overview of such issues.)

Researchers sometimes address the issues of rare relevant instances and high data heterogeneity by drawing more narrow samples—for example, debates with a certain hashtag, topic, or time span that offer more potential for controversy. Thus, the proportion of relevant cases is often higher, and the text data are not that heterogeneous. In this way, a classifier becomes more likely to learn an actual relationship between words and meaning, though such a classifier would probably not be applicable to other contexts. Luckily, all this struggle is not a problem for phony classification—rather, the opposite applies. We only need one further condition in the sample, which, fortunately, is usually a consequential problem of rare events: an *imbalanced* (unbalanced) derivation of the classes in our sample. *Imbalanced data* here means that comments that include no hate speech are clearly overrepresented (e.g., Stoll, 2020). In conclusion, we are once again blessed with the classification of hate speech since it appears infrequently (at least in manageable sample sizes) and seems rather infrequent compared to instances that do *not* include hate speech.

4 Step 3: Choose a weak classification function

In SML, the classification function models the relationship between features (independent variables) and a category (dependent variable). For text classification, these functions must be able to handle a huge amount of data and a huge number of features at the same time. In SML, a classification function can usually be described as a decision boundary between the instances of Category A (e.g., HATE) and Category B (e.g., NO HATE). Obviously, the classification approach is

promising when documents of Category A can be distinguished from documents of Category B, given the words that the text documents (e.g., user comments or Tweets) include. For hate speech, as already discussed in this chapter, this distinction is not always the case since documents of Category A and Category B have many words in common. This problem will most often lead to confused classifiers and unsatisfactory results (Davidson et al., 2017; Waseem & Hovy, 2016).

Luckily, the uncertainty of prediction is a fortune for a phony classifier. In addition to the confusion caused by word overlap between the categories, the small number of relevant instances causes the classifier to finally quit. As Step 2 described, training the classifier on an imbalanced sample where the relevant category (HATE) is underrepresented is a crucial requirement because many classification functions—including *support vector machines*, *logistic regression,* and *decision trees*—tend to predict the major category in the training data in uncertain cases. Indeed, the smaller the data basis, the smaller the chance to find some significant word distribution patterns and the higher the chances that a classifier gives up (e.g., Haixiang et al., 2017; Denil & Trappenberg, 2010; Stoll, 2020). From a statistical perspective, that strategy is straightforward because, in this way, a model will predict the right category in most cases—namely, the overrepresented category in the training data (NO HATE). If we have only 10% of instances annotated as hate, a classifier would be right in 90% of cases if it always predicted NO HATE. However, that rate does not mean hate speech is predicted correctly in 90% of cases. Sometimes, the relevant category HATE will not be predicted at all, which would be an unsatisfactory result, of course, if we were interested in building a model that can actually detect hate speech. During a decent and transparent model evaluation, the scam would be noticed quickly unless we rely only on the popular *accuracy* metric, as the next step describes.

5 Step 4: Stick to the accuracy evaluation metric (only!)

Classification models are usually trained on a subset of an annotated sample, called the *training set*. Then, they are tested and evaluated on a separate sample, called the *test set*. The model performance is measured by how well the predicted values match the true (manually annotated) values on the test set. The most obvious measurement for model quality is the *accuracy*, meaning the

percentage agreement between the predicted and the annotated values for the dependent variable—here, *hate speech*. The higher the agreement between human annotation and classification, the better. If, for example, only 10% of comments in the sample are hate speech, a classifier could achieve 90% accuracy by only predicting the major category NO HATE without detecting one single hate speech comment. Thanks to a long journey through the method of manual content analysis, communication scholars are already critical of the percentage agreement and prefer the relentless *Krippendorff's alpha*, which also reveals disagreement in rare categories (Krippendorff, 2018; Lombard et al., 2002; Vogelgesang & Scharkow, 2012). Luckily, Krippendorff's paper is yet unknown in the research field of machine learning, so we will probably not be obliged to consider it for reliability measurement. Nevertheless, other established measurements and procedures in the research field of machine learning are quite capable of circumventing the accuracy trap. But do not worry, these other options still do not mean we must give up.

In SML, common measures to evaluate a model are *recall*, *precision*, and the *F1 score*, as a balanced average of both measures (e.g., Powers, 2011). In default setting, all of these measures are used to evaluate a classifier's performance in *one* category, meaning HATE and NO HATE each. A high recall value for the hate category would actually be nice for a hate speech classifier because it would show that many of the instances that have been manually labeled as *hate speech* could have been identified. A phony model, on the contrary, would always have a low recall for the relevant category HATE since it would not really learn how to detect hate speech. Furthermore, acceptable precision in the relevant category would be preferable for an actual hate speech classifier since it would show that the model is not always wrong when it classifies an instance as HATE. A phony model, however, would not learn how to identify hate speech and would, therefore, make many mistakes, which would be reflected in low precision for the category HATE (e.g., Stoll, 2020). So, just reporting recall and precision for the relevant category would be a safe and easy way to expose a phony model. In other words, we certainly do not want to do that! However, if we have followed steps 1 to 3, the recall and precision values for the NO HATE class will most often will be quite nice. To make an impression, these values should be reported in any case, alongside a remarkable accuracy score (how sneaky!).

Not only the evaluation of the test set, which is part of the sample, can reveal an unreliable model. Also, the evaluation on a new data set can be dangerous. Communication studies, meanwhile, have established applying and rechecking a developed instrument for automated content analysis on a completely new data set (e.g., Grimmer & Stewart, 2013). This demand also concerns classifiers—thus, we are still not off the hook. Indeed, this demand is very reasonable since phony models (or models that only learned hate speech from a certain debate) can be exposed without much consideration of in-depth model evaluations. Because we do not want our classifier to be busted, this demand is—of course—annoying. Fortunately, not all data sets considered for external evaluation are actually a cause for concern. If a phony classifier is applied to a new data set, which includes also only a few relevant cases (here, hate), it will be accurate to a high percentage again! Since the model would not have learned to identify hate speech, it would have learned the derivation of hate speech in the training data. If the new data set had a similarly imbalanced derivation, there is nothing to worry about. Good luck!

6 Conclusion

As this guide shows, building a phony classifier that looks outwardly powerful but has learned nothing about hate speech detection at all is fairly simple. Many of the important criteria for phony classification and stepping into the accuracy trap come with the hate speech phenomenon itself. First, people can eternally debate whether a statement should be categorized as *hate speech*—most of all because important information is captured in a context or personal perspective. Second, however, for an ordinary text classifier, none of this information is available, only text. Third, in a random sample of user comments, Tweets, or related data sources, the number of relevant instances from which a machine learning model could learn hate speech is rather small and—in relation to instances that do not include hate speech—rather underrepresented. As a result, classifiers often come out poorly equipped from the training process, having learned hardly more than the imbalanced class derivation in the training data. If we ignore all these flaws, we can still achieve impressive-looking results (see Step 4) that we legitimately expected from something called *machine learning* instead of boring *statistics*. This is because the described circumstances lead to model results, which—

when measured with the right metrics—look like an amazing performance. And at first glance, one could almost think the problem of automated hate speech identification has been effectively solved with logistic regression.

The bad news is that, upon a closer look, the machine learning method *is* just statistics. And, consequently, we are still stuck with the same questions and pitfalls that social scientists already know well enough: *Which information do I need to explain a phenomenon?* versus *These are the independent variables that I am capable of measuring*, or, *Which sample would be suitable for my research questions?* versus *I can only afford a student sample*. Nevertheless, this realization also shows us that machine learning is not far from well-known inferential statistics and, therefore, is predestined to be a further comfort zone for social scientists—only without *p*-values and SPSS.

Anke Stoll is a research associate at the Institute for Social Sciences at the Heinrich Heine University in Düsseldorf, Germany.

References

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512-515.

Denil, M., & Trappenberg, T. (2010). Overlap versus imbalance. In A. Farzindar & V. Kešelj (Eds.), *Advances in artificial intelligence. Canadian AI 2010. Lecture notes in computer science* (vol. 6085) (pp. 220–231). Springer. https://doi.org/10.1007/978-3-642-13059-5_22

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication, 38*(5), 624–646. https://doi.org/10.1080/10584609.2020.1830322

Früh, W. (2015). *Inhaltsanalyse: Theorie und Praxis* [Content analysis. Theory and practice]. UTB.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. *Structure of Language and Its Mathematical Aspects, 12*, 190–219.

Müller, A. C., & Guido, S. (2017). *Einführung in Machine Learning mit Python. Praxiswissen Data Science* [Introduction to machine learning with Python: A guide for data scientists]. O'Reilly Media.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://arxiv.org/abs/2010.16061

Risch, J., Stoll, A., Wilms, L., & Wiegand, M. (2021). Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In J. Risch, A. Stoll, L. Wilms, & M. Wiegand (Eds.), *Proceedings of the GermEval 2021 Workshop on the identification of toxic, engaging, and fact-claiming comments. 17th Conference on Natural Language Processing KONVENS 2021* (pp. 1–12). Netlibrary. https://doi.org/10.48415/2021/fhw5-x128

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Eds.), *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication (Bochum)* (pp. 6–9). Bochumer Linguistische Arbeitsberichte. https://doi.org/10.48550/arXiv.1701.08118

Scharkow M. (2013). *Automatische Inhaltsanalyse* [Automated content analysis]. In W. Möhring & D. Schlütz (Eds.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* [Handbook of standardized survey methods in communication science] (pp. 289–306). Springer VS.

Scharkow, M. (2017). Content analysis, automatic. In *The international encyclopedia of communication research methods* (pp. 1–14). John Wiley & Sons. https://doi.org/10.1002/9781118901731.iecrm0043

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

Sommer, K., Wettstein, M., Wirth, W., & Matthes, J. (Eds.). (2014). *Automatisierung in der Inhaltsanalyse* [Automation in content analysis]. Herbert von Halem.

Stoll, A. (2020). Supervised Machine Learning mit Nutzergenerierten Inhalten: Oversampling für nicht balancierte Trainingsdaten [Supervised machine learning with user generated content: Oversampling for imbalanced training data]. *Publizistik*, *65*(2), 233-251.

Vogelgesang, J., & Scharkow, M. (2012). Reliabilitätstests in Inhaltsanalysen: Eine Analyse der Dokumentationspraxis in *Publizistik* und *Medien & Kommunikationswissenschaft* [Reliability tests in content analyses: The documentation of reliability in *Publizistik* and *Medien & Kommunikationswissenschaft*]. *Publizistik*, *57*(3), 333–345. https://doi.org/10.1007/s11616-012-0154-9

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 38–142). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5618

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93). https://aclanthology.org/N16-2013.pdf

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*, 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Zampieri, M, Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 75–86. https://doi.org/10.18653/v1/S19-2010

Digital
Communication
Research.de

Recommended citation: Laugwitz, L. (2023). The right kind of explanation: Validity in automated hate speech detection. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 383–402). Digital Communication Research.
https://doi.org/10.48541/dcr.v12.23

Abstract: To quickly identify hate speech online, communication research offers a useful tool in the form of automatic content analysis. However, the combined methods of standardized manual content analysis and supervised text classification demand different quality criteria. This chapter shows that a more substantial examination of validity is necessary since models often learn on spurious correlations or biases, and researchers run the risk of drawing wrong inferences. To investigate the overlap of theoretical concepts with technological operationalization, explainability methods are evaluated to explain what a model has learned. These methods proved to be of limited use in testing the validity of a model when the generated explanations aim at sense-making rather than faithfulness to the model. The chapter ends with recommendations for further interdisciplinary development of automatic content analysis.

DOI 10.48541/dcr.v12.23

Laura Laugwitz

The Right Kind of Explanation

Validity in automated hate speech detection

1 Automated content analysis: Mind the epistemological gap

As one of the core methods of communication research, content analysis has, for decades, provided the tool to describe, analyze, and compare content conveyed by the media (Krippendorff, 2019; Lacy et al., 2015). In conjuncture with growing amounts of digital communication, accessible tutorials, and evolving computing capacities (van Atteveldt & Peng, 2018), content analysis of text in particular is increasingly supported by computational methods (boyd & Crawford, 2012; Trilling & Jonkman, 2018) to analyze larger amounts of data faster and in a more standardized fashion. Hate speech detection is one of the fields in which automated content analysis stands to reason, since it is not simply the subsequent analysis of hateful communication that is of interest in research but also its quick identification (e.g., Davidson et al., 2017), moderation (e.g., Paasch-Colberg et al., 2020), and prevention (e.g., Schmitt et al., 2018). As other contributions in this collection show, the identification of hate speech is a challenge due to varying definitions (see Part 2), context (see Litvinenko), latent features (see Baider; Becker & Troschke), linguistic limits (see Baden), and bias (see Kim & Stoll). Computational methods add another level of challenges, as researchers not only

need to learn how to choose and implement these methods with regard to hate speech detection but also to understand and evaluate their results. While it is reasonable to collaborate with machine learning experts or computer scientists to implement computational methods well, there are fundamental differences in how computer science and social sciences approach the production of knowledge and, thus, how each evaluates the models they build.

Validity and reliability in computational methods are key issues for communication research (Scharkow, 2013; van Atteveldt & Peng, 2018), whereas machine learning experts are focused mainly on a sub-category of reliability, namely reproducibility, of their work (Henderson et al., 2018; Lipton & Steinhardt, 2019; Stodden, 2010). Since there is no clear equivalent for testing validity in text classification, communication researchers risk drawing the wrong inferences from automatically labeled data if they do not develop methods to ensure that such labels are based on "what words mean in the context of their use" (Krippendorff, 2019, S. 218).

These different focal points on criteria for research quality are rooted in different epistemologies: Communication research is largely conducted through the lens of critical rationalism in which a hypothesis's acceptability is tested *a priori* through logic and comparison with other theories as well as *a posteriori* through empirical tests (Chauviré, 2005). Machine learning, by contrast, mainly operates within a technocratic paradigm, dismissing the idea that *a priori* knowledge about the behavior of a program is possible and instead relying on gaining *a posteriori* knowledge through testing (Eden, 2007). It follows that the former understands quality as the production of reliable, valid, and intersubjectively comprehensible knowledge (Brosius et al., 2009), and the latter understands quality as the development of a satisfactory, reusable application (Stodden, 2010). Thus, these different approaches result not only in different quality criteria for research but also lead to various points of friction in interdisciplinary work.

In this chapter, I will examine the consequences of such epistemological differences, then focus on different quality criteria and how this may be alleviated when using supervised text classification for hate speech detection. In gathering various established and novel methods to establish validity in supervised text classification, I will show that current explainability approaches in development by computer scientists provide a useful starting point for deepening the understanding of a model's decision process. However, only a few of these approaches satisfy social science's imperative to examine a model's validity. This leaves

a void that ought to be filled via diligent collaboration between communication and computer scientists.

2 Validity of automated content analysis for hate speech detection

Automated content analysis can merge communication studies' manual content analysis and machine learning's supervised text classification; a model is trained to reproduce the labeling of concepts developed for and coded within a manually created corpus (Boumans & Trilling, 2016; Scharkow, 2013). In manual content analysis, human coders are given instructions in the shape of a codebook and are trained thoroughly for their classification task. The main strategies to ensure content validity—suggesting that the theoretical constructs are exhaustive and adequate (Krippendorff, 2019) – involve creating these codebooks based on a comprehensive literature review, training coders, improving the codebooks through their feedback, and checking for the use of catch-all or open categories after classification. These strategies, as well as intercoder reliability scores, reasonably create qualitative and quantifiable confidence that coders have integrated the knowledge derived from theory and research into their mental models. Additionally, the results may be compared with results from similar studies using the same or other methods (Krippendorff, 2019). Ensuring validity in content analysis specifically for hate speech is a challenge due to different reasons. First, hate speech is a complex construct, and separating it from adjacent concepts, such as incivility, toxicity, or offensive speech in theory, is still in progress; doing so in practice has only been attempted by a few researchers (e.g., Stoll et al., 2020). Second, some dimensions of hate speech show a manifestation in specific words (Davidson et al., 2017), whereas others are more latent (Nielsen, 2002) or contradictory (van Aken et al., 2018), such as generalizations or irony. Even in thorough manual content analysis, intercoder reliability can vary immensely (Poletto et al., 2020; Ross et al., 2017), which poses a challenge in reliably measuring the difficult concept of hate speech. Third, and most important to this paper, translating any attributes or dimensions of hate speech into technologically traceable features is a challenge that is rarely recognized. In a literature review, Fortuna and Nunes (2018) showed different strategies currently in use for hate speech detection. One strand leverages existing generic methods from natural language processing,

such as topic classification, sentiment analysis, named entity recognition, and deep learning, to identify hate speech. A second, much smaller strand identifies specific linguistic features, such as othering language, objectivity-subjectivity, declarations of superiority of an in-group, and particular stereotypes. The latter strand approximates operationalizations from manual content analysis closer than the former; however, the former is much more common.

In supervised text classification, no fixed rules or instructions are given to the machine learning model. Rather, it derives classification rules inductively from previously coded material (Scharkow, 2013). While multiple strategies to measure reliability for automated content analysis and to increase reproducibility have recently been suggested (Krippendorff, 2019; Mitchell et al., 2019; Pineau, 2020; Scharkow, 2013), testing the validity of supervised text classification has yet to be expanded. Most simply, some researchers rely on manual coding as the gold standard, and rule that validity is established if the automated results are similar enough to the manual classification (Lee et al., 2020). The assumption here is that quality can be assured through the creation of a valid and reliable manually coded dataset, since an algorithm will then simply reproduce these classifications. However, computer science itself is currently raising doubts about whether models actually learn content-related features at all, or instead are trained on spurious correlations and artifacts in the dataset (Lapuschkin et al., 2019), raising the issue of validity without explicitly naming it. Other scholars, then, suggest applying their model to a second dataset to test the validity of inferences (Pilny et al., 2019). Beyond these, some communication scholars have also attempted to examine content validity. To show whether a model has learned to identify concepts previously derived from theory, the weights for individual features can be examined (Stoll et al., 2020).

Examining examples from sentiment analysis, which aims to automatically identify mood in text and is occasionally used as a proxy to identify hate speech, Liu and Avci (2019) claimed that models may assign a negative mood to text as soon as it contains identity terms, such as "Jew" or "Black." Similarly, a hate speech classifier may learn to classify any sentence containing the term "Islam" as toxic (Waseem & Hovy, 2016). While this type of error may be acceptable for an application designed to help companies in the private sector identify potentially problematic discussions, such a lack of validity is fatal in communication research, as inferences based on invalid classifications will result in flawed inferences. More

recently, it has been shown that the reasoning of well-performing models for hate speech classification tasks does not necessarily align with the reasoning that coders for manual classification have provided (Mathew et al., 2020), adding to the limited trust in the validity of these models. Relying on a manually coded dataset with high validity is not sufficient to examine whether the theoretical constructs informed the model's classification decision. Applying the same model to another dataset may hint at its capacity to generalize. However, trying to understand how a model made a decision, for example, via feature weights, appears to be the most fruitful and necessary strategy to date in examining the content validity of supervised text classification. In the following chapter, I present strategies currently explored in machine learning that intend to show the reasoning behind a model's classification decisions.

3 Explaining supervised text classification

A recurring theme in machine learning is the question of why a model made a specific decision (local explanation) or how it works in general (global explanation), which is currently mostly found under the umbrella term "explainability." Efforts to increase the explainability of automated results (Samek & Müller, 2019) have gained more relevance in machine learning since the wider use of automated decision-making in business, banking, and the public sector (Mittelstadt et al., 2016). Following the introduction of the general data protection regulation in 2018, users even have a right to be provided with an explanation for an automated decision (§ 14 2 g GDPR). Two main strategies are currently pursued in explainability research (Vilone & Longo, 2020): Model-agnostic methods are meant to provide generic solutions, creating explanations that do not require access to the model itself. In using solely the input and output of a model, they can be considered reverse engineering. Model-specific methods examine particular aspects of a model, such as revealing word relations in different layers of a neural network. They require access to the model and are dependent on its functionality. Beyond explainability strategies, interpretable methods use *ab initio* algorithms that can be understood by humans (Rudin, 2019), such as linear classifiers, Bayesian classifiers, or support vector machines. Based on the systemic literature reviews by Guidotti et al. (2019) and Vilone and Longo (2020), five model-agnostic methods,

five methods specific to neural networks, and three interpretable models can be identified. These solutions aim to explain models for supervised text classification. They will be summarized to show the general breadth of solutions and give communication scholars an idea of the state of research in machine learning.

3.1 Model-agnostic methods

Developing an additional simplified model based on the input and output of the original model is the general strategy for model-agnostic methods. More specifically, the partition aware local model (PALM) consists of two kinds of models: a meta-model partitions the training dataset, and then individual sub-models approximate local patterns of the partitions (Krishnan & Wu, 2017). The meta-model is a decision tree that can be used to compare single misclassifications with the relevant training data (if available), and thus offers human users an intuition for the relation between training input and classification.

With a similar focus on proximity, local interpretable model-agnostic explanations (LIME) trains a linear classifier for a single classification by approximating further cases from the immediate neighborhood of the example (Ribeiro et al., 2016). Thus, complex models are broken down into single, locally interpretable models. Anchors is an extension of LIME that leverages if-then-rules that anchor an explanation locally to a point at which a change of values to other features of that instance does not lead to a different classification (Ribeiro et al., 2018). Similar instances almost always share the same classification, thus providing examples of how features are relevant to a classification.

Simple rules are also used in a model explanation system (MES), which assumes that explanations are simple logical statements and uses a Monte Carlo algorithm to find the best explanation for a single classification via a scoring system (Turner, 2016). Although it is intended to work for text as well, Turner (2016) has only provided examples of computer vision and credit scoring (tabular data). Whether meaningful automatically generated explanations for text can be achieved with MES is an open question.

The last model-agnostic approach is based on game theory, using the idea that each feature represents a player, and each classification represents the profit. Shapley values indicate how this profit must be fairly distributed between features.

For this purpose, every possible feature combination and its effect on classification are compared. Thus, if all classifications are considered, a statement can actually be made about the global relevance of each feature as well as the local relevance of the feature in a single classification. Practically, these values are impossible to compute due to the large set of features and their possible combinations; however, simplified versions of this approach are used as intuition in various explainability methods (e.g., Chen et al., 2019).

3.2 Model-specific methods

Methods that inspect or retrace the partial mechanics of a model are called model-specific. Each approach depends on the model itself; thus, there is no general solution that can be transferred to a different type of model. However, all relevant solutions from the literature have been developed for some versions of a neural network. For example, a rationale generator is trained in parallel with a neural network, learning to select a subset of the input sequence as an explanation for classification (Lei et al., 2016). The rationale then contains a reduced set of meaningful words, which should result in the same classification as the original input sequence if given to the classifier.

This strategy to identify salient input features is also applied in DeepLIFT (Deep Learning Important FeaTures). Here, the principle of layer-wise relevance propagation traces a single classification of a neural network backwards through said network. DeepLIFT then analyzes the difference in the activation values of single neurons for that input-output pair compared to a reference input-output pair (Shrikumar et al., 2017) and indicates which features (e.g., individual words) were most in favor of a classification or its opposite. This approach has the potential to provide counterfactual explanations if the reference input-output pair is intentionally chosen. However, for the application to text, it seems customary to simply choose zero values (Lertvittayakumjorn & Toni, 2019; Sundararajan et al., 2017).

Integrated gradients explain a neural network by analyzing its sensitivity to differences in input (Sundararajan et al., 2017). They create a sequence of gradients leading from the baseline to the input and compute their average, thus measuring the correlation between the uncertainty in the output of a classifier and its input.

With the recent and extremely rapid success of transformer models, the visualization of attention and hidden states has gained popularity for explanatory purposes (see van Aken et al., 2020). Transformer models are a special form of deep neural networks that first learn basic language structures before being trained in specific tasks (see Minaee et al., 2020). Van Aken et al. (2020) proposed considering the feature embeddings of individual layers to visualize the learning process of a transformer model for individual classifications. For this purpose, the vectors with which the individual inputs are technically represented are reduced in dimension after each layer with principal component analysis and mapped on a two-dimensional surface. Across the layers, the proximity of different words becomes apparent (see van Aken et al., 2020), exposing the inner structures of the neural network. It is then left to the researcher to decide on the layer in which a structure is clear enough to be used as an explanation for a classification.

Another approach that employs human interaction is concept-based explanations. In testing with concept activation vectors, people are asked to choose examples and counterexamples for certain concepts (e.g., stripes in pictures) after training a model (see Kim et al., 2018). An additional linear classifier will then be trained to discriminate between activations for each set of examples, generating global explanations for the influence of concepts on classes. However, these concepts depend on what the researcher chooses in terms of content, and it is unclear whether they cover all concepts relevant to the model. Further development of this method adds a step of unsupervised learning, automatically extracting concepts that are sufficiently predictive for classification (see Yeh et al., 2019).

Communication research might benefit from experimenting with a combination of the analysis of different layers in neural networks, where different linguistic concepts are also recognized in different layers (van Aken et al., 2019), and the concept-based analysis of Yeh et al, (2019). Different linguistic layers could be responsible for different concepts. To detect hate speech, for example, it would be possible that manifest insults could be identified at early levels, while latent concepts such as dehumanization would only be identified in later layers. Empirical testing of this assumption could be extremely valuable.

3.3 Interpretable models

Instead of using complex neural networks or proprietary systems for text classification, researchers can also make use of models that are interpretable by default. *Ante hoc* methods are intended to keep models explainable from the beginning and are therefore also called white box or transparent models (Rosenfeld & Richardson, 2019). They include decision trees and decision rules or k-nearest neighbors, as well as discrete choice probabilities, such as logistic regression or Naive Bayes (Molnar, 2020). However, these models are typically heavily domain-specific and, if the data are not well structured and clear, they can require an enormous amount of computational effort (Rudin, 2019). In a recent and commendable study on incivility and impoliteness in text, Stoll et al. (2020) used Naive Bayes for global explanations that give weight to individual words. The weight of a feature is calculated by its probability of appearing in a given classification. The global explanation thus outputs a list of features that are relevant for a class. Risch et al. (2020) also used this strategy in comparison with explainable models and found that the Naive Bayes model showed the lowest performance. Unlike Stoll et al. (2020), who used various preprocessing methods, however, Risch et al. (2020) did not show whether further steps were taken to improve the data, which would be necessary for interpretable models according to Rudin (2019).

Instead of these weighted features, interpretable models can also be used to create prototypes. Bien and Tibshirani (2011) developed the prototype selection approach in which, instead of focusing on reducing the number of features to a manageable amount, the data itself is bundled by selecting a prototype from the neighborhood of each instance that has the same label. The authors aimed to have as few prototypes as possible and ensure that no instance had a prototype with a different label. Prototype selection requires inference from the researchers, since it does not show which specific features of the prototype were relevant for its selection.

This lack of causal explanation is addressed in the Bayesian case model (BCM) by first clustering the data and then generating prototypes as well as feature weights for these clusters (Kim et al., 2015), thus providing global explanations. However, if too many clusters are formed, both the computational time and the number of explanations are too high, which in turn no longer allows for interpretability. Guidotti et al. (2019) pointed out several improvements for

BCM: humans can interact with the model to improve the prototypes (Kim et al., 2015), or instances in which the classification does not fit well into the model can provide counterexamples (Kim et al., 2016). Subsequently, the overall strategies will be investigated with respect to how they may be leveraged to examine content validity in automated content analysis.

4 Using explanations to examine validity

Model-agnostic methods, such as LIME, Anchors, and MES, aim to explain individual classifications, whereas PALM identifies partial patterns, and Shapley values have the potential to trace the weight of features across the entire model, where it is not for the computational limits. However, given the fact that these solutions do not actually inspect or retrace the mechanics of the initial model, their usefulness regarding validity is limited. Whether the model has identified the same concepts technologically that have previously been defined for manual analysis remains unknown. While they may be useful for identifying discrepancies in classifications, they do not make use of but instead approximate the initial model (Rudin, 2019), thus creating an additional layer of uncertainty instead of alleviating it.

Model-specific methods provide a tool to partially inspect the model's validity; the fact that they create insight into the internal mechanics of a model suggests that they may be used to examine whether the theoretical concepts have been transferred to the technological operationalization. However, it is not sensible to infer how the model works as a whole from explanations for individual classifications (Mittelstadt et al., 2019), which would be an inductive fallacy. In fact, these methods also do not contribute to giving users a more comprehensive understanding of model behavior (Lertvittayakumjorn & Toni, 2019), may also give misleading explanations (Rudin, 2019), and should thus be used only with proper contextualization and caution.

Interpretable models provide "their own explanations, which are faithful to what the model actually computes" (Rudin, 2019, p. 1) and are thus especially interesting to researchers already competent in statistics. Their simplicity can offer insight into how the model has transferred theoretical operationalizations into technical features so that their explanations can actually act as indicators

for content validity. Note that some researchers have also critiqued Rudin's assumption that models can be inherently interpretable yet do not provide any data to substantiate their critique (Jacovi & Goldberg, 2020). Nevertheless, well-performing interpretable models also require time and effort, so the costs and benefits of the research project in question must be weighed. Due to the data structure and the inherent ambiguity of text, interpretable models for text classification currently do not receive much attention. Even Kim et al. (2015) who clearly advocated for interpretable models in 2015 and 2016, have moved on to developing model-specific methods by 2018. Although interpretable methods show the most promise for validity checks, interpretable methods in general are underrepresented in explainability research (Vilone & Longo, 2020). The aim of machine learning to develop generalized solutions (Fortuna et al., 2020) that can be applied to many problems does not necessarily overlap with that of the social sciences to consider problems in context.

In summary, existing strategies developed to explain the overall functioning or individual decisions of a text classification model offer limited help in examining a model's validity. Model-agnostic explanations may be used to gain some intuition when models are complex or proprietary but can be considered insufficient for a validity check. Similarly, model-specific explanations do not satisfy this use case either. While they access the model itself to provide explanations, they rarely explain it in its entirety, and local explanations should not be used to infer the functionality of the model as a whole. Interpretable models show the most promise for our use case. If trained carefully and with sufficient domain knowledge, they perform well and provide explanations that are appropriate for testing content validity. Nonetheless, since both explainability methods for text (as opposed to images or tabular data) and interpretable models are rare in the current body of research, an opportunity to collaborate beyond a simple splitting of tasks in automated content analysis emerges for communication scholars and computer scientists.

5 A call to develop methods to establish validity of automated content analysis

A critical rationalist perspective on automated content analysis substantiates the need to explain how a model works to examine its validity. Failure to provide adequate explanations creates opacity within the scientific process, preventing researchers from ensuring that their model has learned on informative features that sufficiently consider context instead of learning on artifacts or spurious correlations. Hence, only validated models should lead to inferences about the data's context. Whereas standardized content analysis has established strategies to strengthen and examine content validity, no such strategies have been established for supervised text classification. In creating a codebook informed by theory and empirical research, comprehensive coder training, feedback loops, and discussions in training sessions, as well as reliability scores, researchers gain confidence about the validity of their data and subsequent inferences. An automated model, however, is not involved in gaining a shared understanding of what is supposed to be coded; instead, it merely aims to mimic. The strategies to strengthen and examine validity thus look different for supervised text classification. As argued above, validity can be strengthened by using interpretable methods and examining whether the features that a model has learned preserve the context of meaning. Explainability methods partially enable such an examination; however, their current applications are not specific enough for scientific use.

The development of the explainability methods discussed above has mostly been motivated by the need to establish trust, identify bias or errors, and prevent damage by a faulty system. The quality of these methods, in line with a technocratic paradigm, tends to be evaluated *a posteriori*, for example, via a user's reaction, feedback, or subsequent performance (cf. Gilpin et al., 2019)—framing quality as the *plausibility* of the explanation. However, to verify the validity of a model, explanations cannot be measured with regard to their effects on users (see Herman, 2017). What matters in examining validity is not an explanation's effect on a user but that it explains a model concisely. Jacovi and Goldberg (2020) identified a difference between the plausibility and *faithfulness* of an explanation, which describes "how accurately it reflects the true reasoning process of a model" (p. 4198). In the context of using explanations as a tool to examine a model's validity, the

distinction between plausibility and faithfulness is especially valuable: here, faithfulness is not a measurement of explanation quality but a prerequisite.

To fully leverage the advantages of supervised text classification in automated content analysis, profound collaboration and innovation are needed from communication and computer scholars. Through the example of hate speech detection, this contribution has posited a need for quality control and has shown that adequate methods to establish the validity of a model are rare. While a concept such as hate speech presents a rather extreme example due to its complexity, it nonetheless illustrates the intricacies of two disciplines joining one method rather well. Since research on this specific topic is currently growing in both fields, the outlook of building better-performing and explainable models may motivate closer collaboration despite the additional effort. Scholars should collaborate on theoretical and empirical work to resolve epistemological differences, align research processes, develop joint measures for quality, and collect requirements for models that show what they actually compute in a way that is seminal to automated content analysis. This could, for example, result in the development of standardized strategies and criteria for validity in automated content analysis, specific interpretable models, and faithful explanations. As much as the research community and practitioners in the realm of hate speech will benefit from this work, we shall not underestimate how it may contribute to methodological improvements in computational communication studies in general.

Laura Laugwitz is a PhD candidate at the Institute for Journalism and Communication Studies at Universität Hamburg, Germany. https://orcid.org/0000-0001-8527-2504

References

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *Annals of Applied Statistics*, *5*(4), 2403–2424. https://doi.org/10.1214/11-AOAS495

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

boyd, danah, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brosius, H.-B., Koschel, F., & Haas, A. (2009). *Methoden der empirischen Kommunikationsforschung [Methods of empirical research]* (5th ed.). VS Verlag für Sozialwissenschaften.

Chauviré, C. (2005). Peirce, Popper, abduction, and the idea of a logic of discovery. *Semiotica*, *4*(153), 209–221. https://doi.org/10.1515/semi.2005.2005.153-1-4.209

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2019, May 6-9). *L-Shapley and C-Shapley: Efficient model interpretation for structured data* [Conference presentation]. 7th International Conference on Learning Representations, New Orleans, LA, Unites States.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media*, 512–515. https://ojs.aaai.org/index.php/ICWSM/article/view/14955

Eden, A. H. (2007). Three paradigms of computer science. *Minds and Machines*, *17*(2), 135–167. https://doi.org/10.1007/s11023-007-9060-8

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Fortuna, P., Soler-Company, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 6786–6794. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.838.pdf

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *Proceedings of the 5th International Conference on Data Science and Advanced Analytics*, 80–89. https://doi.org/10.1109/DSAA.2018.00018

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *32nd AAAI Conference on Artificial Intelligence*, 3207–3214. https://doi.org/10.48550/arXiv.1709.06560

Herman, B. (2017, December 7). *The promise and peril of human evaluation for model interpretability* [Poster presentation abstract]. Conference on Neural Information Processing Systems, Long Beach, CA, United States. https://arxiv.org/abs/1711.07414

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4198–4205). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.386

Kim, B., Glassman, E., Johnson, B., & Shah, J. (2015). *iBCM: Interactive Bayesian case model empowering humans via intuitive interaction.*(Report No. MIT-CSAIL-TR-2015-010). DSpace@MIT Computer Science and Artificial Intelligence Lab. https://dspace.mit.edu/handle/1721.1/96315

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. In D. D. Lee, M. Sugiyama, U. V Luxburg, I. Guyon, & R. Garnett (Eds.), *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2280–2288). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3157096.3157352

Kim, B., Rudin, C., & Shah, J. (2015). The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Proceedings of the 27th International Conference on Neural Information Processing Systems* (vol. 2, pp. 1952–1960). MIT Press. https://dl.acm.org/doi/10.5555/2969033.2969045

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018): Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning: Vol. 80* (pp. 2668-2677). PMLR. http://proceedings.mlr.press/v80/kim18d.html

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). Sage.

Krishnan, S., & Wu, E. (2017). PALM: Machine learning explanations for iterative debugging. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* 3(1), 1–6, ACM Press. https://doi.org/10.1145/3077257.3077271

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, *92*(4), 791–811. https://doi.org/10.1177/1077699015607338

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1–8. https://doi.org/10.1038/s41467-019-08987-4

Lee, L. W., Dabirian, A., McCarthy, I. P., & Kietzmann, J. (2020). Making sense of text: Artificial intelligence-enabled content analysis. *European Journal of Marketing*, *54*(3), 615–644. https://doi.org/10.1108/EJM-02-2019-0219

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 107–117). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1011

Lertvittayakumjorn, P., & Toni, F. (2019). Human-grounded evaluations of explanation methods for text classification. In K.Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 5194–5204). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1523

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship. *Queue*, *17*(1), 1–15. https://doi.org/10.1145/3317287.3328534

Liu, F., & Avci, B. (2019): Incorporating priors with feature attribution on text classification. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6274–6283). Association for Computational Linguistics. http://doi.org/10.18653/v1/P19-1631

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). *HateXplain: A benchmark dataset for explainable hate speech detection*. ArXiv. http://arxiv.org/abs/2012.10289

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2020). Image segmentation using deep learning: A survey. https://arxiv.org/abs/2001.05566v5

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM Press. https://doi.org/10.1145/3287560.3287596

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 279–288). ACM Press. https://doi.org/10.1145/3287560.3287574

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Lulu.

Nielsen, L. B. (2002). Subtle, pervasive, harmful: racist and sexist remarks in public as hate speech. *Journal of Social Issues*, *58*(2), 265–280. https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/1540-4560.00260

Paasch-Colberg, S., Strippel, C., Laugwitz, L., Emmer, M., & Trebbe, J. (2020). Moderationsfaktoren: Ein Ansatz zur Analyse von Selektionsentscheidungen im Community Management [Moderation factors: an approach to analyzing selection decisions in community management]. In V. Gehrau, A. Waldherr, & A. Scholl (Eds.), *Integration durch Kommunikation: Jahrbuch der Publizistik- und Kommunikationswissenschaft 2019* (pp. 109–119). DGPuK. https://doi.org/10.21241/ssoar.67858

Pilny, A., McAninch, K., Slone, A., & Moore, K. (2019). Using supervised machine learning in automated content analysis: An example using relational uncertainty. *Communication Methods and Measures*, *13*(4), 287–304. https://doi.org/10.1080/19312458.2019.1650166

Pineau, J. (2020). The machine learning reproducibility checklist (Version 2.0). McGill. www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation, 55*(1), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. In B. Kim, D. M. Malioutov, & K.R. Varshney (Eds.), *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning* (pp. 91–95). arXiv. https://arxiv.org/abs/1606.05386

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 1527–1535. https://ojs.aaai.org/index.php/AAAI/article/view/11491

Risch, J., Ruff, R., & Krestel, R. (2020). Explaining offensive language detection. *Journal for Language Technology and Computational Linguistics*, *34*(1), 29-47. https://doi.org/10.21248/jlcl.34.2020.223

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 673–705. https://doi.org/10.1007/s10458-019-09408-y

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Eds.), *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication* (vol. 17, pp. 6–9). Bochumer Linguistische Arbeitsberichte. https://doi.org/10.48550/arXiv.1701.08118

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samek, W., & Müller, K. R. (2019). Towards explainable artificial intelligence. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11700 LNCS*, 5–22. https://doi.org/10.1007/978-3-030-28954-6_1

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*(2), 761–773. https://doi.org/10.1007/s11135-011-9545-7

Schmitt, J. B., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as prevention or promotion of extremism?! The potential role of YouTube. *Journal of Communication*, *68*(4), 780–808. https://doi.org/10.1093/joc/jqy029

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In D. Precup, & Y. W. The (Eds.), *Proceedings of the 34th International Conference on Machine Learning:* (vol. 70, pp. 4844–4866). PMLR. http://proceedings.mlr.press/v70/shrikumar17a.html

Stodden, V. (2010). The scientific method in practice: Reproducibility in the computational sciences. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1550193

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research*, *2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks In D. Precup, & Y. W. The (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (vol. 70, pp. 5109–5118). PMLR. http://proceedings.mlr.press/v70/sundararajan17a.html

Trilling, D., & Jonkman, J. G. F. (2018). Scaling up content analysis. *Communication Methods and Measures*, *12*(2–3), 158–174. https://doi.org/10.1080/19312458.2018.1447655

Turner, R. (2016). *A model explanation system: Latest updates and extensions*. arXiv. https://arxiv.org/abs/1606.09517

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online* (pp. 33–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/w18-5105

van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2020). VisBERT: Hidden-state visualizations for transformers. In A. El Fallah Seghrouchni, G. Sukthankar, T. Liu, & M. van Steen (Eds.), *Companion Proceedings of the Web Conference 2020* (pp. 207–211). Association for Computing Machinery. https://doi.org/10.1145/3366424.3383542

van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Vilone, G., & Longo, L. (2020). *Explainable artificial intelligence: A systematic review*. arXiv. https://arxiv.org/abs/2006.00093

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In J. Andreas, E. Choi, & A. Lazaridou (Eds.), *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-2013

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019). On the (in)fidelity and sensitivity for explanations. https://arxiv.org/abs/1901.09392v4

Digital
Communication
Research.de

Recommended citation: Leerssen, P., Heldt, A., & Kettemann, M. C. (2023). Scraping by? Europe's law and policy on social media research access. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 405–425). Digital Communication Research. https://doi.org/10.48541/dcr.v12.24

Abstract: This chapter discusses the legal aspects of researchers' access to social media data, focusing in particular on recent developments in European law. We see law as playing both an enabling and a restrictive role in facilitating platform data access. Identifying a number of shortcomings in current legislation, we argue for the creation of a sound legal framework for scholarly data research. The new Digital Services Act makes some promising first steps towards regulating programmatic data access through APIs, but many obstacles and ambiguities remain. Furthermore, a clear vision on the legal status of public interest scraping projects is still lacking. In the teeth of private ordering by global platform companies, as new gatekeepers in academic research, ensuring fair and rights-sensitive data access must be a priority for the (European) legislator.

DOI 10.48541/dcr.v12.24

Paddy Leerssen, Amélie Heldt & Matthias C. Kettemann

Scraping By?

Europe's law and policy on social media research access

1 Introduction: Research access as a regulatory problem

Over the past decade, social media research has become a point of controversy in legal and regulatory discussions. In our burgeoning platform society (Van Dijck et al., 2018), access to social media data has grown to be increasingly essential for all sorts of social science research, including the analysis of hate speech. And yet as demand grows, platforms have generally restricted their research access policies over the past decade, rather than expand them. Without clear incentives for platforms to support public interest research, they have instead tended to give precedence to user privacy and data protection concerns. Such concerns may be warranted to some extent, but also risk being exaggerated and weaponized in service of platforms' more self-interested motives in avoiding independent scrutiny of their policies (Ausloos & Veale, 2020). As tensions with platforms escalate, researchers are increasingly turning to courts and legislatures to preserve their existing data access and to demand new, legally-binding access frameworks.

This chapter discusses the legal aspects of researchers' access to social media data, focusing in particular on recent developments in European law. It follows

Cohen's (2019) observation that the law plays both a restrictive and a facilitating role for platform data access: it includes information-blocking rules that constrain data access, as well as information-forcing rules that support it. Accordingly, we start this chapter by discussing the access barriers that researchers currently face and the role of the laws in constructing them, including aspects of contract, data protection, and intellectual property. Second, we review recent legal developments with an information-forcing component, which might offer pathways towards more effective and sustainable research methods. We discuss takedown reporting requirements, GDPR data access rights, as well as recent proposals to regulate platform APIs in the Digital Services Act and related plans to draft Codes of Conduct for platform researchers.

2 How we got here: The techno-legal precarity of platform data access

As communications researchers have pointed out, the problem of platform data access exacerbated rapidly after the Cambridge Analytica scandal (e.g., Bruns, 2019; Puschmann, 2019; Freelon, 2018). In response, several platforms severely restricted researcher access through APIs, in a development described by Bruns as the "APIcalypse" and leading to what Freelon termed the "post-API age." Some platforms responded more extremely than others: for instance, Instagram shut down its research API entirely while YouTube continues to allow relatively generous access (Munger & Philips, 2020). Twitter also recently expanded its accommodations for academic researchers, including a dedicated API and access to a full archive of tweets (while at the same time, however, introducing yet more restrictions on their standard API). Still, the current situation has resulted in a drastic reduction of data access opportunities for researchers. A related concern is that differences in data access between platforms can distort research agendas, by nudging researchers towards the most open and accessible platforms. A recent literature review of research on racism and hate speech on social media supports this, showing that Twitter is "far overrepresented, especially considering its relatively small user base" (Matamoroz-Fernández & Farkas, 2021, p. 215).

Researchers have responded in various ways to this new "post-API age." Some have tried to cooperate with platforms in self-regulatory arrangements (e.g., Puschmann, 2019; King & Persily, 2019; see also Jünger in this volume), some

have introduced method innovations (Münch et al., 2021), whereas others have started to rely on platform-independent data collection methods (e.g., Freelon, 2018) and others still have adopted a "data-activist" stance with the hope of lobbying governments to regulate a privacy-compliant re-opening of APIs (Bruns, 2019). The law, including but not limited to data protection, plays an important role in each of these developments.

Perhaps the most prominent effort at self-regulation in this space is Facebook's Social Science One, a partnership with US academics launched in early 2019 aiming to provide a secure and confidential access regime for researchers, who would be vetted through an independent application process (King & Persily, 2019). Unfortunately, the project was initially hamstrung by repeated delays and complications, which, according to Facebook, were the result of legal compliance concerns related to US privacy and EU data protection laws. However, many researchers did not take these claims at face value and criticized the project as an attempt to stave off binding regulation by governments with a (ultimately inadequate) promise of voluntary access (Bruns, 2019). In December 2019, the co-chairs and European advisory body issued a damning public letter expressing their frustration with the lack of progress, concluding that "we are mostly left in the dark, lacking appropriate data to assess potential risks and benefits" and expressly inviting public authorities to step in (The Co-Chairs and European Advisory Committee of Social Science One, 2019). Funders threatened to pull out of the project. This being said, the project has since then started to produce its first dataset—a database of URL information—as well as assisted in broadening and improving research access to tools such as the CrowdTangle and Ad Library APIs. However, the dataset has been criticized, due to the extensive use of "differential privacy" anonymization method that limit its accuracy and utility (mainly for qualitative research), and so have the API tools for various reasons. Access to CrowdTangle is only possible with Facebook's permission, raising questions about gatekeeping and academic freedom. Overall, then, the record is mixed at best, with some researchers more optimistic about this self-regulatory approach than others. Cornelius Puschmann, who was involved in the Social Science One project, noted: "Facebook improved access through [Social Science One] by a lot and has been very cooperative ever since" (Heldt et al., 2020; King & Persily, 2020).

Overall, self-regulatory projects such as Social Science One projects have thus moved the debate forward, but have not fundamentally reduced the impetus, at least in Europe, for more far-reaching, legally binding reforms.

Independent data collection methods have also taken flight in the "post-API age." With the help of sock-puppet accounts, crawlers or real-world volunteers using browser plugins, for example, researchers can observe platforms directly and assemble their own datasets. However, these methods face important limitations in terms of cost, sample size and bias, operating system restrictions, and so forth. Furthermore, platforms can take legal and technical actions to restrict these projects. Unauthorized data collection can potentially run afoul of many different laws, including anti-hacking laws, intellectual property, contractual restrictions in Terms of Service, and privacy and data protection laws. Indeed, researchers have reported on the complexity of data protection in this space, though compliance is certainly possible (Bodo et al., 2018). If brought to court, favorable rulings for researchers are entirely plausible or even likely based on public interest and fundamental rights defenses (see, for instance, the US ruling in *HiQ v LinkedIn*). The problem, however, is that platforms can often enforce their anti-scraping policies through extra-legal means, simply by blocking the relevant plugins or activities through technical measures and thus foreclosing the possibility for researchers to appeal to relevant constitutional defenses and public relevant interest exceptions (e.g., in data protection law). In these ways, law and technology work together to enable what Cohen (2019) terms the "de facto propertization" of platform data.

Some data scraping activities are tolerated by platforms, in what Rieder and Hofmann (2020) term "implicit acquiescence." Others are not so lucky. One notorious case involved New York University's Ad Observatory project, a collaboration between journalists and academics seeking to collect information about political advertising via a volunteer-installed browser plugin. Mere weeks before the US election, Facebook sent them a cease-and-desist letter, threatening to block the plugin if they did not comply (Horwitz, 2020). Facebook cited its Terms of Service as well as its obligations under US privacy law, which require the platform to prevent unauthorized access to user data. Critics objected that NYU's plugin only collects personal data from their volunteers, who have consented to share the data, and not from third party users, and furthermore that academic research is justified on public interest grounds (e.g., Doctorow, 2020). The broad

permissions that users usually (have to) give to extensions mean that they might be authorizing the collection of more sensitive data, even when that is not what the researchers end up collecting.

NYU has now joined forces with the Knight First Amendment Institute to challenge Facebook's actions in court, but it will likely take many years before legal certainty is obtained. More fundamentally, existing laws do not clearly explain whether or when researchers can go further than NYU's example and collect information *without* users' prior consent; something that may be particularly important in the context of hate speech, where speakers may be unlikely to volunteer their participation. Experts including the European Data Protection Board have pointed towards the many public interest exceptions in the GDPR that could possibly support research on other grounds than consent, but these questions remain clouded in uncertainty. Certainly, platforms cannot be relied on to make this determination by themselves, if only because they may lack the necessary information about the background of data scrapers. And waiting for such conflicts to make their way through the court could take decades.

It may be easy to criticize platforms for undermining public interest research, but it must be kept in mind that independent data collection also presents very real risks. The same methods used by researchers to collect data can be abused by commercial and political actors to the detriment of user privacy. In addition to the Cambridge Analytica scandal, mentioned previously, another chilling reminder is the mass scraping of facial image data by ClearView AI, used to develop (likely unlawful) facial recognition technologies. The largest social media platforms such as Facebook, Twitter, and YouTube accused ClearView AI of violating their policies. In this light, the problem is not so much that platforms restrict independent data collection, but rather that these policies are enforced across the board without an adequate public interest exception. Vetting public interest researchers, however, is a task that platforms are ill-positioned to perform, both operationally and politically. It would be a clear threat to academic freedom if platforms were responsible for deciding which researchers were permitted to study them.

These incidents underscore the fundamental precarity of developing research methods and tools for platform services. Whether relying on self-regulatory data-sharing arrangements, independent plugins, or tools built on platform APIs, researchers operate at the pleasure of platforms who maintain at all times the technical and legal power to alter, restrict or shut down entirely their access—and

who may do so at the slightest threat of legal or political risk. According to Rieder and Hofmann (2020), this techno-legal precarity requires an institutional response, focused on creating more dependable modes of access:

> A common characteristic of the data collecting projects mentioned above is their ephemeral, experimental, and somewhat amateurish nature. While this may sound harsh, it should be obvious that holding platforms to account requires 'institution-building,' that is, the painstaking assembly of skills and competence in a form that transposes local experiments into more robust practices able to guarantee continuity and accumulation. (p. 23)

This institution-building, according to Rieder and Hofmann (2020), would need to be paired with regulatory measures aimed at enhancing the "observability" of platform, for instance by regulating platform APIs: "The main goal, here, is to develop existing approaches further and to make them more stable, transparent, and predictable" (p. 22). Such demands bring us to recent debates in European law, where governments have increasingly sought to impose information-forcing rules on platforms. These rules may help to create the conditions for more robust and dependable data access frameworks and institutions to develop, although, as will be discussed below, these are early days still.

3 Regulating research access: Recent developments in European law

This section provides an overview of legislative and regulatory initiatives that enable access to platform data for research purposes. As will be shown, current efforts are both disparate and initial. With few exceptions, it concerns drafts and proposals rather than in-force measures. We start with one of the most widespread types of transparency regulation, content moderation reporting, followed by discussions of GDPR data access rights, the API-related rules from the Digital Services Act proposal, and the European Digital Media Observatory's proposal for a Code of Conduct.

3.1 Mandatory content moderation reporting (in the NetzDG and elsewhere)

One of the most common modes of data access regulation is the so-called "Transparency Report": the periodical, public reporting of aggregate data about content moderation actions. This practice originates in self-regulation, where it has long served as a rallying point for civil society initiatives such as Ranking Digital Rights and the Manila Principles. Over the past decade, platforms have gradually begun to concede to these demands and release transparency reports, which have gradually grown in scope and detail (Keller & Leerssen, 2020). In recent years, governments in Europe and elsewhere have sought to regulate transparency reporting practices.

Transparency reporting obligations can be found in numerous laws and proposals. The majority focus on moderation related to hate speech and related topics, including Germany's Netzwerkdurchsetzungsgesetz (NetzDG), France's Loi Avia, Austria's Communications Platform Law (Fischer et al., 2020), and the EU's proposed Terrorist Content Regulation. The EU's recent Digital Services Act proposal also includes expansive transparency reporting rules, with escalating levels of disclosure applied based on the size of the platform service.

Most of these instruments have not yet passed into law and/or entered into force, with the exception of the NetzDG. In force since January 1, 2018, the NetzDG offers insights into the practical impact and utility of transparency reporting regulation. Thus far, eight different platforms have released one or more semi-annual reports under this framework: Twitter, Reddit, Facebook, TikTok, Change.org, Jodel, Google/YouTube, and Soundcloud.

Overall, the response from researchers to this data has been muted at best and dismissive at worst (Suzor et al., 2019). Researchers' critiques of NetzDG transparency reporting are several (Heldt, 2019). Most fundamental, however, is the criticism that aggregate data offered by transparency reports leaves researchers without content-level insights into particular cases. As a result, researchers are unable to independently assess platforms' content classifications, and thus to determine the quality of content moderation decisions and its impacts on various groups. For instance, the fact that Google has removed x pieces of content due to hate speech between June and September 2020 does not tell us whether this content concerned, for instance, white supremacy, radical Islam, or some other variant of hate speech; whether it targeted its victims based on gender, race, or

some other protected category; whether the removed content was classified correctly (i.e., false positives); how much non-removed content was reviewed but ultimately left up (i.e., false negatives); and so forth. All of these questions require, at a minimum, access to the actual content at issue (Keller & Leerssen, 2020) and to the practices in use when enforcing content standards against hate speech.

A related criticism is that content removal reporting cannot be assessed meaningfully without robust indicators of the overall prevalence of this content across the platform. For instance, Facebook might report a bi-annual increase in hate speech removals of 15 percent, suggesting an improved detection rate. Even assuming that Facebook's classifications are correct (which we cannot, as discussed above), the opposite could still be true if overall prevalence of hate speech posts simultaneously increased by over 15 percent. In a bid to address these concerns, Facebook has since November 2020 become the first platform to publish prevalence estimates regarding hate speech (Kantor, 2020), though robust comparisons over time are difficult to make since comparable data is lacking and the special situation of an increase in automation in content governance during the COVID-19 crisis caused changes in platform moderation practices (see also Ahmad in this collection).

Another problem is that Facebook undermined the functioning of NetzDG by making their complaint mechanism difficult to access. This has had the effect of discouraging users from submitting complaint, such that Facebook received significantly fewer complaints relative to its size. Since the NetzDG transparency obligations only cover formal notices submitted within its framework, this reporting can paint a distorted picture by omitting content moderation practices initiated under platforms' self-regulatory flagging systems. Facebook in particular was removing significantly more content based on these self-regulatory systems than under the official framework, but the same problem also applies to other platforms and their self-regulatory flagging mechanisms. German authorities have fined Facebook for its practices, and recently proposed amendments to the NetzDG would require platforms to make their NetzDG complaint mechanisms easily accessible. More fundamentally, the problem remains that most takedown reporting rules may fail to capture the totality of moderation actions undertaken by the platform.

Of course, transparency reports have some (limited) utility in tracking trends in content moderation over time. For instance, NetzDG transparency reports

give a high-level view on how much data is removed, which removal grounds are triggered most frequently, and so forth. Indeed, Facebook's transparency reports under NetzDG provided empirical support for the critique that their implementation of this law discouraged users from submitting complaints, by showing that they received substantially fewer than Twitter and Google.

As of May 2020, the German government is amending the NetzDG. The legislator has acknowledged the need for researchers to access data in order to better understand platform practices, but unfortunately this finding was not put into practice. The legislator could have added an access to data provision for research purposes, but the amended version of § 2 (2) NetzDG only stipulates an obligation to report on whether and to what extent relevant insights were granted to members of the scientific and research community. It does not specify *how* researchers will get these "relevant insights" or impose any obligation on platforms to provide them.

Another proposed amendment is to add a new section to § 2 (2), which requires platforms to disclose the use of automation for content moderation purposes, regardless of whether the content was removed because it was considered unlawful or because of a violation of the platform's own content rules. This information could be valuable to further understand how hate speech is detected by platforms, although the information provided here is likely to remain of a rather general nature.

3.2 Copyright

In general, copyright law is rather perceived as an obstacle in the overall attempt to gather third-party data—even for research purposes. But new reforms are underway to relieve some of these constraints. Researchers might infringe the platforms' rights when collecting policies and documents. Recently, legislators have recognized the need to re-adapt to the new possibilities for research and innovation via digital technologies. In 2017, Germany passed a provision for text and data mining in order to bring copyright law in line with the needs of the information society. Under § 60d (1) German Copyright Act, one may collect and duplicate automatically and systematically data in order to create a corpus for research purposes. Similarly, Article 3 of the Digital Single Market Directive makes it mandatory for Member States to provide for an exception allowing text

and data mining "for the purposes of scientific research." The provision does not, however, provide access to data in itself. Instead, the scope of application is restricted to works to which researchers have "lawful access." In Germany, for instance, scraping might infringe the platforms' exclusive rights to reproduce, distribute and publicly reproduce under Section 87b (1) German Copyright Act when third-parties repeatedly and systematically reproduce the "database." However, this restriction will, generally, not affect researchers because of the non-commercial nature of their action.

3.3 GDPR data subject rights

The GDPR does not only block data access; it can also force data access by virtue of its transparency provisions. The GDPR offers a number of data access rights regarding personal data held by the platforms, including the right of access, the right to data portability, and the right to an explanation regarding automated decision making. These rights are granted to data subjects, rather than researchers per se, but Ausloos and Veale (2020) demonstrate that they can nonetheless be repurposed as research tools by enlisting data subjects as volunteers. Their work explores some of the ethical considerations involved and outline a number of use-cases, including research into content moderation, online tracking, the use of biosensors, and digital labor issues. They do not address hate speech research in particular beyond the general issue of content moderation, and further exploration of use-cases in this space would likely be fruitful.

In theory, other user-facing rights could potentially also be retooled for research purposes. For instance, in the context of self-regulation, researchers have crowdsourced the explanations that Facebook offers their users regarding their microtargeted advertisements under their "Why Am I Seeing This" feature, in order to gain insights into targeting practices (WhoTargetsMe, 2020). Rules and proposals for user-facing information rights abound under European law, including the rules on recommender systems in Article 30 of the Digital Services Act. For the most part, however, these rules focus on easy-to-digest, broadly understandable explanations for a general audience, which may only offer marginal benefits to specialized researchers (Leerssen, 2020).

3.4 Digital Services Act: Data access for "vetted researchers"

Perhaps the most significant data access proposal for hate speech research access regulation is Article 31 of the EU's newly-proposed Digital Services Act. Titled "Data access and scrutiny," this article authorizes local platform regulators, so-called "Digital Services Coordinators" (DSC), to compel platforms above a certain size to disclose relevant data to "vetted researchers." The DSA has not yet been finalized. Our discussion focuses on the text of European Commission's original proposal of 15 December 2020.

Many of the details of this article will likely change, since this concerns a first draft proposal with a long and controversial legislative process ahead of it. As of mid-2021, however, the scope of Article 31 is relatively restrictive in terms of its subject matter as well as eligible researchers. In terms of its subject, Article 31 only applies to research conducted for purposes of risk assessments related to the platform service, including but not limited to the following: (a) the dissemination of illegal content, (b) effects on fundamental rights including privacy and freedom of expression, and the rights of the child, and (c) inauthentic usage of the service, "with an actual or foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security" (Articles 31 and 26(1)). This scope clearly enables research into hate speech, but may cut off other fields of inquiry.

For researchers to qualify as "vetted," they must be "affiliated with academic institutions, be independent from commercial interests, have proven records of expertise in the fields related to the risks investigated or related research methodologies, and shall commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request." The restriction to academic institutions risks excluding NGOs and other third parties, unless they partner with vetted academics with a view to gaining access. To comply with the requirement of data security, researchers will likely be required to produce a data management plan demonstrating, at a minimum, GDPR compliance and perhaps the observance of other ethical or scientific standards. At present, the details of these rules remain unspecified, but the European Commission is tasked with developing guidance to ensure compliance with the GDPR. An interesting question is how this standard-setting activity will interact with other

delegated rulemaking and standard-setting ongoing in this space, including the Research Code of Conduct in production at EDMO discussed below.

Article 31 also contains ambitious but as of yet unspecified rules about disclosure formats: subparagraph 3 requires that platforms "shall provide access to data [...] through online databases or application programming interfaces, as appropriate." This clause seems to respond to ongoing debates about the governance of research APIs outlined above. Yet, it leaves many questions open as to how and when APIs or databases would be "appropriate"—again, matters for further standard setting by regulators. The provision does signal, however, that the DSA proposal envisages broadly accessible forms of data-sharing and not merely singular data grants to individual research groups; in some cases, "where appropriate," authorities might demand that data is made available programmatically to a broader pool of researchers. It could arguably provide the basis for regulators to expand and improve existing self-regulatory efforts, such as Facebook's CrowdTangle and Twitter's academic research API, and enable monitoring and scrutiny by larger sets of (vetted) researchers in real-time. The current limitations on 'vetted researchers' could however pose an obstacle to creating truly inclusive resources.

Another blind spot is Article 31 (6) DSA: according to the current proposal, platforms shall have a right to request an exemption whenever they do not have access to data. Because platforms are supposed to act against illegal content under Article 14 and 15 DSA, it might not be available for later research. That is a problem raised by journalists and prosecutors investigating war crimes: once the platforms remove the content, it is almost impossible to retrieve it (or highly dependent on the platforms' goodwill). If this is not policed properly, important material for the study of hate speech and other illegal phenomena, as well as the gatekeeping function of platforms, might be destroyed. This same data may also be an important ingredient in the training of AI tools for the detection of hateful content.

Notably absent from Article 31 is a procedure for researchers to petition either platforms or regulators for access. In the current draft, it seems, access depends on the initiative of the regulator. There is a risk here that researcher access becomes subservient to the goals and aims of regulatory investigations, instead of setting its own scientific agenda. To preserve academic freedom in this regime, regulators would ideally devise independent and objective procedures to vet and prioritize researchers and their projects.

3.5 Digital Services Act: Mandatory ad archive APIs

The DSA proposal also contains specific data access rules related to online advertising in Article 30. Microtargeted online advertising has been the subject of many controversies and policy concerns, including the dissemination of hate speech through channels that are difficult for third parties to trace or respond to (e.g., Wong, 2020). Here too, platforms above a certain size are required to provide some programmatic access to relevant researchers via an API. The requirements here are significantly more detailed than the generic data access framework of Article 31 outlined above.

This rule builds on existing self- and co-regulatory practices, currently known as "ad archives" or "ad libraries," which have already been implemented in some form by most major advertising platforms and are increasingly subject to regulatory requirements in Europe and elsewhere (Leerssen et al., 2019). Ad archives may be valuable for hate speech research because they allow researchers to trace the use of hate speech (and other speech) within ad ecosystems and their interaction with non-ad content.

The DSA largely mirrors these existing practices in requiring that the following information is made available: the content of the ad, the name of the ad buyer, the advertising period, the total number of views, and demographic information about the audience reached. Existing self-regulatory practices for advertising continue to exhibit many errors and shortcomings (Leerssen et al., 2019), and these new binding rules may provide an impetus for platforms to invest in more rigorous implementations.

We also see remarkable differences compared to self-regulatory standards. The most significant change by far is that the DSA's rule applies to *all* advertisements sold on the service, whereas platform projects have been far more narrowly targeted to (varying definitions) of political campaign and issue ads. This broader approach covering all ads has been endorsed by many researchers and activists, who objected that platforms failed to reliably define and detect political ads—thus creating sampling problems and undermining the research utility of their data—and that non-political, commercial ads also deserve scrutiny. The new approach leaves it to researchers themselves to define and operationalize their own interest categories.

The metadata about advertisements required by the DSA proposal also differs on two points. First, the DSA is more expansive in that it also requires that platforms disclose their *targeting criteria* for each ad: "whether the advertisement was intended to be displayed specifically to one or more particular groups of recipients of the service and if so, the main parameters used for that purpose." Again, this change responds to widespread criticism from researchers about the lack of such data in the existing databases (Leerssen et al., 2019). Platforms have objected that disclosing targeting criteria may run afoul of user privacy, which may indeed place limits on the documentation of Facebook's custom audience targeting methods, but is not evidently compelling for other aspects of targeting. Furthermore, the requesting of "main parameters" suggest that platforms will not have to be exhaustive in their documentation. Thus, the further interpretation and implementation of this rule remains subject to debate. In January 2020, only one month after the DSA was proposed, Facebook did announce that it would be making targeting data available on a limited basis to academic researchers in the US, related to the US elections. We cannot assess at this time what the value of these disclosures will be, but the lessons learned here will certainly be instructive for the future of Article 30 DSA.

Second, the DSA also takes a large step *backwards* by omitting advertisement spending data. Spending data has been standard inclusion in all self-regulatory ad libraries (albeit in general ranges rather than precise amounts), and it remains unclear why it has been omitted here.

As noted, Article 30 DSA requires large platforms to disclose their ad archive data through public APIs, enabling programmatic access by researchers as well as other third parties. It should be noted here that Facebook's existing Ad Library API has been criticized extensively by researchers, due to inconsistency, performance issues and bugs, and a lack of user-friendliness (Mozilla, 2019; Rosenberg, 2019). This is another failure mode for ad archive regulation, which might require further regulatory standard-setting to address. An alternative approach would be to demand that platforms disclose their data to an independent third party, which would be entrusted with designing and operating an effective researcher API. For instance, the EU's Data Governance Act Proposal provides "Data Altruism Organisations" (chapter IV) that would "lead to the establishment of data repositories" and "facilitate cross-border data use" (Nr. 36 of the DGA's explanatory memorandum). Such registered third-parties would be subject to strict transparency obligations

and specific requirements under Article 19, making them trusted intermediaries for a general interest data access.

3.6 The EDMO Code of Conduct

A final development worth noting is the push to develop a Code of Conduct for researchers handling platform data, spearheaded by the Commission-funded European Digital Media Observatory (EDMO). This procedure is based on Article 40 of the General Data Protection Regulation, which allows stakeholders involved in the processing of personal data to design voluntary codes specifying GDPR compliance methods in their particular field of activity. These Codes can then be approved by Data Protection Authorities (DPAs) in order to create legal certainty about the requirements of data protection law (which can otherwise be highly general and ambiguous). EDMO's mandate is largely centered on combating disinformation, but they have already announced that their Code of Conduct initiative is not intended to be limited to this subject matter. It therefore bears relevance to other fields of social media research, including the analysis of hate speech.

EDMO's proposal, like most discussed here, is at an early stage: their Article 40 Working Group was officially announced in November 2020, with an official call for comments soliciting input from relevant stakeholders. The Working Group now has the task of processing these comments and further specifying their approach.

Since Article 40 GDPR merely serves to clarify existing law, it cannot create new obligations on platforms to share data with researchers or other third parties, beyond what they voluntary commit to when signing up for the Code of Conduct.

One role the Code could play is to clarify how data protection law should apply in new data-sharing arrangements such as the DSA access frameworks outlined above. For instance, the European Commission could draw on an academic Code of Conduct in assessing who qualifies as a "vetted researcher," and evaluating their data management plans under Article 31 DSA's data access framework. A related role that a Code of Conduct might play is clarifying when and how independent data collection efforts comply with the GDPR—a matter which continues to raise legal uncertainty for researchers and platforms alike. By creating a procedure to certify the GDPR compliance of independent data collection projects, the Code could help to operationalize public interest exceptions without forcing platforms

to act, as Mathias Vermeulen puts it, "as de facto gatekeepers who decide on the validity of specific research proposals and methods" (Vermeulen, 2020, p. 21). Such an institutionalized, vetted approach has the advantage of greater accountability for both platforms and data recipients, although an overly bureaucratic access procedure could discourage buy-in from researchers and may risks privileging certain forms of research over others.

Similarly, Article 35 DSA proposes "the drawing up of codes of conduct at Union level to contribute to the proper application of this Regulation, taking into account in particular the specific challenges of tackling different types of illegal content and systemic risks, in accordance with Union law, in particular on competition and the protection of personal data." Finally, these Codes might also be a venue for platforms, in light of the mounting public pressure, to make certain data access commitments, including proactive data-sharing with compliant researchers as well as non-interference with compliant data scraping projects. Overall, a key question remains the interaction between the GDPR and DSA codes of conduct in this space; whether EDMO will choose to focus on supporting and facilitating the DSA's (future) access rules, or rather to create an independent, GDPR-based framework of its own.

4 Outlook: First steps taken, long read ahead for responsive API regulation

Clearly, these are heady times for the regulation of research access. Quite suddenly it has become a hot topic for lawyers and policymakers—hot, if not overheated. The result has been a spate of different proposals and initiatives, some more promising than others. Many of these plans are still at an extremely early stage, and may still take years to come to fruition. But experience shows that the early stages of drafting are often pivotal, since it is then that concepts, frames, ideas can become anchored in legislative minds and texts. All the more important, therefore, for communications researchers and other social scientists to involve themselves in these discussions and demand rulemaking that actually responds to their research needs.

If European policymakers were to listen more closely to the research community, they might for instance realize that their recurrent emphasis on aggregate

takedown reporting rules, without insights into the underlying content, may be somewhat misplaced. Such rules continue to proliferate in various instruments, despite offering a rather minimal benefit to the scientific understanding of the topics they regulate, including hate speech. At the same time, policymakers still lack a clear policy vision on what many researchers find most urgent: tools to study the actual spread of harmful content, and the substance of what is ultimately being flagged and removed. Or indeed, on academic research unconstrained by governments' particular interests or agendas. A clear stance on the status of independent scraping projects has also not emerged yet, and efforts to regulate APIs are still in their infancy. National laws fail to protect researchers against overbroad Terms of Service that jeopardize good-faith research efforts, despite the significant public interests often implicated in this activity. Collecting the pictures of the January 6, 2021, attacks on the Capitol through scraping the social media app Parler, for instance, has been an invaluable source for public interest-based reporting.

While the DSA is still in the making, it is encouraging to see that it contains a clear statement in favor of mandatory procedures for researcher data access, including the regulation of automated disclosure via public databases and APIs. Also promising are the DSA's rules on Ad Archive APIs, the Commission's backing of a GDPR Code of Conduct, and new experimentation with data subject rights as a tool for researchers.

Whilst these efforts appear well-intentioned, the devil remains in the details. Regulating the design of APIs in particular is a complex and relatively unprecedented issue, raising questions as to whether governments will be up to the task. To ensure that researchers' access to user data via APIs is GDPR-compliant, compliance-by-design solutions could be explored. One possibility is pseudonymized/anonymized data outputs, which could eliminate the need for substantial vetting procedures for certain APIs. Recent developments in self-regulation, such as Facebook's attempts at differential privacy, seem to point in this direction. Approaches that allow access to more sensitive data would likely require more extensive vetting procedures, at the possible cost of scalability and uptake amongst researchers. Generally speaking, reproducibility and reliability of the data produced remains a concern.

Perhaps the most feasible approach, at least in the short term, might be to develop certification schemes or safe harbors to protect independent scraping

efforts from restrictive platform policies; this issue is not currently addressed in any relevant legislation, but the EDMO Code of Conduct and other GDPR standard setting could already be an important first step towards creating greater certainty in this space, so that ethical and privacy-conscious research, in compliance with researchers' special duties of care, is not restricted unnecessarily. There is no doubt that privacy and academic research can be reconciled, but particularly in sensitive areas such as hate speech, safeguard procedures are crucial to prevent abuse and preserve the rights of users and victims. Just like ethical tests for medical trials or trials involving humans, data use audits might have to precede large-scale API uses by scientists.

In the longer term, however, there is no way around establishing a clear and sound legal framework for scholarly data access; independent scraping is not enough, and there is a clear need—and political will—to also regulate API access and data grants. The more social interaction happens in the digital sphere, subject to the private ordering of global platform conglomerates, the more should legislators protect the lawful access to research data.

Paddy Leerssen is a PhD candidate in Information Law at the University of Amsterdam, Netherlands, and a non-resident fellow at the Stanford Center for Internet and Society, USA.

Amélie Heldt is a researcher at the Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg, and associated with the Humboldt Institute for Internet and Society, Berlin, Germany. https://orcid.org/0000-0002-1910-9925

Matthias C. Kettemann is senior researcher at the Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg, Germany. He is research group leader at the Humboldt Institute for Internet and Society, Berlin, Germany, and the Sustainable Computing Lab at the Vienna University of Economics and Business, Austria. https://orcid.org/0000-0003-1884-6218

References

Ausloos, J., & Veale, M. (2020). Researching with data rights. *Technology and Regulation,* 136-157. https://doi.org/10.26116/techreg.2020.010

Bodo, B., Helberger, N., Irion, K., Zuiderveen Borgesius, K., Moller, J., ..., & de Vreese, C. (2018). Tackling the algorithmic control crisis: The technical, legal, and ethical challenges of research into algorithmic agents. *Yale Journal of Law & Technology, 19*(1), 133–180.

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society, 22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Cohen, J. (2019). *Between truth and power: The legal constructions of informational capitalism.* Oxford University Press.

Doctorow, C. (2020, November 20). Facebook is going after its critics in the name of privacy. *Wired.* https://www.wired.com/story/facebook-is-going-after-its-critics-in-the-name-of-privacy/

Fischer, G., Ketteman, M. C., & Rachinger, F. (2020). Così fan tutte: Some comments on Austria's draft communications platforms act (Graz Law Working Paper No 05-2020). https://doi.org/10.2139/ssrn.3731593

Freelon, D. (2018). Computational research in the post-API age. *Political Communication, 35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Heldt, A. (2019). Reading between the lines and the numbers: an analysis of the first NetzDG reports. *Internet Policy Review, 8*(2). https://doi.org/10.14763/2019.2.1398

Heldt, A., Kettemann, M., & Leerssen, P. (2020, November 30). The sorrows of scraping for science: Why platforms struggle with ensuring data access for academics. *Verfassungsblog*. https://verfassungsblog.de/the-sorrows-of-scraping-for-science/

Horwitz, J. (2020, October 23). Facebook seeks shutdown of NYU research project into political ad targeting. *Wall Street Journal*. https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533

Kantor, A. (2020, November 19). Measuring our progress combating hate speech. *Facebook.* https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/

Keller, D., & Leerssen P. (2020). Facts and where to find them: Empirical research on Internet platforms and content moderation. In N. Persily & J. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 220–251). Cambridge University Press.Formularende

King, G., & Persily, N. (2019). A new model for industry-academic partnerships. *PS: Political Science and Politics, 53*(4), 703–709. https://doi.org/10.1017/S1049096519001021

King, G., & Persily, N. (2020, February 13). Unprecedented Facebook URLs dataset now available for academic research through Social Science One. *Social Science One*. https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one

Leerssen, P. (2020). The soap box as a black box: Regulating transparency in social media recommender systems. *European Journal of Law and Technology, 11*(2). https://doi.org/10.2139/ssrn.3544009

Leerssen, P., Ausloos, J., Zarouali, B., Helberger, N., & de Vreese, C. (2019). Platform ad archives: Promises and pitfalls. *Internet Policy Review, 8*(4). https://doi.org/10.14763/2019.4.1421

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Mozilla (2019, March 28). Facebook and Google: This is what an effective ad archive API looks like. *The Mozilla Blog*. https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like

Munger, K., & Phillips, J. (2020). Right-wing YouTube: A supply and demand perspective. *The International Journal of Press/Politics*. Advanced online publication. https://doi.org/10.1177/1940161220964767

Münch, F. V., Thies, B., Puschmann, C., & Bruns, A. (2021). Walking through Twitter: Sampling a language-based follow network of influential Twitter accounts. *Social Media + Society, 7*(1). https://doi.org/10.1177/2056305120984475

Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society, 22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review, 9*(4). https://doi.org/10.14763/2020.4.1535

Rosenberg, M. (2019, July 25). Ad tool Facebook built to fight disinformation doesn't work as advertised. *The New York Times.* https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html

Suzor, N. P., Myers West, S., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13, 1526–1543.

The Co-Chairs and European Advisory Committee of Social Science One (2019, December 11). Public statement from the Co-Chairs and European Advisory Committee of Social Science One. *Social Science One*. https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Vermeulen, M. (2020). The keys to the kingdom. Overcoming GDPR-concerns to unlock access to platform data for independent researchers. Draft paper. https://doi.org/10.31219/osf.io/vnswz

WhoTargetsMe (2020). Our research. https://whotargets.me/en/our-research/

Wong, J. C. (2020, January 29). One year inside Trump's monumental Facebook campaign. *The Guardian*. https://www.theguardian.com/us-news/2020/jan/28/donald-trump-facebook-ad-campaign-2020-election?CMP=Share_iOSApp_Other

Digital
Communication
Research.de

Recommended citation: Jünger, J. (2023). Scraping social media data as platform research: A data hermeneutical perspective. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 427–441). Digital Communication Research. https://doi.org/10.48541/dcr.v12.25

Abstract: Working with social media data is a hermeneutic procedure systematically guided by doubts about the meaning of data at all stages of the research process, from data collection and preparation to data analysis and publication. A short walk through the automated data collection workflow, as it is implemented in the open-source software Facepager, highlights some of the epistemic peculiarities of the process. The paper encourages researchers to deal with technical details, errors, and restrictions in order to gain a deeper understanding of the organizing principles of the web. Technical limitations and hurdles should not solely be considered as problems to be solved, but also as indicators of social processes on online platforms. Scraping social media data touches on key aspects of platformization and, therefore, is not merely a data collection method, but also a means of examining the online world through a data hermeneutical lens.

DOI 10.48541/dcr.v12.25

Jakob Jünger

Scraping Social Media Data as Platform Research

A data hermeneutical perspective

1 Social science researchers and the platform ecosystem

While participation, inclusion, and empowerment were dominant topics in the early years of internet research (e.g., Scherer, 1998), the last decade has seen a focus on hostile, uncivilized, and deceptive behaviors (e.g., Ben-David & Matamoros-Fernández, 2016). To understand prosocial and antisocial behaviors, researchers have been working with data from social media platforms, including Facebook, Twitter, and YouTube, which provide application programming interfaces (APIs) that allow large-scale analyses of textual data (such as user comments), metrics (such as like and share counts), and network data (based on followers and hashtags). These data are not merely traces left behind by users; they are co-produced by users, platforms, and researchers (Driscoll & Walker, 2014; Vis, 2013).

In general, using social media data for research is not a neutral process—it promotes or hinders the development of platforms as researchers become part of the platform ecosystem. Reactivity and interactivity are embedded in scientific data collection and analysis processes (Marres, 2017, p. 190) both on a surface and a structural level. For example, on the surface level, every click on a YouTube or Tik-

Tok video by a researcher increases the view count. On the structural level, using APIs to amass large datasets increases the attention paid to the platforms studied. In fact, some researchers have stated that platform research has "facilitated these platforms' gradual societal acceptance" (Bruns, 2019, p. 1553). Furthermore, the findings from studies analyzing disinformation campaigns or hate speech can inform public debate and policy making as well as platform organizations and can eventually change the platform ecosystem.

Within this context, a firm grasp of social media data collection processes is crucial in order to understand how online platforms, users, and scientists shape communication datasets. The decisions made in this process have consequences for the interpretation of scientific findings in at least two ways:

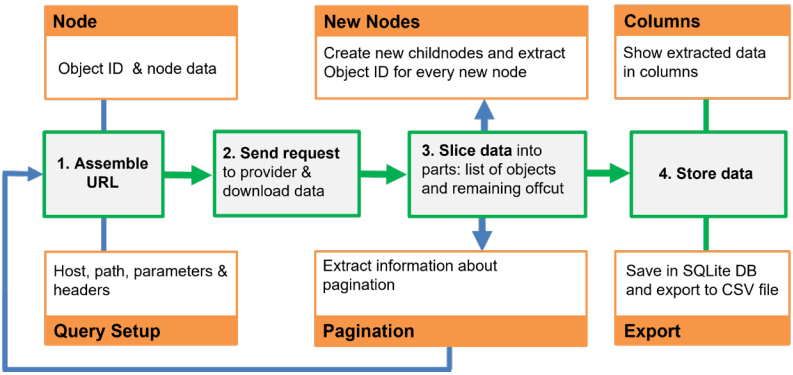
1. *Sampling:* No matter how much effort is exerted, samples of online content, to a certain degree, are always black boxes. For example, technical obstacles cause data dropouts without the exact causes always being known. In addition, populations are usually unknown – a list of all contents is not available – or cannot be defined because, for example, the boundaries of all possible communication situations are not sharply delineated. In this respect, it becomes necessary to assess what a sample actually represents.

2. *Operationalization:* The data structures that can be collected are prescribed by online platforms. Even though a multitude of data traces may be available, their meanings and creation contexts are more diverse than can be expressed, for example, by the number of likes. The data found are not necessarily the best, but only the best available indicators of theoretical constructs, such as communities or discourses.

Such uncertainties must be taken into account in the interpretation of research results. The more is known about the background conditions of the data-generating processes, the more stably the results can be interpreted. Working with social media data is a hermeneutical procedure systematically guided by doubts about the meaning of data at all stages of the research process, from data collection and preparation to data analysis and publication. Furthermore, the paper suggests a change of perspective, viewing technical limitations not solely as problems to be solved but also as indicators of social and organizational processes on online platforms.

In order to higlight some of the hermeneutical challenges, the following sections describe the automatic data collection workflow as we implemented it in Facepager (Jünger & Keyling, 2019). Facepager is a tool that can be used by non-programmers for automated data collection. By design, it is not a one-click-and-you-get-it-all solution; instead, it encourages researchers to deal with low-level API details, errors, and restrictions in order to gain a deeper understanding of the organizational and technical conditions of online platforms. The basic workflow consists of four steps: (1) assembling uniform resource locators (URLs), (2) downloading resources, (3) slicing and extracting data, and (4) storing and exporting data. The sketch of the workflow shown in Figure 1, and outlined in the following chapters, provides the background for delving into the epistemic dimension of social media data.

Figure 1: The Facepager process model



2 Step 1: Assembling URLs – indications about users and content

Whether they are implemented as classical webpages or originate from APIs, resources on the web are usually identified by URLs. When browsing the web, URLs are visible in the address bar. For example, the address of an Instagram page consists of the base path, "https://www.instagram.com," followed by a path containing a handle, such as "smartdatasprint." The dual function of URLs has been described within the context of semantic web applications (Sauermann et al., 2008). First, they are so-called endpoints for requesting documents or webpages

containing information about users or posts. Second, they identify the described entities, such as the users, organizations, or posts.

Due to this dual function, scraping social media data always involves dealing with representations of entities instead of the entities themselves. Requested documents are representations of the platform's database content, which represents social entities. The situation is further complicated when an organization or a human is active under different accounts. Therefore, data accessed on the web provide indicators about behavior without a clear concept of what those data represent. As in Plato's allegory of the cave, we see the shadows of actors and must build hypotheses about their meanings based on the combination of the actors' moves and the platforms' infrastructure. We can only deal with the artifacts of data-generating processes leading to representations of something unknown.

Requesting the URL mentioned above will lead to a hypertext markup language (HTML) page that is rendered in the browser and shows information about a user profile. Different representations of the same data are usually attached to different URLs (Figure 2). For example, when adding the parameter "__a=1" to the Instagram URL a document containing JavaScript object notation (JSON) data instead of HTML data is delivered. These formats differ in important ways. HTML contains markup that is used to assemble the visual (or auditive) representation of a page; thus, the document contains the data that users see (or hear) on the user interface. JSON is a human- and machine-readable format containing data structured according to key-value fields. JSON is usually provided by API endpoints to enable the development of third-party apps in order to enhance platform functionality (Jünger, 2018).

While the structure of HTML pages changes frequently and must be explored by researchers to extract relevant data, API endpoints are documented on providers' pages and are relatively stable over time. The difference between the two access types has consequences for social media research because the documents (as well as the providers' databases) may contain different data points. Moreover, significantly different relations between the data points and different data contexts may become salient and eventually guide the process of knowledge production. As an example, conversation structures (e.g., threads containing replies to comments) are visible on the user interface. In contrast, reconstructing nested threads from API data gathered from platforms such as Facebook or VKontakte is partially impossible, although responses to hate

Figure 2: Three representations of the same Instagram page

Source: https://www.instagram.com/smartdatasprint/?__a=1

speech, for example, are important for analyzing toxic discourse dynamics, and conversations between users are essential for tracing community formation.

Reverse engineering the URLs of HTML documents or reading API endpoint documentations is not merely informative from a technical point of view. The organizational principles of the platforms become visible, such as when usage scenarios for data processing are described in API references. In such scenarios, numerous references to marketing purposes and the data-centric business models of the providers appear. For example, Instagram provides two use cases for its API:

> The API is intended for *Instagram Businesses and Creators* who need insight into, and full control over, all of their social media interactions. If you are building an *app for consumers* or you only need to get an app user's basic profile information, photos, and videos, consider the Instagram Basic Display API instead. (Instagram, 2021, emphasis added)

In contrast, academic research does not seem to be a relevant use case from the providers' perspective. In recent years, APIs have become gradually more restrictive (Jünger, 2021), with some scholars even talking about the "Post-API Age" (Freelon, 2018) or the "APIcalypse" (Bruns, 2019). Although Facebook has launched research partner programs, initiatives investigating disinformation and related issues, such as the Ad Observer (Edelson & McCoy, 2021) and the Instagram monitoring project

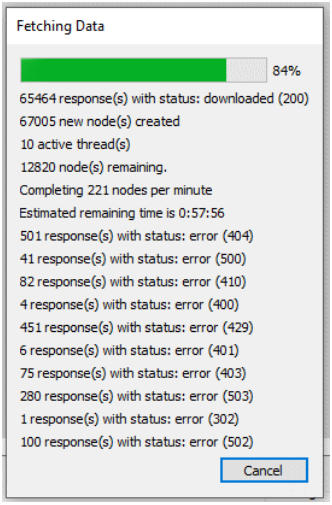
of AlgorithmWatch (Kayser-Bril, 2021), reported they were shut down by Facebook. In consequence, as long as online platforms do not accept their ethical obligation to open up research that serves the public interest, researchers are forced to put even more effort into understanding the various pathways to online platforms' data.

3 Step 2: Downloading data and platform mediation

In a broad sense, the entire web can be considered an API (Fielding, 2000) since downloaded resources are processed by other software, such as a browser or dedicated research tools and scripts. Viewing an interface from the perspective of media theory (Marres & Gerlitz, 2016) highlights the fact that APIs are not solely technical infrastructures but also involve social processes, especially in terms of the processes of the provider organization and the rules governing usage. The interface is under the control of the API provider and is usually not specifically designed to serve scientific purposes. Thus, like the behavior of users, the analysis of content on social media platforms is mediated through these platforms.

This mediation has limitations, especially when rate limits slow down the collection process or when certain content is not available. Each platform has its own set of rules. In general, access is smooth if researchers behave like humans and download data slowly. However, efficiency is restricted with this method, and the research process gains little from automation. For example, Twitter restricts the number of requests for a list of followers to 15 per 15 minutes. Each batch of data contains up to 5,000 IDs; in the next step, profile information can be requested for up to 100 IDs at a rate of 900 calls per 15 minutes (Twitter, 2021a). Thus, crawling the followers of followers for network analyses can become a tedious task and should be carefully planned. Moreover, when crawling the web, a variety of status codes, such as redirects (302), rate limits (429), and server errors (500), are encountered (see Figure 3 for examples). The larger the dataset, the higher the chance of encountering errors. The status codes tell a story about how the web works and highlight the dynamic nature of changes on the web. What works today may already be obsolete tomorrow, and pages that were not working a moment ago may begin delivering data again a few seconds later. In other words, the dataset is a time-traveling slice of information with barely known parameters.

Figure 3: Typical web scraping status codes on German news pages (error rate = 2%; screenshot of Facepager)



Access restrictions apply to the content as well. For example, on Facebook, access to posts in groups and on pages has to be reviewed by the platform; and even then, the names of comment authors are not available through the basic API. Furthermore, the API only provides a sample of posts per page and "will return approximately 600 ranked, published posts per year" (Facebook, 2021). The sampling criteria are opaque, and posts with more likes and shares are presumably preferred (Ho, 2020). Researching highly active accounts, such as those of news media outlets or politicians, is thus potentially biased toward popular content. Moreover, even though deleted posts and moderation practices are crucial for the analysis of antisocial behaviors, these details are usually hidden from the interface. Careful reflection in terms of the platform architecture is indispensable when assessing the scope of research findings.

Although access restrictions can be study limitations, insights into the platforms can be gained when scraping social media data. For example, API results from Telegram (2021) include flags for restricted users, and placeholders for deleted content can be retrieved from Disqus (2021), a comment plugin occasionally embedded in news websites. Therefore, dealing with errors at a low level of data collection can offer fruitful insights into the platformization of human behavior.

4 Step 3: Slicing and extracting data: A data hermeneutical perspective

After accessing and downloading resources, data wrangling begins. Web scraping involves extracting snippets of interest from many data fields and organizing them appropriately for analysis. In the case of HTML, data, such as the dates of posts, are often deeply nested in the hierarchical structure generated by the content management system of the platform. Boilerplate removal involves cutting away unnecessary content and omitting elements, such as webpage footers, headers, and metadata, to reduce the data to units of analysis, such as posts or comments. Alternatively, the elements of interest can be cut out from the data and transferred into a database file. Thus, data wrangling is a multistep process of slicing and extracting data.

The techniques used for data scraping "follow the medium" (Rogers, 2009) when selector languages are used to address the elements in the source code, because these languages also play a role in building webpages. One of the core technologies is cascading style sheets (CSS) selectors, which, on the production site, are used to specify the appearance of specific elements, such as the size, colour, and font of a comment text box. The same selectors can be used with R or Python packages or tools, such as Facepager, to grab content. In general, selectors define a path in the hierarchy of the HTML or JSON document to obtain data, such as the date field inside a comment element that is nested on the page. Sometimes different techniques and intermediate data conversion steps need to be combined. For example, collecting Twitter replies by scraping the interface is not straightforward. Progressing from an undocumented API endpoint to the date of a Twitter reply can be accomplished with Facepager by using a chain of modifiers, including transformation from JSON into HTML, and then parsing the timestamp into a formatted date object ("items_html|css:div.js-stream-tweet|xpath://div[@class='stream-item-header']//@data-time|timestamp").

In general, the hierarchical and technical structures of social media data pose a challenge since scientific data analysts are more used to working with tabular data. Shaping the data is the first step of the analysis, and it defines the units of analysis (cases) and their properties (variables). Even though data formats can be transformed into each other, the shape of the data may frame how researchers think about the world and what research questions are raised. Different data analysis frameworks require different data preparation steps. A multilevel re-

gression problem implies the assembly of different levels of standardized data in the same dataset, a network analysis problem requires relational data, and a hermeneutical problem involves a rich textual representation of the same data. These perspectives come from different research frameworks with different epistemological foundations, such as interpretive and normative paradigms (Wilson, 1973). Going through the steps of shaping the data makes it clear that transforming the world under investigation into a research problem is not merely a measuring procedure but also a knowledge generation process. Considering data wrangling as reconstructive data hermeneutics can bring fruitful irritations with regard to scientific thinking and strengthen the link between one's own analysis and the analyzed artifacts.

Along with investigations into source code and data structures, insights into website architectures can be gained from the data wrangling process. Against the backdrop of static content in the early days of the web (O'Reilly, 2005), interesting issues arise, such as how interactivity and real-time responses are built with dynamic programming languages. Furthermore, the division of labor between diverse roles, like database engineers, frontend designers, and marketing officers, is inscribed into the source code. By following links to content delivery networks and metadata containing semantic web markup in the header, one can see how a page is embedded in a web of services. These metadata often follow Twitter or Facebook standards and are used, for example, to create previews of shared links. In this way, a simple webpage documents the infrastructure of the online ecology from the infrastructural roots to the data leaves of the platformization tree (van Dijck, 2020).

Amid all these issues, the interplay of creativity versus standardization stands out as a dominant theme, and it can be illustrated in the case of emojis. Emojis may become a nuisance when scraping data because their encoding goes beyond the range of standard codepoints used for representing alphanumeric signs. Starting as a small proprietary list of pictures on Japanese mobile devices in around 2000, big tech companies (e.g., Google and Apple) pushed for emojis to finally be included in the Unicode Standard in 2010 (Bergerhausen et al., 2011; Pardes, 2018). However, despite the standardization of code points, emojis are challenging in at least four ways. First, when transferring data between software or devices, care must be taken to choose the right encodings; otherwise, the output will contain cryptic letters or empty boxes. Some functions, for example, in R under Windows, still have limited Unicode support. Second, emojis and colored and animated variations

are developed over time, and new emojis, such as the transgender pride flag, are constantly proposed (Unicode, 2021), mirroring societal developments. After new emojis are included in the standard, font designers, device manufacturers, and software developers lag behind and must decide whether, when, and how they will update their products. Third, the concrete representation of the emojis is left to the vendors, and there are diverse stylings across platforms. Fourth, even though the Unicode standard includes textual descriptions of the emojis, the interpretation is obviously open to users. For example, the "Folded Hands" emoji is known under the names "Thank You," "Please," and "Prayer" (Emojipedia, 2021), all of which bear quite different meanings. Overall, the emoji-related technical issues encountered during web scraping evoke a broad range of semiotical and social issues in the tension between standardization and innovation.

Taken together, the various challenges in data processing encourage a shift in perspective. The first reaction to technical problems might be an urge to fix the problem at hand. If one sits back for a moment, one can see through the code and the data into the social and organizational world of online platforms. From a data hermeneutic perspective, technical hurdles, because they are traces of social processes, become a subject of social science research.

5 Step 4: Storing and exporting data: Addressing replicability and platform rules

Data storage decisions have long-term consequences. The first decision to be made is whether to archive downloaded JSON or HTML data or extracted tabular data. Saving downloaded data can lead to large repositories, especially if media files have been collected. However, refinements and secondary analyses are possible if it becomes clear only later which data fields need to be analyzed. Facepager stores downloaded JSON data in an SQLite database. SQLite is an open-source database management system, and the files can easily be accessed with R or Python packages. Downloaded HTML data can be saved as files. The difference between the data formats for storage and analysis further demonstrates that data are always representations and lack a unique reference. In this sense, there is no such thing as raw data (Gitelman, 2013).

Since APIs and websites are constantly changing, corresponding documentation for downloaded and processed data needs to be prepared. Just a few months later, the structure and meanings that were obvious during the collection stage are often no longer apparent. A simplified compilation of the extracted data has the advantage that common data formats, such as CSV files, can be used; furthermore, documentation complexity is reduced. The reduction and documentation steps are fruitful for reducing errors and understanding the data. For example, in this step, it becomes apparent that Twitter IDs are very large and cannot be handled as numbers by Excel. Without being sensitive to such details, confusing paradoxes can sneak into the analysis. Automated data collection should, therefore, not be rashly outsourced to service providers. Even though this first decision about storage formats and documentation involves some effort, in the context of scientific analysis, it is important for the reproducibility and comprehensibility of the subsequent findings.

Another decision related to data formats concerns what is stored and for how long. Social media data often originate from users and demand thoughtful handling to balance legal regulations, platform terms, and ethical principles with the scientific research mandate. Data collection triggers complex considerations about the interplay between the involved actors and the processes of knowledge production in the context of social systems. Carefully reading platforms' terms of service, ethical guidelines, and copyright and data protection regulations can be inspiring, as more questions are raised than answers are given. For example, what can and should be done about deleted content is not obvious. On this point, the Twitter developer terms include the following regulation:

> If *Twitter* Content is deleted, gains protected status, or is otherwise suspended, withheld, modified, or removed from the Twitter Applications (including removal of location information), *you* will make all reasonable efforts to delete or modify such Twitter Content (as applicable) as soon as possible, and in any case within 24 hours after a written request to do so by Twitter or by a Twitter *user* with regard to their Twitter Content, unless prohibited by applicable *law or regulation* and with the express written permission of Twitter. (Twitter, 2021b, emphasis added).

Once collected, data are arranged into academic datasets. The removal of cases, as demanded by the Twitter terms, potentially obstructs reproducibility and destroys findings. Especially in research fields dealing with antisocial behaviors,

censorship, and platform regulation, it is expected that content will constantly appear and disappear—the (dis)appearance itself is part of the research interest. If the research outcomes are not merely filed away, they will eventually change the world under investigation, for example, by fueling political debates. When contrasted with ongoing discourses about user privacy, the replicability of research, and political regulation, the quoted Twitter terms illustrate how the four mentioned actors—platforms, users, legal regulators, and "you"—struggle with their roles in the platform economy. Who can legitimately make what claims and who bears what responsibility when handling social media data is subject to permanent negotiation.

6 Conclusion

Careful reflection on the interplay between users, platforms, and researchers is essential to making sound sampling decisions based on online traces and to finding interpretable operationalizations of theoretical concepts. A short walk through the automated data collection workflow offers a vague idea of the epistemic puzzles and peculiarities to be explored. At first glance, assembling URLs appears to be nothing more than a technical process. However, if one takes a closer look, questions arise as to what these addresses actually locate and the kinds of realities that different data formats represent. Although download errors and access restrictions can be perceived as annoyances, they also invite researchers to reflect on the social and organizational conditions of the web. Meanwhile, data wrangling—reconciling data structures with academic thinking—makes the tension between creativity and standardization visible. Finally, deciding on storage options is accompanied by considerations of replicability and the rules of data ownership. Thus, scraping social media data touches key aspects of platformization and, therefore, is not merely a method of collecting data but also a means of studying the online world through a data hermeneutical lens.

Jakob Jünger is Junior Professor for Digital Media and Computational Methods at the Institute of Communiction at the University of Münster, Germany. https://orcid.org/0000-0003-1860-6695

References

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Bergerhausen, J., Poarangan, S., & Anderson, D. (2011). *Decodeunicode – die Schriftzeichen der Welt* [Decodeunicode – characters of the world]. Schmidt.

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Disqus. (2021). *API documentation*. https://disqus.com/api/docs/forums/listPosts/

Driscoll, K., & Walker, S. (2014). Big data, big questions. Working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication*, *8*, 1745–1764. https://ijoc.org/index.php/ijoc/article/view/2171

Edelson, L., & McCoy, D. (2021, August 10). We research misinformation on Facebook. It just disabled our accounts. *New York Times*. https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html

Emojipedia. (2021). *Folded hands*. https://emojipedia.org/folded-hands/

Facebook. (2021). *Page feed*. https://developers.facebook.com/docs/graph-api/reference/v10.0/page/feed

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. University of California.

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Gitelman, L. (Ed.). (2013). *Infrastructures series. "Raw data" is an oxymoron*. MIT Press.

Ho, J. C.-T. (2020). How biased is the sample? Reverse engineering the ranking algorithm of Facebook's graph application programming interface. *Big Data & Society*, *7*(1), 1–15. https://doi.org/10.1177/2053951720905874

Instagram. (2021). *Instagram graph API*. https://developers.facebook.com/docs/instagram-api/

Jünger, J. (2018). Mapping the field of automated data collection on the web. Data types, collection approaches and their research logic. In M. Welker, C. Stützer, & M. Egger (Eds.), *Computational social science in the age of big data: Concepts, methodologies, tools, and applications* (pp. 104–130). Halem.

Jünger, J. (2021). A brief history of APIs. How social media providers shape the opportunities and limitations of online research. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), *Handbook of computational social science* (pp. 17–32). Routledge. https://doi.org/10.4324/9781003025245-3

Jünger, J., & Keyling, T. (2019). *Facepager* (Version 4.0.4) [Computer software]. https://github.com/strohne/Facepager/

Kayser-Bril, N. (2021). *AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook*. https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/

Marres, N. (2017). *Digital sociology: The reinvention of social research*. Polity.

Marres, N., & Gerlitz, C. (2016). Interface methods: Renegotiating relations between digital social research, STS and sociology. *The Sociological Review*, *64*(1), 21–46. https://doi.org/10.1111/1467-954X.12314

O'Reilly, T. (2005). What is web 2.0. design patterns and business models for the next generation of software. *O'Reilly*. https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Pardes, A. (2018, January 2). The WIRED guide to emoji. *Wired*. https://www.wired.com/story/guide-emoji

Rogers, R. (2009). *The end of the virtual: Digital methods*. Amsterdam University Press.

Sauermann, L., Cyganiak, R., Ayers, D., & Völkel, M. (2008). Cool URIs for the semantic web: W3C interest group note 03 December 2008. *W3C*. https://www.w3.org/TR/cooluris/

Scherer, H. (1998). Partizipation für alle? Die Veränderung des Politikprozesses durch das Internet [Participation for all? Changes in political processes through the Internet]. In P. Rössler (Ed.), *Online-Kommunikation: Beiträge zu Nutzung und Wirkung* [Online communication. Contributions about its use and effects] (pp. 171–188). Westdeutscher Verlag.

Telegram. (2021). *API > TL-schema > restrictionReason*. https://core.telegram.org/constructor/restrictionReason

Twitter. (2021a). *Rate limits: Standard v1.1*. https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits

Twitter. (2021b). *Developer terms. Developer agreement and policy*. https://developer.twitter.com/en/developer-terms/agreement-and-policy

Unicode. (2021). *Submitting emoji proposals*. https://unicode.org/emoji/proposals.html

van Dijck, J. (2020). Seeing the forest for the trees: Visualizing platformization and its governance. *New Media & Society, 23*(9), 2801–2819. https://doi.org/10.1177/1461444820940293

Vis, F. (2013). A critical reflection on big data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10). https://doi.org/10.5210/fm.v18i10.4878

Wilson, T. P. (1973). Theorien der Interaktion und Modelle soziologischer Erklärung [Theory of interaction and models of sociological explanation]. In Arbeitsgruppe Bielefelder Soziologen (Ed.), *Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit, Band 1, Symbolischer Interaktionismus und Ethnomethodologie* [Everyday knowledge, interaction, and social reality, volume 1, symbolic interactionism and ethnomethodology] (pp. 54–79). Rowohlt.

Digital
Communication
Research.de

Recommended citation: Fortuna, P., Soler-Company, J., & Wanner, L. (2023). Dataset annotation in abusive language detection. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 443–464). Digital Communication Research. https://doi.org/10.48541/dcr.v12.26

Abstract: The last decade saw the rise of research in the area of hate speech and abusive language detection. A lot of research has been conducted, with further datasets being introduced and new models put forward. However, contrastive studies of the annotation of different datasets also revealed that some problematic issues remain. Theoretically ambiguous and misleading definitions between different studies make it more difficult to evaluate model reproducibility and generalizability and require additional steps for dataset standardization. To overcome these challenges, the field needs a common understanding of concepts and problems such that standard datasets and different compatible approaches can be developed, avoiding inefficient and redundant research. This article attempts to identify persistent challenges and develop guidelines to help future annotation tasks. Some of the challenges and guidelines identified and discussed in the article relate to concept subjectivity, focus on overt hate speech, dataset integrity and lack of ethical considerations.

DOI 10.48541/dcr.v12.26

Paula Fortuna, Juan Soler-Company & Leo Wanner

Dataset Annotation in Abusive Language Detection

1 Why is the sharing of concepts and datasets important?

The last decade saw a rise in research in the area of hate speech and abusive language detection. The early period of this research was thereby characterized by a low number of publicly available datasets. A corresponding survey (Fortuna & Nunes, 2018), in which works on the topic until mid-2017 are reviewed, points out that the majority of the studies describe the collection and annotation of new datasets, but that only a few of those datasets were made available to the community. This is certainly a problem since progress in a field depends to a large extent on a critical comparison between different approaches and thus requires the sharing of resources. In the years following this survey, a lot of research has been conducted, with more datasets being introduced and new models put forward, and, as shown by a more recent survey (Vidgen & Derczynski, 2021), fortunately, the tendency to keep datasets locked away has changed: By 2020, 63 datasets were publicly available, making research more comparable and the types of data available for the detection of online hate speech, abuse, and harm more diverse. However, contrastive studies on the annotation of different datasets also revealed that other issues remain (Fortuna et al., 2020). In particular, the excessive amount of idiosyncratic interpretations of common terminology in the context

of abusive language analysis has been identified as problematic since multiple, often ambiguous, definitions and different interpretations of the same terms led to fragmented research and difficulties in data reuse. In more generic terms, ambiguous definitions make it more difficult to evaluate model reproducibility and generalizability and require additional steps for dataset standardization (Fortuna et al., 2021; Kumar et al., 2018; Vidgen et al., 2019). To overcome these challenges, the field needs a common understanding of concepts and problems, such that standard datasets and different compatible approaches can be developed to avoid inefficient and redundant research.

Sharing concept definitions and datasets is an essential component of mature research areas, and there are several reasons for this. *First*, without a common conceptual framework it is not possible to establish a dialog between the different research contributions. *Second*, annotating new data and sharing them promotes the study of new phenomena or different aspects of the same phenomena. Usually, gathering new data implies new annotation schemes and guidelines. These should be carefully constructed and documented when annotating broad concepts, such as hate speech. *Third*, and as already mentioned, shared datasets are essential for comparing results between different experiments and models. Hence, it is common that benchmarking datasets are established, which serve as a baseline for comparison and model evaluation. However, in contrast to other research areas, so far, no datasets have been established as standard in the field of abusive language detection. *Fourth*, data quality is of primary relevance and should not be taken for granted, so, along with shared data, it is necessary to provide evidence of its quality evaluation. Thus, it has been observed that several hate speech datasets pose issues regarding bias (Sap et al., 2019). *Finally*, from a pragmatic point of view, resource sharing avoids repeated work, namely in the form of concept definition and data annotation.

In view of the fact that for further advances in the field of abusive language detection, we urgently need to establish a common understanding of the basic concepts we are working with and share datasets that are based on these concepts, this article attempts to identify persistent intra- and inter-dataset challenges and develop guidelines to help future annotation tasks.

2 Challenges related to the annotation of abusive language

In what follows, the central challenges related to the annotation of abusive language datasets are analyzed. In this context, it is useful to distinguish between *intra*-dataset and *inter*-dataset quality challenges. The first concerns topics related to the annotation criteria and annotation procedure within one dataset, while the second concerns topics related to the coherence and compatibility of the annotation across different datasets. Note that this distinction also implies that a dataset that covers all the intra-dataset quality requirements may still be problematic when used and analyzed in comparison with other datasets. Table 1 lists these challenges and the guidelines aligned with them. The intra-dataset challenges concern concept definition, bias, data sharing, and ethics. As concept definition challenges, concept subjectivity, unclear definitions, and the varying generalization potential of coarse-grained categories can be identified. Bias related to dataset composition may originate in the focus on overt hate speech, in a reduced number of authors of the posts in the dataset, and in a lack of information on the annotation procedure. Data sharing is often an important challenge because of, for example, author privacy or copyright issues (the authored statements may not be shared because this would violate the rights of the author or infringe the copyright terms) or dataset integrity (the data may have undergone unwanted alterations, may not be accessible via the provided link, or may simply have been removed from the repository). The last of the intra-dataset challenges concerns ethics in the sense that the composition of a dataset may be used for harmful actions. The inter-dataset challenges concern the introduction of redundant and contradicting features across datasets.

In what follows, we address both the intra- and inter-dataset challenges (with a particular emphasis on concept definition, bias, data sharing, and ethics) in relation to abusive language.

2.1 Challenges in intra-dataset quality

In this section, the challenges related to intra-dataset quality in hate speech, as identified in Table 1, are discussed.

Table 1: Abusive language annotation challenges and the guidelines aligned with them

Level	Topic	Challenges	Guidelines
Intra-dataset	Concept definition	Concept subjectivity	Adopt a problem-driven approach
		Unclear definitions	Motivate definitions, tasks, and datasets socially
		Varying generalization potential of coarse-grained categories	Use clear category definitions
			Use coarse-grained categories with caution
			Prioritize fine-grained categories
			Match collected and targeted data
	Bias	Focus on overt hate speech	Increase versatile message search
		Reduced number of authors	Control communities, message threads, and author distribution
		Lack of information on annotation	Control covered time spans
			Provide information on the annotator profiles
			Define precise guidelines for annotation
	Data Sharing	Author privacy and copyrights	Protect the identity of the authors of the data and comply with copyright legislation
		Dataset integrity	Ensure data preservation and availability
			Include a data statement
	Ethics	Lack of ethical consideration	Follow ethical principles

Inter-dataset	Redundant data features	Diversify data characteristics
	Redundant and contradicting labels	Avoid redundant labels
		Provide definitions, examples, and justifications
		Position new concepts on the map of standardized categories

Challenges related to concept definition

Concept subjectivity. Defining the meaning of online hate speech, abuse, or harm is not a trivial task. The difficulty in providing definitions for these concepts arises from their subjectivity and the dependence of the connotation in the context in which a statement is made (see Litvinenko in this volume). For instance, according to some authors, the "N-word" is a slur no matter the context, while its intra-group usage may be considered harmless (Weir-Reeves, 2010); "*You son of a b****" is offensive in a neutral context but can be meant as an expression of admiration between friends; and the mention of the cultural background of an individual can be interpreted as racist or hate speech in some contexts. Furthermore, we cannot ignore that the public interpretation of the concepts of hate speech (or abusive language, in general) is also predetermined by legal regulations, such as the European Union Code of Conduct on Countering Illegal Hate Speech (European Commission, 2017), or the terms of use of social media platforms. That is, the determination of the interpretation and scope of the concepts of abusive language, offensive language, or hate speech underlying the research on their detection requires a thorough assessment of the different points of view and different contexts in which these may occur.

Unclear definitions. The worst lack of clarity with respect to central concepts or categories in the field is when data and annotation schemata do not provide any definitions at all, leaving what a certain data category actually represents open to interpretation. However, even when present, definition characteristics may not comply with the best standards. Low-quality definitions are vague, suffer from

being too generic, are defined in terms of negation of other categories (e.g., when "covert aggression"[1] is simply defined as the negation of "overt aggression"), or make an assumption with respect to the sensibility of the audience. This is, for example, pointed out by Vidgen et al. (2019) with respect to the concept of "offensiveness" by Davidson et al. (2017), which implies the question, "Offensive for whom?", because what is considered offensive by one audience, or in one context, might not be offensive elsewhere.

Varying generalization potential of coarse-grained categories. Categories are coarse-grained if they contain other subcategories (e.g., "hate speech" is a coarse-grained category when compared to "sexism" as a subcategory). In previous works, coarse-grained categories such as "hate speech" proved to be difficult to be generalized across datasets, while others like "toxicity" generalized well (Fortuna et al., 2021).

Challenges concerning bias

Online abuse is a rather sparse phenomenon if we consider the total volume of data on social media. This makes data collection a laborious task. Different strategies are applied to overcome this problem. However, these strategies may imply biases that are discussed in the next paragraphs.

Abusive message collection using keywords. The sparsity of online abuse data has led researchers to develop specific sampling techniques to increase their chances of retrieving abusive messages. The most common technique is to apply specific keywords for abusive message searches (e.g., derived from the Hatebase resource[2]). However, the use of specific keywords (and thus training on explicit abusive language posts) leads to a poor identification of posts with covert abusive language messages (Fortuna et al., 2021). More generally, the use of keywords, the focus on messages in communities or threads with a likely high percentage of abusive content, or sampling over relatively short time periods (Poletto et al., 2020) will necessarily generate datasets that have very specific characteristics, such that the modules trained on them are likely to perform less well on datasets with other characteristics.

1 Please note that we use single quotes for the names of abusive language categories.

2 https://hatebase.org/

Reduced number of message authors. Datasets related to hate speech or abusive language are often collected from a limited set of authors (Arango et al., 2019). If this fact is not taken into account during the partitioning of the information into training, development, and test data, messages from the same author can be randomly divided and may appear in both the training and the test data, which distorts the evaluation of the quality of the classification task. In other words, in this case, it is not possible to distinguish whether a model is capable of classifying hate speech or the content of particular authors.

Biased annotation. Guidelines for annotation are central when creating a new dataset as they will condition the data classification. Apart from the fact that when a hate speech dataset is published, the corresponding guidelines, if provided at all, are not always sufficiently clear and rigorous, the annotation will reflect the socio-cultural backgrounds of the annotators (see Kim in this volume). In the case that this is inevitable or tolerated, the socio-cultural bias characteristics should be well documented.

Challenges for data sharing

Once the data have been collected and annotated, nowadays, it is common practice to share the obtained dataset. However, dataset sharing may also put at risk the viability of the dataset.

Violation of authors' privacy and copyright legislation. In the majority of cases when social media text data are collected, no permission can be granted by the authors of the messages. This constitutes a potential violation of the authors' privacy and copyright issues. In order to comply with the legal regulations on data protection and privacy and not to invalidate the dataset as a whole, it is of utmost importance to strictly observe the data sharing and data use policies of the corresponding social media platforms. It is also essential to comply with the legislation related to copyrights for digital content. It is important to note that there are differences between the European and US legislation in this respect.

Loss of dataset integrity. A common practice is to provide only annotations and original IDs of the messages on the platforms where they have been spotted, with no direct access to the posted content of the dataset. However, in this case, there is a substantial risk that the content will be removed from the platform sooner or later, and thus not be accessible anymore, which is often the case when dealing

with abusive and harmful content. When this happens, the proportion of positive and negative classes changes, a new version of the dataset has to be created, and the advantages and purpose of sharing datasets get lost.

Ethical concerns

Another challenge that the area of hate speech automatic detection faces is that researchers do not always address *ethical concerns* related to their resources.

Lack of ethical considerations and data statements. Technological solutions need to adhere to ethical principles in order to ensure that harmful side effects are avoided. The main issues related to ethical principles are user privacy, bias, and dual use. User privacy and bias have already been discussed above. Let us thus focus on dual use. Here, dual use refers to the repurposing of abusive language detection technologies such that they cause harm. In the case of automatic hate speech and abusive language detection, the deployment of such technologies has already resulted in mistakenly flagging non-hateful discourses (Sarkar & KhudaBukhsh, 2021) and, even worse, the marginalization of some minority groups (see, e.g., Oliva et al., 2021).

To avoid pitfalls, the authors are morally obliged to anticipate how a new annotated dataset could be repurposed in a negative way and to design their data model in a way that does not cause harm. In the case of abusive language detection, research has not always been accompanied by proper ethical reflections, considerations, and terms of use. We discuss some possible solutions in the corresponding guidelines section.

2.2 Challenges of coherent annotation across different datasets

Abusive language occurs in different forms, thus potentially in different styles, and in different languages. Therefore, it is crucial that the research community can count on diverse datasets so that a representative sample of the spectrum of abusive language is covered. Furthermore, when annotating a dataset, it is important to be consistent with existing research in the field in order to avoid research duplication and contradiction. This results in at least two challenges.

Diversity of collected data across datasets

The majority of the available collected datasets in the field share a data modality, language, and platform. Thus, data are shared in text format, are mostly in English, and in their majority, stem from Twitter (see, e.g., Davidson et al., 2017; Waseem & Hovy, 2016). This reduces the variability of the available data and, if the data overlap, also results in data redundancy with respect to repetitive data.

Redundant and contradicting labels

There has been confusion in terms of concept definition and the usage of the terms related to hate speech, abuse, and harm. This is critical since the use of different terms for equivalent categories hampers the reuse of resources. For instance, it is not clear what the differences between generic concepts, such as "toxicity," "offensive" and "abusive" are. Moreover, almost equivalent concepts such as "sexism" and "misogyny" are not always used in the same way. Detailed analyses of the diverging terms in abusive language dataset compilation and the consequences of this divergence are discussed in detail in Fortuna et al. (2020, 2021).

3 Towards transparent dataset construction and annotation guidelines

As seen in the previous section, there are a series of challenges that may undermine the quality of resources in the field of abusive language detection and analysis. Inspired by the study of these challenges, we propose, in what follows, a set of guidelines for leveraging quality resources. As in the previous section, we distinguish between intra-dataset and inter-dataset aspects.

3.1 Intra-dataset quality guidelines

As already pointed out above, in order to reduce the subjectivity and ambiguity of *concept definitions* used in the field it is important to follow certain guidelines.

Guidelines for concept definition

Adopt a problem-driven approach. Task definition should follow, as much as possible, a holistic, problem-driven approach, rather than a data-motivated approach. In other words, the task formulation should motivate data collection, instead of the task being defined based on the available data (Gudivada et al., 2015).

Motivate definitions, tasks, and datasets socially. Online hate speech, cyberbullying, abuse, and harm infliction are inter-personal behaviors with a strong social component. As these behaviors are within the scope of different academic disciplines, researchers in the field of abusive language detection should reach out to other relevant disciplines before defining the task and annotating data material. Literature from humanities and social science (including, for example, law, sociology, and anthropology) may become an important source of insight, together with existing surveys in the field (e.g., Fortuna & Nunes, 2018; Poletto et al., 2020; Schmidt & Wiegand, 2017; Vidgen & Derczynski, 2021).

Use clear category definitions. One of the goals of future annotation tasks should be to establish a clear taxonomy with meaningful and theoretically sound categories. Several theoretical studies already outline possible procedures concerning how this can be done (Vidgen & Derczynski, 2021). In this context, we should aim for explicit, precise, and universal category definitions. Such clear category definitions are instrumental in high-quality annotation.

Use coarse-grained categories with caution. In view of the challenges of using coarse-grained categories, we may conclude that such categories should be used with caution. In the case that they serve a given task or purpose well, they need to be clearly defined (see above), and the phenomena that they are supposed to cover should be clearly delimited. Along with each coarse-grained category, more specific categories, which further detail this category, should be spelled out and annotated such that an error analysis on the model performance can be conducted in order to assess whether it equally detects all the subcategories of a generic class (Fortuna et al., 2020, 2021; Pamungkas & Patti, 2019).

Prioritize fine-grained categories. Irrespective of the guideline above, previous research suggests that in the case of hate speech, fine-grained categories are better suited than coarse-grained categories. Experiments show that when a model is trained and tested on fine-grained categories, such as "sexism" or "racism," better levels of generalization are achieved. This also further buttresses the argumentation

in favor of more fine-grained taxonomies of abusive language categories. Despite this general rule, it is also important to note that excessively detailed taxonomies may lead to an unbalanced distribution of the data across categories, such that certain categories may end up with only a few samples. This would, obviously, also be problematic for standard supervised machine learning models. A compromise between taxonomy detail and data availability is thus necessary, and the granularity of the taxonomy should be carefully analyzed and justified from the perspective of the goals of a given experiment or study (Fortuna et al., 2021).

Match training and target data. For machine learning models to work, data collected for training and data to which the model is then applied (either for testing or, in the case of practical applications, during routine use) should share properties. One basic requirement is to observe that both share similar features, such as text length, style, and topic. Otherwise, the model generalizability capacities are put at risk.

Guidelines for bias mitigation

Data sampling techniques may involve decisions and the application of strategies for data collection that imply bias. If such decisions or strategies cannot be avoided, the focus should be on the minimization of their negative consequences and on the documentation of the data collection procedure, such that researchers in the field can select the datasets and procedures that best fit the application they are targeting. In what follows, we outline some guidelines for bias mitigation.

Increase versatile message searches. As already mentioned above, a common practice during data collection is the use of explicit keywords for the identification of relevant messages. While this practice has the advantage that it ensures the presence of abusive content, it has the drawback of introducing vocabulary bias into the collected material. The use of explicit keywords should thus be avoided and replaced by more versatile methods of message identification. Should keyword-based searches be necessary, a list of the keywords used should be provided in the data statement.

Control communities, message threads, and author distribution. Another strategy for data collection is to gather data from specific threads or from profiles that belong to authors previously identified as posting a higher rate of abusive content. This type of sampling procedure also has limitations since, if not controlled, the number of message authors will be small, and, as a result, the dataset

will be biased with respect to their writing style. A possible way to control this problem is to make sure that a reduced number of messages per author is collected. Another alternative is to ensure that the distribution of the authors in the data collection is balanced. In other words, text from the same author should not be present in both training and testing sets (see also, e.g., Arango et al., 2019). Controlling this type of bias will improve the model generalization as the model will be less tuned to a very reduced number of authors.

Control covered time spans. The data should cover a broad time span. Thus, while obtaining, for instance, feedback on an election, it makes sense to only gather data over a short period of time, to obtain a realistic picture of the use of abusive language in a social medium, the data should cover a longer time period since samples with narrower timeframes will be more affected by exceptional events. Again, the covered time span should be protocolled in the data statement, and in the case that any societal events influence the tenor and content of the data, this should be recorded as well.

Provide information on the annotator profiles. For the annotation of abuse language datasets, annotators with different backgrounds can be drawn upon (see Vidgen & Derczynski, 2021):

- trained specialists (one of the most common options that, however, usually provides little information on the type of annotators' expertise);
- crowdworkers (an option that is prone to trade quality for quantity);
- professional moderators (usually employees of a social medium platform who annotate following the platform's policies);
- a mix of crowdworkers and experts; and
- synthetic data creation (less representative of real-world data).

The profile of any of these types of annotators will necessarily influence the way an annotation will be carried out (and thus what the final annotated dataset will look like). Therefore, the profiles of the annotators involved should be properly described, such that biases can be measured and counter-balanced. The main information to be recorded in a strictly anonymized way concerns:

- demographic features (age, gender, nationality),
- annotation expertise, and

- personal experience with abuse (i.e., whether the annotator was a victim of online abuse).

It is furthermore important to add relevant attitudes and beliefs. Thus, attitudes toward discrimination and political orientation are closely related to the capacity of evaluating online abuse; annotating racist material in research should not, for example, be left to the discretion of a prejudiced individual.

Define precise guidelines for annotation. The annotation guidelines should be transparent and comprehensive. Rules should account for difficult or counter-intuitive cases, and a set of shared practices should be developed. The rules should be enriched with clear and easy-to-understand examples. Ideally, experienced annotators will be involved in the development of the guidelines since only they know the language used in the material and can thus ensure that it is captured in appropriate consistent categories (Vidgen & Derczynski, 2021).

Guidelines for data sharing

Let us now have a look at the guidelines for data sharing to ensure data preservation over time.

Protect the identity of the authors of the data. The identity of the authors of the data related to abuse language research must be protected, if not strictly anonymized, during data collection and also during the training, evaluation, and sharing of the material. With regard to a published dataset, IDs or user names that allow for the direct retrieval of the material from social media should not be freely published in an open repository. When it is necessary to share this type of information, the data should be kept private and only be accessed strictly in accordance with the terms of use of the data of the social medium in question and the relevant legal regulations.

Ensure data preservation and availability. As already mentioned, sharing or making data publicly available risks violating terms and conditions of social media platforms. On the other hand, using IDs instead of providing actual data poses data integrity risks. If both types of risks cannot be discarded in a concrete case, a possible solution is to use synthetic data, which would also solve the issue of data privacy and offers the advantage of allowing a better control of data quality. The disadvantage synthetic data brings is the loss of variability.

Data donations by social media platforms are another alternative, as are data trusts, which also provide a framework for storing and accessing data and respective terms and contracts for data access (Vidgen & Derczynski, 2021).

Include a data statement. When making a dataset available, it is important to provide detailed information on all stages of the dataset creation. This includes information on the following:

- task definition (concept definition, taxonomy, related concepts, targeted groups);
- decisions taken with respect to the data collection;
- data sampling procedure (social network, socio-historic data context, e.g., comments on news about politics or sports, the time and location of the data collection);
- researchers' and annotators' backgrounds;
- annotation guidelines (interviews, steps, task design on platforms); and
- class-balancing procedures.

Only with proper data annotation and dataset documentation will it be possible to achieve more standardization in the field.

Guidelines ensuring ethical principles

Last but not least, dataset creation must follow guidelines that ensure that the compilation procedures and the obtained dataset are in line with ethical principles.

Follow ethical principles. As already pointed out, technological solutions need to adhere to ethical principles and ensure that the harm done when developing a technology is minimized. These principles also apply to dataset collection and annotation. In this case, the main issue concerns bias, as discussed in the previous paragraphs (Bender et al., 2020; Tomašev et al., 2020), and the privacy of the message author and target. Datasets should be accompanied by a data statement in which the procedure followed to compile the dataset, the introduced bias, and the dataset purpose are described. It is only recently that some researchers in the field have started to adopt this practice of automatic hate speech detection (e.g., Sap et al., 2020).

3.2 Guidelines for inter-dataset coherence

The guidelines related to inter-dataset coherence concern, first of all, four aspects that are discussed below.

Diversify data characteristics. English is the most common *language* for the analysis of hate speech, but since hate in social media is a global phenomenon, other languages have to be considered as well, prioritizing under-represented languages. Due to the increased popularity of multilingual approaches, it would also be valuable to annotate equivalent phenomena in different languages at the same time. Code-mixed textual material has been collected in the community as well, which is adequate to represent online communication using more than one language at the same time.

The most common *source* of hate speech material with which the community works is Twitter. However, this also raises the question of platform diversification—especially in view of the specific characteristics of Twitter messages. In the future, platforms other than Twitter should be studied such that abusive language of the communities that use other platforms is also captured.

Regarding the *modality*, the majority of datasets only contain textual material, while image, audio, or multimodal data can also be relevant. Furthermore, it is necessary to keep in mind that the context of the collected material provides essential clues for the assessment of whether certain data are abusive. In the case of texts, this can be achieved by collecting complete conversation threads, including the main stimulus invoking a thread (e.g., news, a comment, a video) and replies to it. For instance, certain communities use slurs as a sign of identity. Multimodal context information can help to identify them.

As far as the *dataset size* is concerned, supervised machine learning-based techniques require large amounts of high-quality annotated data. Automatically annotated datasets may help to create bigger data collections. However, even if quantity matters, it is important to ensure annotation quality—for instance, by contrasting a manually annotated data sample with the corresponding automatically annotated one. Finally, it would be advantageous to also annotate other dataset characteristics in terms of linguistic features, including, for example, "overture" ("covert abuse" vs. "overt abuse"), "irony," "sarcasm," "adversarial," etc. From the previous literature (Caselli et al., 2020; Fortuna et al., 2021; Sanguinetti et al., 2018), we know that online abuse correlates with these characteristics, and their annotation would help to better understand this correlation.

Avoid redundant labels. Given the amount of ongoing research and available datasets in the field of abusive language detection and classification, it is also necessary to position a new dataset in the context of the datasets that already exist. The community should avoid creating new categories to refer to concepts already present in the literature and move toward dataset standardization. Previous work has shown that categories such as "toxicity," "offensive," and "abusive" correlate well with each other and lead to good cross-dataset generalization when used as training categories (Fortuna et al., 2021). With this in mind, it is appropriate to introduce a generic category term, "abuse and harms" to replace "toxicity," "offensive," and "abusive." This term also captures the recent insight into the community reflected by the change in the title of the most popular workshop in the area from *Abusive Language Workshop* to *Workshop on Online Abuse and Harms*. In Fortuna et al. (2021), it was also observed that classifiers trained on the categories of "sexism" and "misogyny" achieved a cross-dataset generalization between both concepts, indicating that using the label of "sexism" to refer to both would avoid the need for an extra label.

Provide definitions, examples, and justifications. In the case that a new category is identified, clear examples and justification of why a new category is needed and in what way it enriches the field should be provided. Due to its importance, again, the process of the definition of a new category should be well documented and grounded, based on the insights from social sciences.

Position new concepts on the map of standardized categories. Previous research provides standardized categories that allow for the conversion between different datasets (Fortuna et al., 2021). In this study, different publicly available datasets on abuse in English are annotated with respect to their similarity and compatibility. In the future, studies on other, new, datasets should conduct the same type of analysis. However, the question of how to ensure dataset standardization may remain, and there is no simple answer to it.

In any case, existing dataset definitions and surveys of existing datasets should be taken into account, and already introduced notions and categories should be adopted whenever possible. For instance, if a dataset contains and is annotated with "sexism" and "racism" categories, the creators of the dataset may compare these categories with more generic categories, such as "hate speech," "abuse and harms" and assess to what extent the targeted phenomena relate to one of these categories (obviously, this does not mean that coarse-grained categories are to be preferred (see also our discussion above).

Careful data analysis can help detect similarity between datasets. This can consist of a comparison of dataset feature descriptions, application algorithms for text similarity detection, topic extraction, and class comparison or cross-dataset classification (Fortuna et al., 2020, 2021).

4 Prospects of uniform annotation across abusive language detection applications

The evolution from the early to the latest research in the field of abusive language detection shows that it is difficult to predict in advance all the problems and nuances related to defining tasks and collecting and annotating data in this field. However, the field has also advanced considerably. While in the early era, proprietary datasets were created, and rarely generalizable models were developed, this tendency has changed in recent years. Now it is time to identify the remaining challenges and to agree collectively on strategies aimed at achieving a more mature research area.

In this article, we enumerated what we consider to be the central challenges of the field, which include the need for better and clearer concept definitions; the lack of data diversity in terms of languages and the platforms analyzed; the introduced bias when collecting, annotating, and publishing data; and the creation of new data resources that are compatible with the previous research in the field. To address these challenges, guidelines, which are summarized in the following set of instructions, were discussed and proposed:

- find solid theoretical ground (from social sciences and previous research in the field) and prefer clear fine-grained definitions;
- diversify data (e.g., find new data source languages and provide the data context);
- mitigate bias by controlling the message search, data properties, and data annotation (e.g., provide information on authors, topics, dates, annotation procedures);
- ensure data availability, but at same time, protect the authors of the data;

- Well document the data and the methodologies followed to compile them (e.g., include a data statement); and
- follow ethical guidelines.

Steps toward more maturity with respect to dataset collection and annotation can be observed. Datasets are becoming more diverse, with new languages and modalities being annotated (de Gibert et al., 2018; Suryawanshi et al., 2020). Data quality is being discussed (Vidgen & Derczynski, 2021), and datasets and annotation schemas (Fortuna et al., 2020, 2021) are being compared in search of good practices. Platforms that ensure data availability while observing content author privacy are also beginning to emerge.[3]

Another tendency that can be observed involves gathering and merging existing resources and building new annotation schemes based on this material, instead of always collecting and annotating new datasets, as was done in earlier research. This leads to more extensive and alternative collections of data (Sap et al., 2020).

Paula Fortuna is a final year PhD student at the Department of Information and Communication Technologies of the Universitat Pompeu Fabra of Barcelona, Spain. https://orcid.org/0000-0002-2306-9276

Juan Soler-Company is a data scientist in Pepsico. Previously he was a postdoctoral researcher at the Natural Language Processing group of Universitat Pompeu Fabra of Barcelona, Spain. https://orcid.org/0000-0002-8645-0162

Leo Wanner is ICREA Research Professor at the Department of Information and Communication Technologies of the Universitat Pompeu Fabra of Barcelona, Spain. https://orcid.org/0000-0002-9446-3748

Acknowledgments

The first author is supported by the research grant SFRH/BD/143623/2019, provided by the Portuguese National Funding Agency for Science, Research, and Technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of the *Human Capital* Operational Program (POCH), supported by the European Social

3 https://hatespeechdata.com/

Fund and by national funds from MCTES. The work of the second and third authors has been supported by the European Commission in the context of the H2020 Research Program under contract numbers 700024 and 786731.

References

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 45–54).

Bender, E. M., Hovy, D., & Schofield, A. (2020). Integrating ethics into the NLP curriculum. In A. Savary & Y. Zhang (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6–9). https://www.aclweb.org/anthology/2020.acl-tutorials.2/

Caselli, T., Basile, V., Mitrovic, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, ..., & S. Piperidis (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020* (pp. 6193–6202). https://www.aclweb.org/anthology/2020.lrec-1.760/

Davidson, T., Warmsley, D., Macy, M. W., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media* (pp. 512–515). https://doi.org/10.48550/arXiv.1703.04009

de Gibert, O., Pérez, N., Pablos, A. G., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In D. Fiser, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online* (pp. 11–20). https://doi.org/10.18653/v1/w18-5102

European Commission. (2017). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Tackling illegal content online, towards an enhanced responsibility of online platforms.* Reference: *COM* (2017)555.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computer Surveys*, *51*(4), 1–30.

Fortuna, P., Soler-Company, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6788–6796).

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing Management*, *58*(3), 102524. https://doi.org/10.1016/j.ipm.2021.102524

Gudivada, V. N., Baeza-Yates, R., & Raghavan, V. V. (2015). Big data: Promises and problems. *Computer*, *48*(3), 20–23. https://doi.org/10.1109/MC.2015.62

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, 1-11.

Le, T., Wang, S., & Lee, D. (2020). Malcom: Generating malicious comments to attack neural fake news detection models. *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)* (pp. 282–291).

Oliva, T. D., Antonialli, D. M., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LQBTQ voices online. *Sexuality & Culture*, *25*(2), 700–732.

Pamungkas, E. W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 363–370).

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, *55(2),* 1–47.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, ..., & T. Tokunaga (Eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018*. http://www.lrec-conf.org/proceedings/lrec2018/summaries/710.html

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics* (pp. 1668–1678). https://doi.org/10.18653/v1/p19-1163

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5477–5490). https://doi.org/10.18653/v1/2020.acl-main.486

Sarkar, R., & KhudaBukhsh, A. R. (2021). Are chess discussions racist? An adversarial hate speech data set. *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35 (pp. 15881–15882).

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In L.-W. Ku & C.-T. Li (Eds.), *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media* (pp. 1–10). https://doi.org/10.18653/v1/w17-1101

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). Multimodal meme dataset (multioff) for identifying offensive content in image and text. In R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, & D. Kadar (Eds.), *Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying* (pp. 32–41). https://www.aclweb.org/anthology/2020.trac-1.6/

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C., Ezer, D., van der Haert, F. C., Mugisha, F., Haert F.C., Mugisha F., Abila G. (2020). AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, *11*(1), 1–6.

Vidgen, B., & Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, *15*(12), 1–32. https://doi.org/10.1371/journal.pone.0243300

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *Proceedings of the 3rd Workshop on Abusive Language Online* (pp. 80–93).

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *In Proceedings of the NAACL Student Research Workshop* (pp. 88–93).

Weir-Reeves, J. (2010). Is the N-word acceptable? *The Temple News*. https://temple-news.com/is-the-n-word-acceptable/

Digital
Communication
Research.de

Recommended citation: Kirtz, J. L., & Talat, Z. (2023). Futures for research on hate speech in online social media platforms. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 467–482). Digital Communication Research. https://doi.org/10.48541/dcr.v12.27

Abstract: This chapter provides an overview of the various themes and points of connections between the various chapters in this section and outlines the current limitations as well as the major social and technical issues that still need to be addressed in hate speech detection. In particular, Kirtz and Talat discuss the ways in contexts—from legal contexts such as laws determining data collection methods to sociocultural contexts like annotator knowledge—affect the possibilities for the machine learning pipelines. Along with identifying current issues and limitations, Kirtz and Talat delineate future avenues for hate speech detection research.

DOI 10.48541/dcr.v12.27

Jaime Lee Kirtz & Zeerak Talat

Futures for Research on Hate Speech in Online Social Media Platforms

1 Introduction

Given their networked structure and the various affordances such as sharing features and algorithmic recommendation systems, social media platforms make it easier for users and communities to connect, organize, and share content. However, these same affordances and structure also enable social media platforms to act as effective facilitators for the dissemination and amplification of hate speech and incivility (Matamoros-Fernández & Farkas, 2021; Schmid et al., 2022). Subsequently, researchers have observed the growth of racism, sexism, homophobia, and numerous other discriminatory attitudes and beliefs as more and more abusive and hateful content is circulated and generated by increasingly interconnected users (Massanari, 2017; Matamoros-Fernández, 2017). This rise in abusive content has coincided with political shifts to the right occurring at national and international levels, resulting in the hyperactivity of hate speech (Johnson et al., 2019; Rieger et al., 2021; Mathew et al., 2020; 2019; Bilewicz & Soral, 2020).

As a consequence of the increasing prevalence of discriminatory and hateful attitudes, researchers have turned to the question of online hate speech from a number of different disciplines to propose solutions, many of which rely on machine learning models for hate speech detection such as automated content moderation

systems. Yet, as the texts in this section of the volume have shown, the analysis and detection of hate speech in machine learning faces challenges at every step of the research pipeline: from the legal frameworks for data collection, to the annotation and creation of datasets, and to the evaluation and application of machine learning models in automated content moderation systems. Throughout this section, the authors point out that the limitations for identifying hate speech are not necessarily due to technological restrictions, but rather due to the difficult nature of hate speech, and indeed language itself. Thus, in this chapter, we track these concerns across the various works within this section in order to outline the current limitations and major social *and* technical issues that still need to be addressed while also identifying future avenues for hate speech detection research.

2 Contextualizing context

At first glance, hate speech seems simple: it is the expression of hatred toward someone or some community. But as the chapters in this section discuss, hate speech is anything but simplistic. This is because words are always in relation to one another, to the individual, to the cultural and political modes and structures, to the medium or format. It is this relational quality, something we refer to as context, that makes hate speech so difficult. However, context cannot be eliminated or ignored as context is necessary in order to produce and comprehend the meaning that affords insight into whether speech indeed ventures into the realm of hate speech. This anthology makes apparent the need to further explore and understand how context affects identification at linguistic, semiotic, procedural and technical levels.

Context acts as a type of frame by encompassing that which surrounds a communicative event or text and occurs at moments of production, dissemination and interpretation. It influences how meaning is encoded in the production of text, how the text is disseminated and how the text's meaning is decoded, i.e., interpretated. How an individual produces a communicative event, such as a post on social media, is shaped by numerous intersecting and multi-layered forms of contexts. These include: the rules of language like grammar (linguistic context); the technical infrastructure, i.e., the platform technology, the legal infrastructure, etc. (situational context); and cultural beliefs, social backgrounds

and frameworks of knowledge (sociocultural context). All of these different, intersecting contexts thus have a profound effect on both the expression of hate speech and efforts to combat it.

On a base level, knowledge or awareness of the linguistic context or specific language and grammatical rules is needed in order to be able to read and write and this applies to authors, readers, coders and even in some cases, machine learning algorithms. Beyond a general knowledge of language and its rules, there are necessary contextual requirements for specific invocations and uses of words and phrases. For example, subcultural context is not only necessary for someone posting a hateful message, particularly those that seek to evade content moderation systems, but also is necessary for data annotators or coders to understand the text's intended meaning.

Another complication in hate speech detection research is that it requires an awareness of the legal contexts which dictate where data can be collected and how it may be used, and reused within academic pursuits. The legal frameworks put in place by states and platforms for user-to-user interaction govern the ways in which data may be collected, constructed, and subsequently shared. For instance, where Twitter actively provides an API that allows for scraping and sharing of social media data, other platforms such as Meta's Facebook have gone through several iterations of opening and restricting data access for public research. These distinctions and changes over time have severe impact on the data that can be collected, the legality of collection and sharing, and the possibilities that are afforded by any data that has been collected.

These contexts come to affect the possibilities for the machine learning pipelines. In particular, they are apparent in a) how data can be constructed, and the reductions that are necessary to transform data into machine readable formats; b) the construction of disjointed and incompatible datasets for abuse detection; c) how such contexts limit the choice of machine learning models; and d) the selection of appropriate evaluation metrics for hate speech detection.

In this chapter, we emphasize how the preceding chapters in this section introduce new forms of contexts and problematise current limits in context for hate speech detection. We argue that in order to address the limitations of hate speech detection, particularly around context, future work within the field of hate speech detection must take seriously the questions of sustained data access, annotator knowledge, domain specificity and transfer, and devote resources

towards online learning. By addressing these concerns, the task of hate speech detection can begin to realize its goals of protecting marginalized communities from being subject to hate speech in online spaces.

3 Contexts in limbo

Several of the chapters in this section outline various strategies involving rhetorical and semiotic tropes employed by users/authors to evade content moderation systems. Many of these strategies focus on the use of absent references, such as comments that use only subject nouns and/or no proper nouns or comments that make extratextual references. Thus, these strategies involve the purposeful elision of linguistic and rhetorical context and make it difficult to decode the comment's meaning. These absent reference strategies are highly effective because machine learning models have difficulty in assessing hate speech when the intended target or meaning is not explicit. For example, in their chapter on implicit modes of antisemitism, Becker and Troschke, illustrate how many comments avoid hate speech detection models by using references to subjects in earlier comments or parent threads through subject pronouns like "he" or "they." Because the subject, a known Jewish figure like George Soros, was not explicitly named as such and is instead implicitly referenced through the use of situational anaphora, the subject cannot be recognized as the intended Jewish object and thus, the intended meaning, namely the antisemitic sentiment, is unable to be understood without additional context, i.e., the parent comments that contain the original explicit naming of the subject. However, the questions around appropriate contexts occur at much earlier steps, as the chapter by Leerssen et al. remind us.

In their chapter, Leerssen et al. examine the legal context surrounding data collection and access for social media researchers and the ways in which law helps to both restrict and enable access to important data. The authors argue that because most researchers are only able to obtain data through data-sharing arrangements with platforms, these platforms maintain the legal and technical power to determine what kinds of data is shared and how it is accessed. However, platforms are often resistant to sharing sensitive data such as removed content (i.e., hate speech), thus producing a situation in which researchers like those studying hate speech are routinely denied access to this data with no legal

options for recourse. Such a denial of access has profound impacts on the knowledge that can be produced and held outside of private corporations. Moreover, the denial of access has impact on information that is available to legislators surrounding questions of discrimination and marginalization, and, as Bahador raise in their chapter, how intolerance may be rising towards communities.

Leerssen et al. also discuss how access is legislated or brought into being through proposed laws, such as Article 31 of the European Union's Digital Services Act, which requires platforms to make certain data available to vetted researchers. However, as they point out, many of these initiatives are in early stages and have yet to be fully developed or implemented and, as such, the success of these programs is difficult to assess. Furthermore, many of these proposals raise questions about power and privilege, such as those around the required qualifications of researchers in the vetting process.

These disparities in access risk creating tiered systems in which established researchers and institutions can gain access to information that is otherwise not available for academic scrutiny. In such, certain narratives around appropriate measures and perspectives are likely to have an outsized influence over future research and policy directions. Moreover, this disparity is likely to create a group of second-class citizens, in terms of research methods that are feasible with the data available. Thus, the vetting process, as discussed by Leerssen et al., risks consolidating influence and power over public research and policy within a small set of institutions and individuals, to the detriment of a breadth of research and insight into the issue of online hate speech and its causes.

Once the decision of collecting data and access has been established, technological affordances come to determine how the data is collected, structured, and accessed. Examining the process from decision to storage, Jünger engages a close reading of each stage of the data collection pipeline to make visible the underlying organizational structures and logics. For example, during the data extraction process, there are numerous elements, such as webpage footers or certain metadata, that are deleted or omitted. This practice of omittance can vary depending on the API, often provided by the platform, or through the user initiating the data scraping process; however, either way, there are deliberate choices being made about what is and is not valuable information and this shapes the data that is then used in machine learning. In their chapter, Jünger extends the mission of data hermeneutics from "interpreting, reconstructing and explaining the overarching narratives that

underpin social media conversations" to include the interpretation and explanation of the narratives that underpin the processes of data collection and assembly (Gerbaudo, 2016, p. 100). As such, Jünger's contribution addresses the problem of context through the necessary interrogation of how, where, and why databases for machine learning are formed and shaped by both socio-political and technical structures.

What Jünger's work points to is a necessity for data analytics and machine learning to deeply consider the processes and the affordances that outline, shape, and determine the datasets. Through such analyses, machine learning researchers can come to understand the powers that shape how the technologies may be used and who they serve while also pointing to the particular groups and societies which remain under-served by machine learning technologies. That is, we can come to understand the political life of data and machine learning by understanding the deliberate choices that shape the data that is collected, stored, and used for machine learning.

Once the type of data and the methods for data collection have been decided, it becomes necessary to define hate speech. In their chapter, Bahador turns a critical eye towards the limits of contemporary hate speech definitions and their ramifications for the monitoring of hate speech. Bahador emphasizes the ways in which hate is a product of escalation that ultimately leads to outright hatred. As fascism and the conservative right are on the rise across the globe, so are the precursors to hate speech and violent hatred.

Relying on this, Bahador exposes how the over-emphasis on hatred creates a situation where efforts towards computationally mitigating harms occurs after the basis for hateful rhetoric has already been established. This emphasis on hate speech further leads to an over-emphasis on individual target groups, rather than the social and linguistic commonalities in hate speech and its precursors. Bahador thus offers a recontextualization of hate speech away from hate speech itself and onto the shared characteristics that lead to hate speech.

Such a recontextualization of hate speech very widely opens up new avenues for research into hate speech detection, and in particular provides space for the task of hate speech detection to attend more closely to its mission of protecting those who are at most risk of being targets for online abuse and harassment. In particular, this recontextualization affords developing technologies and strategies to address rising intolerance, which is often directed towards communities that are already marginalized and minoritized.

However, regardless of working within or expanding our current definitions, Kim argues hate speech is highly complex, contextual, and socially determined which factors into the annotation of data for training machine learning models. But as Kim gestures, the solution to biased datasets is not simply introducing new datasets or re-defining hate speech with a new context. In order to combat the issue of bias in hate speech training datasets, Kim advocates for a fundamental shift in both how hate speech is understood *and* how the problem of hate speech is framed. Rather than focusing on what is or is not hate speech (i.e., determinations of hate speech) as much contemporary research tends to do, they utilize an intersectional perspective to reframe efforts around who determines hate speech and how this designation is determined.

Operationalizing this perspective, Kim proposes two principles for researchers, namely: transparency and inclusion. The former emphasizes the contestable nature of hate speech and Kim offers suggestions for researchers such as the inclusion of position statements in publications. The latter principle, inclusion, acknowledges how hate speech detection automation disproportionately impacts certain groups, particularly those with multiple marginalized identities (e.g., Black women) and seeks to include those most likely impacted in the data collection and annotation process.

What this chapter points to is a fundamental issue of objectivity and knowledge production, in that knowledge is never objective, but always grounded in situational context and subjectivity. This is something that critical race scholars, such as Kimberlé Crenshaw (Crenshaw, 1991, cited by Kim), as well as scholars in feminist technoscience, and science and technology studies have extensively written on.[1] As such, it is not enough to understand if data is biased but as Kim argues, we need to interrogate the underlying power relations—both in between and within groups—if we want to truly address the problem of bias.

While creating new datasets itself does not address biases in the datasets, increasing interoperability, and ensuring that new datasets are conjunctive with pre-existing can increase the usability and lifespan of all datasets available. Highlighting how contemporary contexts within data creation are disjunctive, Fortuna et al. argue that the result is methods that are not comparable with one another. In

1 See also Balsamo, 1996; Barad, 2007; Browne, 2015; Bucher, 2018; D'Ignazio & Klein, 2020; Haraway, 1991; McPherson, 2018; Noble, 2018; Suchman, 2008; Wajcman, 1991, 2004.

particular, the authors argue that contemporary datasets, by virtue of incompatible typologies of abusive language provide a challenge for research by not affording a full exploration of the concept of online abuse. In this way, Fortuna et al. draw attention to an inherent tension in the detection of online abuse and hate speech: At which junction does the contextualized and situated experiences of groups and individuals require departing from pre-existing typologies of abuse. In spite of early efforts towards creating unifying typologies (e.g., Talat et al., 2017), a number of different typologies of abuse have been proposed. On one hand, Fortuna et al.'s argument for a consolidation of annotation typologies can provide space for the deeper and wider exploration of abuse within individual contexts. On the other hand, a commitment towards consolidation also forecloses the possibility of disagreement between typologies that reflect the embodied and situated experiences, for instance across different identity groups and their particular needs.

Many datasets for online abuse rely on datasets that are collected with majoritarian perspectives (Davidson et al., 2019; Sap et al., 2019; Thylstrup & Talat, 2020), and most frequently collected for the English language (Vidgen & Derczynski, 2020) in the North American context. In light of the critique in Fortuna et al.'s chapter, the disconnect between different annotation frameworks and typologies has in most cases not been motivated by a distinctive need of individual groups. This disconnect, however, has also afforded a wide variety of positions through which we have come to understand the conceptualizations of hate and abuse of one group, and the fallouts when systems trained on the data of one group has been applied widely across groups with distinct needs and desires. That is, neither conjunctive or disjunctive datasets and annotation typologies are a unilateral good, but must be considered in the moment with attention and respect to the particular goals of the annotation processes.

However, even when operating for only a single group, annotating data provides significant complexities, as Becker and Troschke detail. In their chapter, they perform a case study on antisemitism to identify and address the difficulties in interpretation of implicitly produced meanings and present their approach to developing a differentiated code system for annotation of implicit meaning. One of the most relevant aspects of this chapter is how the authors approach implicit meaning, wherein rather than simply naming the form of implicit meaning, such as irony or anaphora, they classify implicit meaning through the types of knowledge required to extrapolate it. The chapter focuses on three different kinds of

knowledge, namely: language knowledge—knowledge about the structure and rules of language; context knowledge—knowledge about the specific situation, such as the original post an antisemitic comment is responding to; and world knowledge—general cultural and discursive knowledge about social norms, spaces and subject matter. Becker and Troschke then demonstrate how these knowledge areas interact to produce implicit meanings using examples from their research, such as how language and world knowledge interact in irony.

In this, Becker and Troschke show different levels of contexts required to understand and annotate the texts themselves. This is further evidenced by Baden, who argues that content moderation technologies and antagonist users are engaging in an arms race, where ever-more sophisticated computational content moderation methods are met with increasingly sophisticated evasive manoeuvres to avoid detection by such filters. In particular, Baden argues that there is a need for shifting the context of research efforts from explicit hate speech, as computational methods have improved in their ability to detect this form of hate, to more implicit and context dependent forms of hate. With such a context shift in research also comes a distinction in how technologies are situated culturally. Where explicit hatred may be more easily detected across cultural contexts, Baden argues that systems for implicit hate speech will require cultural competency and therefore a requirement that hate speech detection systems are grounded within the cultures that seek to be protected from hate speech. By shifting from general purpose to culturally grounded systems, the evasiveness of language can also be addressed, as the reading and understanding of text and context will be situated within the understanding of the reader. Content moderation systems can thus engage as third parties that act on behalf of the reader—situated within the context of the reader, rather than as an external third party as they currently exist (Thylstrup & Talat, 2020).

In making such a shift in the cultural situatedness of machine learning models, it is also necessary to make appropriate shifts in the methods by which data is made, and the reductions that are necessary for each cultural context. In this volume, Laaksonen outlines the particular methods by which hate speech is made into data for machine learning. Thus, their chapter addresses the linguistic and cultural contexts and complexities that are reduced away, in order to make hate speech a computational concept. Laaksonen's intervention of context builds on that which was proposed by Baden. Rather than understanding hate speech as

an immovable entity, Laaksonen insists that systems for the detection of hate speech must operate iteratively, that is data must continuously be made available for models to remain relevant and applicable to the changes and developments in how hate speech is produced.

Through the emphasis on the reductions in complexity, Laaksonen makes abundantly clear the limitations of the machine learning approach to hate speech detection, which necessitates the loss of the very context that is fundamental to the functioning of hate speech. Without such context, the process and outcomes of predicting hate speech have a vital lack of ability to accurately disentangle the hateful from the non-hateful. Perhaps more critically, machine learning models that are trained without appropriate contextual information will lack the ability to situate correct classifications within the context that they are hateful.

Beyond the contexts of data that have been highlighted, building automated systems for hate speech detection is itself a deeply contextual task as Stoll shows in their chapter. Stoll provides a step-by-step consideration of how machine learning classifiers for hate speech detection can appear to have high performances, while being fundamentally broken. Through a construction of the appropriate and the "phony," Stoll provides a criticism of statistical machine learning-based approaches to hate speech detection arguing that "machine learning *is* just statistics. And consequently, we are still stuck with the same questions and pitfalls social scientists already know about well enough." Thus, Stoll contextualizes statistical machine learning for hate speech as a theoretical research question, rather than the practical question that machine learning researchers often propose.

This challenge to the predominant context in the machine learning literature raises the question of whether machine learning models are at all appropriate for hate speech detection. On the one hand, Stoll's contextualization offers an analytical vision for machine learning models for hate speech detection, which has the purposes of understanding social climates. On the other hand, machine learning's contextualization of content moderation imagines an applied focus, where the purpose is not understanding but social control. Although these two contexts appear, at first glance, to be at odds, we propose that they are complementary. That is, we argue that an automated approach to content moderation cannot stand without the analytical insights of the social phenomena that underlie the need for content moderation systems.

For either the predictive or analytical use of machine learning for hate speech it is necessary to consider the means of validating, evaluating, and explaining machine learning models and their outputs. However, depending on the particular use case, different and discrepant notions of evaluation and validation may be necessary. In their chapter, Laugwitz speaks to the discrepancies between algorithmic and social scientific explanations and rationalization. Laugwitz argues that there is an epistemic gap between the evidence that is offered by hate speech detection models, and the explainability models and methods applied to them, and the burden of evidence required in communications research. The latter operates with a priori rationalization which is tested a posteriori through empirical tests. The former, on the other hand assumes that a priori knowledge is only required to a lesser degree (e.g., a priori considerations are apparent in the development of features or rationalisation over model architecture), shifting its focus to a posteriori analysis of constructed systems. Here Laugwitz argues that contemporary methods for evaluating model validity, through understanding correlations in models or their outputs do not fully satisfy the need for validating models, as these do not concisely or adequately explain model behaviour. That is, Laugwitz argues that the scientific and validation practices of the computational fields and the communication field are complementary and provide distinct insights that are required for effective and productive content moderation systems.

In this recontextualization of validation, Laugwitz comes to offer a mode of operationalizing machine learning technologies as cultural probes, for which a priori hypothesis can be formulated and in which the output is a deeper understanding of the problem of hate speech. This operationalization stands in contrast to contemporary forms of hate speech detection systems, that seek the allure of categorization and sanitization that is offered by content moderation technologies (Thylstrup & Talat, 2020).

4 Futures

Collectively, the chapters emphasize that the problem of hate speech is a social problem, but it has been characterized as a technical problem and been addressed through technical solutions such as hate speech detection tools that employ machine learning models. This results in a problematic scenario in which

unquantifiable, affective discourses are put into discrete terms and as such, context and meaning are lost both at the encoding and decoding stages. This is similar to the conversion of a signal from analogue to digital, where the rounded waveform with continuous values transforms into a stepped function with sharp edges and discrete points.

This treatment of a social problem as a technical problem gives rise to the limitations that the authors highlight in this volume. To address this fundamental mismatch between the task and its operationalization requires starting from the knowledges required to contextualize and understand hate speech. On a higher level of abstraction, researchers in hate speech detection can take from these chapters a need for explicating how data and machine learning models are situated and which perspectives these seek to reproduce. This includes taking an intersectional, critical approach as proposed in the chapters by Fortuna and Kim. This includes a commitment to consolidation that needs to occur within individual demographic groups that have overlapping understandings of abuse and hate speech—and typologies must diverge where one typology cannot account for the particular needs of a group. In addition, it is imperative that future work should treat bias as a question of power and situationality, such that it is clear who is producing models and data, and which perspectives these seek to encode.

Further, as Leerssen argues, there is a need for strong legal protections for hate speech data for research, and researchers can push towards new forms of sharing data and requiring large social media companies to make data available for research purposes. Future directions include building off these nascent initiatives, which seek to inscribe regulatory data access practices into law, this chapter argues that legislating access is a potential path forward for researchers.

In addition to increasing access to data for researchers through legal avenues, many of the chapters point to the need for future interventions at the level of data collection and classification through critical inquiry and reflection. There is an imperative need for considering how data is derived for machine learning, in the process of building such technologies. Future work for hate speech detection should therefore strongly heed Stoll's warning that machine learning efforts are building "phony classifiers" that only have an appearance of working. Attending to this warning, researchers and practitioners must address each step of the machine learning pipeline, such that the methods and data answer active research questions surrounding efforts to understand the social phenomena that give rise

to the need for content moderation, and how that need changes over time, place, and culture. By examining and understanding the power relations and the decisions that give rise to the specific form of data, we can come to understand how technologies for hate speech detection privilege and marginalize communities on the basis of the ways in which researchers and practitioners interact with the larger social, technical and socio-technical structures at hand.

More practically, some chapters call for an increased attention to the annotation processes, with particular emphasis on the interoperability and ambiguity that inherently pose challenges to language technologies and culturally contextual concepts such as hate speech. To be able to situate the data and technologies, and identify when interoperability is appropriate, future research should remain in close dialogue with the communities that are affected by hateful rhetoric. Such close ties with communities are particularly important when addressing the question of rising intolerance towards communities, prior to the establishment of outright hatred towards them. By maintaining close ties to affected communities, researchers can engage in ongoing data making processes which can afford addressing the changing nature of hate speech whilst ensuring that evaluation of machine learning techniques are situated within the needs of individual communities, rather than an imagined universal public. Such community-based evaluation can further allow researchers to engage in-depth with questions surrounding the validity of models, i.e., that they produce correct predictions, and ensure that researchers develop research questions on the basis of the needs of communities and are given direct feedback where model explanations are incongruent with how harm is experienced.

The introduction of context, particularly sociocultural context in machine learning processes is echoed by Laaksonen. While this is an active research field (e.g., Gao & Huang, 2017; Chakrabarty et al., 2019), information beyond what is currently considered is needed. Context such as social and socio-political context, and geographic and cultural information is needed for machine learning models to be able to situate their predictions within the social context in which hate speech is hate speech.

In our reflection on the various contributions to this volume, we have sought to center the question of how each chapter imagines and reimagines context in the frame of hate speech detection. If we want to make lasting interventions into the proliferation of hate speech online, it is imperative that we shift from static to

dynamic, contextual based understandings of hate speech. Future efforts need to move away from technological solutionism and towards multidirectional, collectively driven projects that involve social and technological approaches.

Jaime Lee Kirtz is Assistant Professor of Media Studies in the School of Arts, Media and Engineering at Arizona State University, USA. https://orcid.org/0000-0002-3577-9689

Zeerak Talat is a research fellow at Mohamed Bin Zayed University of Artificial Intelligence, UAE. https://orcid.org/0000-0001-5503-867X

References

Balsamo, A. M. (1996). *Technologies of the gendered body: Reading cyborg women*. Duke University Press.

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology, 41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Browne, S. (2015). *Dark matters: On the surveillance of blackness*. Duke University Press.

Bucher, T. (2018). *If...then: Algorithmic power and politics.* Oxford University Press.

Chakrabarty, T., Gupta, K., & Muresan, S. (2019). Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, 70–79. Florence, Italy: ACL. https://doi.org/10.18653/v1/W19-3508

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against Women of Color. *Stanford Law Review, 43*(6), 1241–1299. https://doi.org/10.2307/1229039

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. Florence, Italy: ACL. https://doi.org/10.18653/v1/W19-3504

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press.

Gao, L., & Huang, R. (2017). Detecting online hate speech using context aware models. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, 260–266. Incoma Ltd. Shoumen, Bulgaria. https://doi.org/10.26615/978-954-452-049-6_036

Gerbaudo, P. (2016). From data analytics to data hermeneutics. Online political discussions, digital methods and the continuing relevance of interpretive approaches. *Digital Culture & Society, 2*(2), 95–112. https://doi.org/10.14361/dcs-2016-0207

Haraway, D. J. (1991). *Simians, cyborgs, and women: The reinvention of nature*. Routledge.

Johnson, N. F., Leahy, R., Johnson Restrepo, N., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature, 573*, 261–265. https://doi.org/10.1038/s41586-019-1494-7

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society, 19*(3), 329–346. https://doi.org/10.1177/1461444815608807

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society, 20*(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, 173–182. Boston Massachusetts USA: ACM. https://doi.org/10.1145/3292522.3326034.

Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., & Mukherjee, A. (2020). Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction 4 (CSCW2)*, 1–24. https://doi.org/10.1145/3415163

McPherson, T. (2018). *Feminist in a software lab: Difference + design. MetaLABprojects*. Harvard University Press.

Noble, S. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (2021). Assessing the extent and types of hate speech in fringe communities: A case study of Alt-right communities on 8chan, 4chan, and Reddit. *Social Media + Society, 7*(4). https://doi.org/10.1177/20563051211052906

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: ACL. https://doi.org/10.18653/v1/P19-1163

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, Advanced Online Publication. https://doi.org/10.1177/14614448221091185

Suchman, L. (2008). Feminist STS and the sciences of the artificial. In E. J. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies* (pp. 139–164). MIT Press

Talat, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, 78–84. Vancouver, BC, Canada: ACL. https://doi.org/10.18653/v1/W17-3012

Thylstrup, N., & Talat, Z. (2020). Detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3709719

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE, 15*(12). https://doi.org/10.1371/journal.pone.0243300

Wajcman, J. (1991). *Feminism confronts technology*. Pennsylvania State University Press.

Wajcman, J. (2004). *TechnoFeminism*. Polity.

About the editors

Martin Emmer is Professor for Media and Communication Studies at Freie Universität Berlin and was one of the founding directors of the Weizenbaum Institute for the Networked Society in Berlin. Among others, his research focuses on the use of digital media in international comparison, political online communication as well as communication policy for the digital society. His latest projects addressed the convergence of internet and television from a user perspective, digital media use by refugees and the development of methods for an automated analysis of online communication using the example of hate speech in social media.

Ulrike Klinger is Professor for Digital Democracy at the European New School of Digital Studies at European University Viadrina in Frankfurt (Oder) and an associated researcher at the Weizenbaum Institut for the Networked Society in Berlin. Her research focuses on political and digital communication. After her PhD on media pluralism in defective democracies (2010), she has worked extensively on the transformation of digital public spheres, the role of digital media in election campaigns, and the impact of technologies on public communication, e.g. algorithms and social bots.

Merja Mahrt is a Research Associate at the Weizenbaum Institute for the Networked Society, where she studies digitalization and its effects on individuals and society. She completed her "Habilitation" at Heinrich Heine University Düsseldorf and received her PhD from the University of Amsterdam, after studying communication and media at Freie Universität Berlin. Since 2020, she is chair of the Digital Communication Section of the DGPuK.

Sünje Paasch-Colberg is a communication researcher with a focus on media content research and works as a Research Associate at the German Centre for Integration and Migration Research (DeZIM) in Berlin. Her research focuses on issues of social cohesion and the media, more specifically on the dynamics of exclusion in online communication, the representation of social groups in the media, and media discourse on social inequality. Sünje studied communication and media at Freie Universität Berlin and Auckland University of Technology and has worked as a Research Assistant at Universität Freiburg/Université de Fribourg (Switzerland) and Freie Universität Berlin before joining DeZIM.

Christina Schumann is a Senior Researcher at the Department of "Empirical Media Research and Political Communication" at the Institute for Media and Communication Studies at Technische Universität Ilmenau. Her research focuses on digital communication as well as media reception and effects research. From 2012 to 2016, she was chair of the Digital Communication Section of the DGPuK.

Christian Strippel is head of the "Weizenbaum Panel" and the "Methods Lab" at the Weizenbaum Institute for the Networked Society, Berlin, Germany. His research interests include digital communication, media use, public theory, and sociology of science.

Monika Taddicken heads the Institute for Communication Science at the Technische Universität Braunschweig. She studied at the Georg-August-Universität Göttingen and received her doctorate at the University of Hohenheim on the subject of method effects in web surveys. Her research focuses on digital communication and science communication, particulary from the audience perspective. In particular, she focuses on science-related communication in new media environments. From 2012 to 2016, she was chair of the Digital Communication Section of the DGPuK.

Joachim Trebbe is Professor for Media and Communication Studies at Freie Universität Berlin. His subject areas are research methods and media content research. After receiving his PhD in Berlin he worked nearly ten years as Professor for communication research in Fribourg / Switzerland before returning to his alma mater. His research focus lies on trends in television programming and streaming/ video on demand. He spend some time on media, migration and social integration, in particular focusing on residents in Germany from Turkish origin. He is currently involved in international health communication research about infection control in the global south. Besides that he is fascinated by machine learning methods for applied communication research.

Martin Welker is Professor of Journalism and Corporate Communication at HMKW Hochschule für Medien, Kommunikation und Wirtschaft (University of Applied Sciences) in Frankfurt am Main. He heads the BA/MA-study program for journalism and communication. He studied at the University of Mannheim, received his doctorate in 2001 and worked as a Deputy Professor for Journalism at the University of Leipzig. His research covers communication practices in social media. Welker is editor of the "Neue Schriften zur Online-Forschung" at Herbert von Halem Verlag.