Futures for research on hate speech in online social media platforms

Kirtz, Jaime Lee; Talat, Zeerak

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.