The datafication of hate speech

Laaksonen, Salla-Maaria

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested 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). Berlin https://doi.org/10.48541/dcr.v12.18

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