Machines Do Not Decide Hate Speech:
Machine Learning, Power, and the Intersectional Approach

Jae Yeon Kim

jkim638@jhu.edu

Assistant Research Scholar, SNF Agora Institute, Johns Hopkins University

In preparation for *Challenges and Perspectives of Hate Speech Analysis*

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 digital hate speech. 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 paying attention to this data-generation process, building a fair automated hate speech detection system is unachievable. This article first examines the relationship between power, hate speech, and machine learning, then 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.

Keywords: Hate speech, machine learning, fairness, power, intersectionality

CV

Jae Yeon Kim is an Assistant Research Scholar at the SNF Agora Institute at Johns Hopkins University, United States. His research agenda is interdisciplinary and lies at the intersection of social sciences and data science. His research centers around two themes: advancing social science theory and methods on power, inequality, and resistance and making social science research more efficient and scalable. He received his Ph.D. in political science from the University of California-Berkeley.

Jae Yeon Kim thanks Thomas R. Davidson, Renata Barreto, two anonymous reviewers, and the editors of this volume for their constructive comments on the earlier draft of this chapter.

1 Introduction

The advent of social media platforms – such as Twitter, Facebook, and YouTube – has increased digital content – and with it, hate speech, defined as highly negative and often violent speech that targets historically disadvantaged groups (Walker, 1994; Jacobs et al., 1998; see also Sponholz in this volume). In response, social media platforms have leveraged machine learning to scale up their efforts to detect and moderate their users' contents (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 the gains in efficiency 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, LGBTQ, 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 for targeting hate speech.¹ The tricky issue is that speech by these historically disadvantaged groups is actually more likely than that of others to be labeled as hate speech (Sap et al., 2019). Ideally, advancements in machine learning should have solved this problem by developing automated hate speech detection systems that are efficient and fair. For instance, the probability of one's speech being labeled hate speech should be independent of whether that person 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,

¹ Twitter's policy on hateful conduct: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>
Facebook's hate speech community standards: https://m.facebook.com/communitystandards/hate_speech/
YouTube's hate speech policy: <https://support.google.com/youtube/answer/2801939?hl=en>

These links were all accessed on August 16, 2021.

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?

This situation is not ironic at all upon closer inspection. 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 the challenge is that hate speech annotation is about deciding whether particular speech violates social norms. However, most norms have boundaries that can vary depending on context. For instance, use of the “n-word” between whites and blacks 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 American speech as offensive than that of whites (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 academics define machine learning performance based on its prediction accuracy, but if the ground truth, which an algorithm predicts, is invalid, whether it predicts it effectively becomes a secondary question.

Without paying attention to this data generation process, building a fair and automated hate speech detection system is unachievable. Moving in this direction 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 our discussion of the obstacles that hinder progress in building a fair automated hate speech detection system and how to overcome these obstacles.

2 Bias in Machine Learning and Hate Speech Detection

Bias in a machine learning application usually is defined as a residual category of fairness (see Gajane & Pechenizkiy, 2017; Corbett-Davies & Goel, 2018; and Mitchell et al., 2021, for a review of the various definitions of fairness in machine learning applications). 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 there are many ways to define how a model's performance is uneven, 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. In this definition, a machine learning model is fair, in a binary classification case, if the 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, p. 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 that of others (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 missing elements in 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), it means that 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 the names of the job applicants varied. The field experiment results indicated that 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 from name attribute in resumes that may cause people to perceive a job applicant's race differently. In contrast, in machine learning applications, if the outcome is biased racially, it means that the model performs poorly for one racial group compared with others. Unlike in the earlier social science example, in this case, we do not focus on what caused the disparity in model performance across demographic groups that potentially could exacerbate existing socioeconomic inequities (Kasy & Abebe, 2020).

To understand bias in machine learning applications and where it originates, we need to 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 that it utilizes and the tasks it tackles. Humans are involved in both generating training data and defining the tasks, which are decisions 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). There is no panacea, and only careful investigation of underlying causes can lead to promising solutions.

Unfortunately, most presented solutions 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 is a laudable effort, but an exclusive focus on training data sets was insufficient to address the issue fully because representation bias is only one element of a broad set of societal and historical biases (Mehrabi et al., 2021). The more fundamental issue is not that training data sets lack sufficient numbers of black women faces, but why this practice was accepted and not questioned in the first place. The main issue is about power, not the size of observations related to different racial groups. Therefore, it is necessary to investigate the larger social environment that defines what kinds of training data and labeling processes are acceptable. This is particularly true when the 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).

An issue related to this label bias, and fundamental to making automated hate speech systems fair, is measurement. It is difficult to define what can be *objective* ground truth in hate speech data sets. If the goal of building a predictive model is differentiating between cats and dogs, a consensus easily could 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 it 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 that of others is not, the odds of my speech eliciting a political and legal toll is higher than those of others. In a hierarchical society, power relations are unequal, and these relations determine who shapes rules

and norms (Lukes, 1974). Thus, there is a greater chance that a subordinate group's speech will be labeled hate speech compared with that of a dominant group (Maass, 1999; Collins, 2002; Campbell-Kibler, 2009).

In the machine learning literature, researchers have circumvented this measurement problem by assuming that either well-defined ground truth exists or that the best approximation is available through social consensus (Dwork et al., 2012, p. 214). This is a convenient assumption upon which to build a compact theory, but an important obstacle to be acknowledged in practice. To acknowledge the relationship between power, normative judgments, and hate speech labeling, researchers should be conscious of 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 not enough to make an automated hate speech detection system fair (Blodgett et al., 2020).

In principle, this objective can be accomplished by promoting greater transparency and inclusion in building an automated hate speech detection system.

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 also should provide a position statement in their research that describes why they define hate speech in a particular way in 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 others. Model cards, a documentation tool for fair machine learning, provide one helpful illustration of this approach (Mitchell et al., 2019).

Inclusion: Furthermore, researchers should include people who would be most likely to be

harmful from automating hate speech detection in the process so that they could 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 an ethical, as well as a scientific, perspective. These individuals possess deep knowledge of what speech targets them and what part of their speech practice potentially could be mislabeled hate speech.

Practicing these principles requires thinking about power in two steps: How does the dominant group define societal norms (between-group power relations), and how do these societal norms marginalize particular segments of subordinate groups (within-group power relations)? Clarifying these points helps to identify which assumptions by researchers should be transparent and which members of marginalized groups should be invited to be research partners. In the next section, I discuss how the intersectional approach helps raise this type of awareness.

3 Why Intersectional Approach?

The intersectional approach is helpful in understanding what shapes people's perception of hate speech and how this bias is baked into data sets used to train hate speech detection algorithms. According to this perspective, hate speech can be intersubjectively defined on the basis of certain characteristics. This approach to hate speech differs from the one that focuses on hate speech analysis at the content level (for a discussion on distinctions between these approaches, see, e.g., Sellars, 2016, pp. 14–18). It is also different from another 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 made by their out-group's members more negatively than that created 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, men and women). Then, when aggregated, the biases of annotators from different backgrounds would cancel each other out.

However, taking this approach raises another question: How should we define diversity? The above approach works only if strong homogeneity of opinion on hate speech exists among members of a particular group. In practice, this means that if one investigates racial bias, then the researcher 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, gender) intersect and make a segment of a subordinate group more marginalized than others.

For instance, Cohen (1999) demonstrated how the intersection of race and sexuality explains African American communities' unwillingness to mobilize against the 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 (lesbian, gay, bisexual, transgender, and queer) 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 U.S., 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 it is highly

likely that hate speech targeting Black LGBTQ communities was not labeled hate speech. Even Black annotators might avoid labeling such attacks as hate speech because their community leaders did not address this problem publicly. These annotators might not be aware of 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 what thoughts, speech, and behaviors are acceptable within subordinate communities if they want to maintain their (moral) reputation 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 a blind spot 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 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 it is crucial to include people who are most likely to be harmed by hate speech in the creation of automating hate speech detection so that they could 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, it is helpful to assume that only one form of bias (e.g., racial bias) exists, while other forms (e.g., gender bias) do not. However, in reality, these various bias axes intersect, creating one segment of a historically disadvantaged group that suffers 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).

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*, 695–695. <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*, 77–91.
- 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. In *Conference on Neural Information Processing Systems*. <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*, 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*, 248–255. <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. <https://doi.org/10.1145/2090236.2090255>
- 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*, 259–268. <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. <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. <https://arxiv.org/abs/1908.07898>
- Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate

- speech detection. In *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. In *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. *Advances in Neural Information Processing Systems*, 3315–3323.
- Hinnefeld, J. H., Cooman, P., Mammo, N., & Deese, R. (2018). Evaluating fairness metrics in the presence of dataset bias. <https://arxiv.org/abs/1809.09245>
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(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 the *International Conference on Artificial Intelligence and Statistics*, 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*, 306–316. <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*, 576–586. <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*, 45–55.
<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*. <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*, 1–4. <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](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*, 220–229.
- 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*, 337–342. <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, 1668–1678. <https://doi.org/10.18653/v1/P19-1163>

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>

Sellers, A. (2016). Defining hate speech. *Berkman Klein Center Research Publication*,

(2016-20), 16–48. <https://doi.org/10.2139/ssrn.2882244>

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*, 53–59. <https://doi.org/10.18653/v1/W17-1606>

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*, 138–142.
- Waseem, Z., Davidson, T., Warmusley, 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. <https://doi.org/10.18653/v1/W17-3012>
- 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*, 88–93.
- 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>
- Zhou, X., Sap, M., Swayamdipta, S., Smith, N. A., & Choi, Y. (2021). Challenges in automated debiasing for toxic language detection. <https://arxiv.org/abs/2102.00086>