Enhancing Hate Speech Detection and Mitigation in Online Social Media Platforms using Natural Language Processing

Vigneshwaran Shankaran
vigneshwaran.shankaran@ut.ee

1 Introduction

1.1 Background and Motivation

Online social media platforms have revolutionized communication, bridging the gap between people across distances and allowing instant interactions. They have transformed the way we share information, opinions and connect with others, shaping a new era of digital socialization. However, alongside the benefits, online social media platforms have also given rise to a disturbing downside: the proliferation of hate speech. Unfortunately, this kind of speech can quickly spread and take on a life of its own, with like-minded individuals reinforcing each other's views and amplifying the message to a wider audience. The anonymity and ease of posting on these platforms have emboldened individuals to spread messages of prejudice, discrimination, and intolerance, fostering a toxic environment that threatens social harmony and undermines healthy communication. This can lead to a toxic and dangerous environment for marginalized groups and can fuel hate crimes and other forms of discrimination. Therefore, it is essential to have effective moderation policies and tools in place to combat hate speech and maintain a safe and inclusive space for all users on social media platforms.

Unfortunately, the definition of hate speech is not precisely defined due to its complex phenomena, organically related interactions between groups, and reliance upon language subtleties. De Gibert et al. (2018) defines hate speech as a deliberate attack directed towards a specific group of people motivated by aspects of the group's identity. Nobata et al. (2016) defines it as "language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity." Social media giants like Facebook (2023), Youtube (2023), Twitter (2023), etc, have their own definitions as well.

1.2 Problem Statement

The intricate and multifaceted nature of hate speech on digital platforms presents a formidable obstacle, rendering its identification, categorization, and comprehension a challenging task. The dynamic nature of hate speech patterns presents a significant challenge in effectively addressing this issue. The crux of the issue concerns the requirement for a thorough and resilient natural language processing (NLP) solution that can effectively tackle the obstacles linked to hate speech and furnish a mechanism for reducing its detrimental effects.

The identification and precise categorization of hate speech is a challenging undertaking due to its dependence on nuanced linguistic indicators, situational factors, and the communicator's purpose. Conventional filtering systems that rely on keywords are inadequate for effectively identifying hate speech due to their inability to comprehend the contextual and nuanced aspects of the language used. Furthermore, hate speech undergoes a process of evolution over time, wherein novel terms, expressions, and covert manifestations frequently arise. The dynamic nature of evolving patterns poses a significant challenge in devising static rule-based strategies to counter hate speech effectively.

Comprehending the linguistic and purposeful aspects of hate speech is pivotal for efficacious moderation and intervention. Hate speech frequently directs discriminatory language towards particular groups on the basis of their race, ethnicity, religion, gender, sexual orientation, or other identifying traits. The use of language that is deemed offensive has the potential to provoke acts of violence, discrimination, and harm toward both individuals and communities. Consequently, it is crucial to devise a resolution that surpasses mere superficial recognition and integrates more profound linguistic scrutiny to apprehend the motives and possible ramifications of hate speech.

Due to the extensive quantity of content generated by users on digital platforms, relying solely on manual moderation is both impractical and inadequate. The implementation of an NLP-based automated solution can enhance scalability and efficiency in the detection and mitigation of hate speech. Through the utilization of computational linguistic methodologies, including natural language processing techniques, machine learning algorithms, and deep learning models, it is possible to create a proficient system for the examination, categorization, and surveillance of hate speech within digital communities.

The objective of this study is to investigate and suggest methodologies that utilize natural language processing and computational linguistic methods to accurately identify, categorize, and address instances of hate speech. The creation of sophisticated NLP models that can effectively capture the nuances of hate speech, contextualize it, and evaluate its potential for harm can equip online platforms with the essential resources to uphold a secure and all-encompassing atmosphere for their user base.

1.3 Research Objectives

The principal aim of this study is to tackle the difficulties related to hate speech in online environments through the utilization of computational linguistic methodologies and natural language processing techniques. The research objectives are outlined as follows:

- **Develop advanced NLP techniques:** One of the objectives of this study is to develop and enhance natural language processing models that can effectively detect instances of hate speech in digital media. The task at hand entails an examination of diverse methodologies, including supervised and unsupervised machine learning algorithms, deep learning architectures, and sentiment analysis techniques. The aim is to improve the accuracy of identifying hate speech by taking into account its complex and changing characteristics, with a focus on increasing both precision, recall and reducing bias towards targeted identity groups.

- **Analyze linguistic and socio-cultural factors:** This study aims to investigate the fundamental linguistic and sociocultural elements that give rise to hate speech and to enhance comprehension of the underlying factors and processes that contribute to the dissemination of hate speech by means of extensive linguistic analysis and examination of the socio-cultural environments in which it arises. The aforementioned analysis has the potential to provide valuable insights for the enhancement of detection and mitigation strategies.

- **Design effective mitigation strategies:** Drawing upon the knowledge acquired through linguistic and socio-cultural examination, the study endeavors to devise and execute efficacious tactics for alleviating the deleterious effects of hate speech within virtual communities. This encompasses the examination of various methodologies, such as counter-speech, user reputation systems, community-based moderation, and proactive intervention techniques. The aim is to devise interventions that can efficiently deter hate speech, foster constructive online communication, and establish a more secure milieu for all users.

This study seeks to contribute to the development of comprehensive and robust computational linguistic approaches for analyzing and mitigating hate speech in online communities by achieving these research objectives. The ultimate objective is to equip social media platforms and online communities with the necessary tools and knowledge to effectively combat hate speech and promote a more inclusive and respectful online environment.

2 Literature Review

One of the first attempts to deal with abusive language was made by Yin et al. (2009), in which the authors used a supervised classification technique along with n-grams, manually created regular expression patterns, and contextual characteristics that take the abusiveness of earlier phrases into account. Röttger et al. (2021) developed a set of carefully defined functional suites (HateCHECK) to test any hate speech classification model to pinpoint its limitations. A different work by the intersecting set of authors proposed a human-in-the-loop approach to generate datasets that fared better in HateCHECK Vidgen et al. (2021). Mishra et al. (2018a, 2019) involved the use of community-based profiling to rely on user and community information rather than solely relying on textual cues. Mishra et al. (2018b) iterates the challenges facing abusive content on social media platforms and how recurrent neural networks fail to address such challenges. Several works like Muti and Barrón-Cedeño (2022); Hartvigsen et al. (2022) have taken advantage of BERT-based models for downstream classification tasks in this domain.

Numerous datasets focusing on vital themes like anti-racism, feminism, anti-misogyny, and more were meticulously curated by scraping diverse online social media platforms such as Twitter and Reddit Waseem and Hovy (2016); Founta et al. (2018); Waseem (2016). Notably, a dataset annotated with instances of personal attacks, toxic messages, and aggression sourced from the English Wikipedia's Talk pages was published by Wulczyn et al. (2017). These valuable datasets were harnessed to conduct comprehensive analyses on pivotal factors such as hate speech, toxicity, and sentiments surrounding these subjects. Furthermore, they shed light on the overall sentiment prevailing within these platform communities. It is worth mentioning that the models trained on these datasets exhibit limited effectiveness when it comes to capturing implicit expressions of hatred. To address this gap, ElSherief et al. (2021) curated a distinct dataset comprising tweets from extremist groups, meticulously labeled to identify implicit forms of toxicity. In addition to such manually scraped data, adversarial text generation and human-in-the-loop methods have been effectively used to create balanced datasets that cater to the complex nature of hatespeech Vidgen et al. (2021); Hartvigsen et al. (2022). Bianchi et al. (2022) takes a holistic approach to hate speech by arguing for a fine-grained multi-label approach that considers different aspects of incivility and hateful or intolerant content.

Researchers have found time and again that machine learning models are subject to bias. In the case of hate speech and abuse detection the effects are pronounced and has serious effect on the practicality. Numerous studies have been conducted on dialects such as African American English and how linguistic markers are subjective and can wrongly contribute to the models output Sap et al. (2019); Davidson et al. (2019); Xia et al. (2020). As a result, researchers have shown keen interest on developing metrics Dixon et al. (2018); Borkan et al. (2019); Röttger et al. (2021) and methods Vaidya et al. (2020); Barikeri et al. (2021) to evaluate and mitigate bias towards certain targeted identities.

Another crucial area of research in the realm of hate speech revolves around the examination of hate, abuse, and profanity dynamics within public environments. This sub-field seeks to comprehend the mechanisms through which toxicity is propagated Mondal et al. (2017), the establishment of societal norms Rajadesingan et al. (2020), and the utilization of author profiling to gain insights into a particular user's intentions regarding hate speech Mishra et al. (2018a). Numerous studies have shed light on pressing issues such as political discourse Solovev and Pröllochs (2022), incidents of Asian hate He et al. (2021); Tahmasbi et al. (2021), and the persistence of anti-Semitism Arviv et al. (2021). Additionally, there is a growing focus on investigating contrasting phenomena like fear speech Saha et al. (2023) and hope speech Palakodety et al. (2020); Chakravarthi (2020). These subjects are being studied in greater depth to better comprehend the multifaceted nature of hate and toxicity. It is important to acknowledge that hate and toxicity are inherently subjective and vary from one individual to another. Exploring the influence of personal characteristics and content shared among social media users provides valuable insights into the perception of hate speech Schmid et al. (2022).

3 Methodology

3.1 Data Collection

In order to investigate and mitigate hate speech in online communities using Natural Language Processing (NLP), an essential aspect of this research is the collection of relevant data. Numerous data collection strategies have been employed in existing related work, offering valuable insights into the nature and extent of hate speech in various online platforms. This subsection outlines the different data collection approaches that have been utilized and their relevance to the study of hate speech in online social media.

Keywords-based data scraping has been a widely employed technique in the field of hate speech analysis. By identifying specific keywords or phrases associated with hate speech, researchers have been scraping data from online platforms such as Twitter, Reddit, 4Chan, Gab, etc. These platforms have been predominantly chosen due to their popularity and the extensive user-generated content they host. Consequently, they provide rich datasets that can be mined for valuable information on hate speech patterns, prevalent hate speech topics, and the dynamics of hate speech propagation within online communities.

Furthermore, data collection efforts have extended beyond social media platforms to include platforms such as Wikipedia and other similar resources. These platforms offer a unique perspective on the presence of stereotypes and gender bias in online content. By analyzing articles, discussions, and edits on Wikipedia, researchers have gained insights into the representation of different genders and the perpetuation of biased narratives. This approach helps shed light on the underlying factors contributing to hate speech and can inform the development of more inclusive and unbiased language models.

Moreover, adversarial text generation techniques can be employed as a vital tool in reducing bias towards targeted identities and enhancing the capabilities of hate speech detection models. Adversarial text generation involves creating examples that challenge the model's ability to correctly identify hate speech. By generating such adversarial examples, researchers can identify potential weaknesses and biases in existing models, thus enabling the development of more robust and accurate hate speech detection systems. This approach contributes to the mitigation of false positives and negatives, which are crucial for effectively addressing hate speech in online communities.

In this proposed research, the data collection process will involve a combination of keywords-based data scraping from platforms such as Twitter, Reddit, and 4Chan to study toxicity, stereotypes, and gender bias. Additionally, the implementation of adversarial text generation techniques will be explored to enhance hate speech detection models. The utilization of these diverse data sources and methodologies will provide a comprehensive and multifaceted understanding of hate speech in online communities, enabling the development of practical computational linguistic approaches to analyze and mitigate hate speech using NLP.

3.2 Hate Speech Detection

Detecting and mitigating hate speech in online communities using NLP approaches requires a robust methodology incorporating various techniques and models. This subsection outlines the key components and considerations involved in hate speech detection, highlighting the advantages of different approaches and the proposed methodology for this research.

Various approaches have been employed in hate speech detection, ranging from rule-based mechanisms to state-of-the-art NLP techniques. Rule-based mechanisms rely on predefined patterns or linguistic rules to identify and classify hate speech. While these approaches are relatively straightforward to implement, they may lack the flexibility to handle complex and evolving forms of hate speech. On the other hand, state-of-the-art NLP techniques, such as deep learning models, leverage the power of machine learning to automatically learn representations and patterns from data, offering the potential for more accurate and adaptable hate speech detection.

Preprocessing and feature engineering methods play a crucial role in transforming raw text into suitable input representations for hate speech detection models. These methods involve tasks such as tokenization,

stop-word removal, and vectorization, which allow the models to capture relevant linguistic information. By effectively preprocessing the text and engineering informative features, the models can better distinguish between hate speech and non-hate speech content.

Machine learning models, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based architectures, have demonstrated success in hate speech detection tasks. RNNs are effective at capturing sequential dependencies in text, while CNNs excels at capturing local patterns and textual features. Transformer-based architectures, such as the popular BERT model, have shown impressive performance in a variety of NLP tasks, including hate speech detection. These models can effectively learn contextual representations and leverage the power of self-attention mechanisms to capture important linguistic information.

Evaluation metrics are essential for assessing the performance and generalizability of hate speech detection models. Commonly used metrics include precision, recall, and F1 score, which measure the model's accuracy, completeness, and balance between the two. However, it is also important to consider newer metrics and functional test suites that evaluate a model's ability to perform bias-free classification and avoid perpetuating harmful stereotypes.

In addition to focusing solely on text-based analysis, a holistic approach will be undertaken in this proposed research. This approach involves considering the wider context surrounding hate speech, including the entire public discourse, user profiles, and the societal environment. By incorporating contextual factors, such as the influence of user interactions and the impact of societal events, a more comprehensive understanding of hate speech in online communities can be achieved.

3.3 Linguistic and Socio-Cultural Analysis

Gaining insights into the linguistic and socio-cultural factors influencing hate speech in online communities is crucial for the development of effective computational linguistic approaches to analyze and mitigate this issue. This subsection presents the planned analysis to understand linguistic and socio-cultural influences, including different dialects and regional variations, as well as the methods for identifying linguistic markers, context-based features, biases, and the incorporation of socio-cultural variables.

This part of the research aims to investigate linguistic markers and features that differentiate hate speech across different dialects and regional contexts, including variants such as African American English, low-resource languages, etc. Analyzing the characteristics of hate speech within specific linguistic communities will provide insights into the diverse dimensions of hate speech.

The analysis will employ various methods to identify linguistic markers and context-based features in hate speech data. Natural Language Processing techniques, including part-of-speech tagging, named entity recognition, and sentiment analysis, will be used to extract linguistic features associated with hate speech. These features will help identify patterns, offensive terms, and discriminatory language commonly found in hate speech. Additionally, contextual features, such as the presence of slurs, offensive references, or derogatory comments within specific contexts, will be examined to uncover the nuances of hate speech usage.

Incorporating socio-cultural variables is essential for a deeper understanding of hate speech dynamics in online communities. Demographic information, such as age, gender, and geographical location, will be considered to identify potential biases and demographic-specific patterns in hate speech. Analyzing social media dynamics, including user interactions, community structures, and influence networks, will provide insights into the spread and amplification of hate speech within online platforms. Examining user interactions and the dissemination of hate speech will contribute to understanding the social dynamics that fuel the persistence and propagation of hate speech.

Moreover, the analysis will focus on studying influential events, such as political posts, social phenomena, or rights rallies, to gain information on the cultural aspects influencing hate speech. By analyzing hate speech data surrounding such events, this research will identify patterns, themes, and cultural triggers contributing

to the occurrence and intensity of hate speech. Understanding cultural aspects will inform the development of targeted interventions and mitigation strategies that address the underlying socio-cultural factors driving hate speech in online communities.

4 Conclusion

This proposal presents a comprehensive research plan that addresses the urgent need for analyzing and mitigating hate speech in online social media platforms using computational linguistic approaches and Natural Language Processing techniques. By leveraging advanced NLP models, the proposed research aims to enhance hate speech detection accuracy and efficiency, enabling more proactive interventions and fostering inclusive online environments.

Through the development of advanced NLP techniques, I seek to improve the precision and recall of hate speech detection algorithms, thereby reducing false negatives and false positives while also concentrating on reducing bias. By leveraging large-scale datasets and employing robust annotation protocols, I will train and validate the models to ensure their effectiveness across diverse platforms and contexts. This will empower platform administrators and moderators with the tools and insights needed to swiftly identify and address instances of hate speech, promoting a safer and more respectful online community.

Furthermore, the research will delve into the linguistic and socio-cultural factors that contribute to hate speech. By analyzing linguistic patterns, syntactic structures, semantic associations, and socio-cultural contexts, I aim to gain a deeper understanding of the underlying motivations and dynamics of hate speech. This multifaceted analysis will enable us to develop more contextually aware and sensitive approaches to hate speech mitigation, addressing not only the explicit expressions of hate but also the underlying root causes. By considering socio-cultural factors such as identity, group dynamics, and power structures, our research will contribute to the development of interventions that effectively counter hate speech and promote inclusivity.

In conclusion, the proposed research holds great promise for advancing the field of computational linguistics and combating the pervasive issue of hate speech on online social media platforms. By combining sophisticated NLP techniques with comprehensive linguistic and socio-cultural analyses, I aim to create a significant impact in the detection and mitigation of hate speech. Ultimately, the research endeavors to foster a digital landscape that is characterized by respect, inclusivity, and constructive dialogue for all users.

References

E. Arviv, S. Hanouna, and O. Tsur. It'sa thin line between love and hate: Using the echo in modeling dynamics of racist online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 61–70, 2021.

S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, 2021.

F. Bianchi, S. A. Hills, P. Rossini, D. Hovy, R. Tromble, and N. Tintarev. " it's not just hate": A multi-dimensional perspective on detecting harmful speech online. *arXiv preprint arXiv:2210.15870*, 2022.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

B. R. Chakravarthi. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, 2020.

T. Davidson, D. Bhattacharya, and I. Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019.

O. De Gibert, N. Perez, A. García-Pablos, and M. Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*, 2021.

Facebook. Report inappropriate or abusive things on facebook (e.g. nudity, hate speech, threats). https://www.facebook.com/help/135402139904490, 2023. [Online; accessed 06-Feb-2023].

A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, 2022.

B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94, 2021.

P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova. Author profiling for abuse detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1088–1098. Association for Computational Linguistics (ACL), 2018a.

P. Mishra, H. Yannakoudakis, and E. Shutova. Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10, 2018b.

P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova. Abusive language detection with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2145–2150, 2019.

M. Mondal, L. A. Silva, and F. Benevenuto. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 85–94, 2017.

A. Muti and A. Barrón-Cedeño. A checkpoint on multilingual misogyny identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 454–460, 2022.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

S. Palakodety, A. R. KhudaBukhsh, and J. G. Carbonell. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 454–462, 2020.

A. Rajadesingan, P. Resnick, and C. Budak. Quick, community-specific learning: How distinctive toxicity

norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568, 2020.

P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J. Pierrehumbert, et al. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 41. Association for Computational Linguistics, 2021.

P. Saha, K. Garimella, N. K. Kalyan, S. K. Pandey, P. M. Meher, B. Mathew, and A. Mukherjee. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120, 2023.

M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019.

U. K. Schmid, A. S. Kümpel, and D. Rieger. How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, page 14614448221091185, 2022.

K. Solovev and N. Pröllochs. Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In *Proceedings of the ACM Web Conference 2022*, pages 3656–3661, 2022.

F. Tahmasbi, L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou. "go eat a bat, chang!": On the emergence of sinophobic behavior on web communities in the face of covid-19. In *Proceedings of the web conference 2021*, pages 1122–1133, 2021.

Twitter. Hateful conduct. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, 2023. [Online; accessed 06-Feb-2023].

A. Vaidya, F. Mai, and Y. Ning. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693, 2020.

B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, 2021.

Z. Waseem. 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*, pages 138–142, 2016.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

M. Xia, A. Field, and Y. Tsvetkov. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*, 2020.

D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2(0):1–7, 2009.

Youtube. Hate speech policy. https://support.google.com/youtube/answer/2801939, 2023. [Online; accessed 06-Feb-2023].