Directions for NLP Practices Applied to Online Hate Speech Detection

Paula Fortuna, Mónica Domínguez
NLP Group
Pompeu Fabra University
Barcelona, Spain
first.last@upf.edu

Leo Wanner
ICREA and
Pompeu Fabra University
Barcelona, Spain
leo.wanner@upf.edu

Zeerak Talat
Simon Fraser University
Burnaby, Vancouver
zeerak_talat@sfu.ca

Abstract

Addressing hate speech in online spaces has been conceptualized as a classification task that uses Natural Language Processing (NLP) techniques. Through this conceptualization, the hate speech detection task has relied on common conventions and practices from NLP. For instance, inter-annotator agreement is conceptualized as a way to measure dataset quality and certain metrics and benchmarks are used to assure model generalization. However, hate speech is a deeply complex and situated concept that eludes such static and disembodied practices. In this position paper, we critically reflect on these methodologies for hate speech detection, we argue that many conventions in NLP are poorly suited for the problem and encourage researchers to develop methods that are more appropriate for the task.

1 Introduction

Online hate speech is the cause for growing concern, due to its social impacts (Soral et al., 2018). In response, automatic detection of hate speech has become a popular research area in Natural Language Processing (NLP), with a substantial number of papers at leading conferences, targeted workshops and shared tasks dedicated to it. Simultaneously, content moderation infrastructures that rely on machine learning technologies have been proposed to address the propagation of online harms.

The increase in research interest to hate speech detection has spurred on a growth and variety in annotated resources for the task created within the academy and industry. However, at the same time, critical work on hate speech detection has found that there are significant challenges related to the published research outcome with respect to the construction of data (Vidgen and Derczynski, 2021; Fortuna et al., 2020), model generalizability (Nozza, 2021; Fortuna et al., 2021), and socially biased effects of models (Talat et al., 2018; Davidson et al., 2019). These challenges indicate that despite a techno-optimist attitude from the NLP community to the task (Talat et al., 2021), the acceleration in research on the topic has impeded a maturation of the basic conventions, principles and practices which this research is based on. Hence, it seems timely to assess whether the NLP conventions and practices that have been followed so far are indeed appropriate for the classification of hate speech.

In contrast to prior work, which has sought to address specific shortcomings of machine learning algorithms for hate speech detection (e.g., Dixon et al., 2018), our work presents a comprehensive analysis of the inadequacies of some of NLP's methodological conventions for algorithmic hate speech detection. That is, in this position paper we reflect on contemporary methodological challenges that arise from applying common NLP methodologies to hate speech detection.

More specifically, we a) reflect on frequently used conventions for classification of text and explore their inadequacies with reference to hate speech detection; and b) discuss the future of current methodologies for this task. On the basis of our analysis, we conclude that current models are incapable of detecting hate speech without harms to marginalized communities. We therefore call for the scientific community to adapt NLP methodologies such that future developments center the impacts that used methodologies may have on marginalized communities. We believe that by critically reflecting on the potential real-world impacts of the methodologies for hate speech detection on marginalized communities, the scientific community can come to identify methodologies that result in more just futures.

2 Reflections On Hate Speech Detection

Hate speech detection is commonly conceptualized as a supervised classification task, with the goal to determine whether content is hateful or not (Yin

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11794–11805
December 7-11, 2022

and Zubiaga, 2021). This setting requires that we i) define the problem; ii) collect, sample, and annotate data to obtain labels;[1] and iii) apply optimization technologies (i.e. machine learning algorithms) to the labeled data. Finally, the resulting methods and models are evaluated using specific metrics and techniques. In this section, we review these stages through the lens of hate speech detection.

2.1 Definitional Challenges of Hate Speech

Defining hate speech is to control the discourse surrounding the phenomena; determine which groups are minoritized, and therefore should be protected; and which patterns of speech should be sanctioned (Gelber, 2021). Coining a definition of hate speech is therefore a political task, in particular due to the implication that each of the many available possible definitions carry (Thylstrup and Talat, 2020). While definitions are subject to the cultural norms of the geography in which they are created (Talat et al., 2022), universalist assumptions surrounding established definitions of hate speech in NLP fail to account for the diversity required for the task. Such universalist assumptions allow for hate speech detection infrastructures as "a third layer of interpretation" between the sender and recipient of a message (Thylstrup and Talat, 2020).

Moreover, hate speech is often categorized under the umbrella terms such as "abusive language", "offensive language", or "toxicity" (Poletto et al., 2021; Jigsaw, 2019), resulting in a concept drift, where hate speech cedes prominence to the more generic concepts, such as generally offensive language.[2] As a result, models are prone to learn patterns that emphasize the more frequently occurring categories (e.g. 'insult') and under-perform on hate speech (Fortuna et al., 2020).

Furthermore, most NLP research exclusively considers textual material, assuming that it provides adequate information. However, hate speech is deeply tied to oppression and it is therefore necessary to understand the speaker and listener's subjectivities to situate the text and adjudicate whether it constitutes hate speech. More often than not, this information is unavailable from the text.

[1]Labels can be understood as facts, for machine learning models as these are unable to verify or contest the veracity of the labels (Talat et al., 2021).

[2]Although "toxicity" and "offensive language" are used as umbrella terms, they are also used as terms for specific types of abuse within NLP.

2.2 On Challenges of Hate Speech Annotation

A common convention in labeling a dataset is to use an odd number of annotations for each text sample. The reliability of the labels in a dataset is often measured by computing the Inter-Annotator Agreement (IAA). In this section, we discuss biases in annotations and the paradoxical search for ground truth within disagreement.

2.2.1 Annotation Bias

Socially biased systems are a growing concern within NLP (Blodgett et al., 2020; Talat et al., 2022). Social biases are particularly apparent in hate speech datasets (and models) as biases are a reflection of wider social tension. Data source selection and sampling strategies can also be contributors to the social biases found in datasets and models. For instance, data samples often skew towards particular perpetrators or keywords (e.g., slurs), resulting in datasets that favor explicit abuse and hate speech (e.g. Davidson et al., 2017; Basile et al., 2019; Founta et al., 2018). Collected data samples are then annotated through the lens of the selected definitions of hate speech, and the IAA is computed. Once satisfactory levels of agreement have been obtained, the "ground-truth" for each text is selected (Pustejovsky and Stubbs, 2012), often by selecting the label chosen by the majority of annotators. Thus, labels embed the subjectivities of the annotators. For instance, if three annotators agree that "cats are better than dogs", the agreement reflects the annotators' subjective preferences, in spite of complete agreement (i.e. $IAA = 1.0$), rather than the inherent value of cats or dogs.

The annotation challenge is further aggravated by the absence of widely agreed-upon annotation criteria (Vidgen et al., 2019a), resulting in unclear categories. For instance, the term "abusive" has been described in terms of the speakers' intent to injure (Pitsilis et al., 2018) and in terms of the assumed impact on the reader (Wulczyn et al., 2017). Further, as annotators most often cannot communicate with the authors or the targeted subjects, annotators must make assumptions with respect to the intention of the authors of a text and its impact on readers – as, e.g., in the case of the above reclaimed slurs (Sap et al., 2019).

The selection of annotators is another source of social biases (Talat et al., 2021). Annotators are often recruited from crowd-working platforms, with little regard to their subject matter expertise or so-

cial and cultural background. However, annotators' subjectivities, expertise (Waseem, 2016), attitudes and beliefs (Sap et al., 2021), and their diversity and variability (Hovy and Prabhumoye, 2021) have been shown to influence annotation results in spite of training and exposure to annotation guidelines.

2.2.2 On Ground Truth and Agreement

The goal for annotation efforts in NLP is to assign a gold label to data (e.g., a document or an entity therein) (Zeinert et al., 2021). The search for a single label in the face of disagreement is based on the assumption that there exists a single *correct* label which can be approximated using agreement aggregation methods. IAA is used as a proxy for the quality, i.e., correctness, of obtained labels. In the context of hate speech, IAA is often very low (Vigna et al., 2017; Olteanu et al., 2018; Poletto et al., 2019). However, researchers often rely on a single label as ground truth, disregarding the absence of agreement, variability, and subjectivity of the obtained *ground truth* (Paullada et al., 2021). The result is that, paradoxically, researchers construct ground truth for inherently subjective questions on the basis of disagreement.

2.3 Model Learning and Evaluation

Once a dataset has been labeled, models can be trained and evaluated. To assess model performance and generalizability, the trained models are evaluated on held-out test sets (Chollet and Allaire, 2018; Pustejovsky and Stubbs, 2012). This paradigm of evaluation assumes that training data and data encountered when a model is deployed are independent and identically distributed (I.I.D.) (Arlot and Celisse, 2010). For hate speech, the I.I.D. assumption means that the annotation of a text is independent of earlier annotations of other texts, and that the data sampled from outside of the dataset will follow the same class distribution that is evident in the labeled dataset. Below, we detail the limitations that can arise from the assumptions made in for model evaluation and the risks from drawing conclusions from them.

2.3.1 On Model Understanding

Although contemporary machine learning models often show an impressive performance when applied to different NLP tasks, they have been criticized for failing to grasp pragmatics due to their reliance on the distributional hypothesis (Bender and Koller, 2020). This is particularly concerning for addressing hate speech as it operates at the linguistic level of pragmatics. It is therefore important to understand how machine learning models make judgments on whether texts are hateful. Deeper assessments of hate speech models suggest that they have a very superficial understanding of language (see appendix A). In fact, prior work has argued that the reported performances for hate speech detection are in part influenced by spurious correlations (e.g. Rahman et al., 2021; Wiegand et al., 2019), and overlapping data in the train and test sets (Arango et al., 2019). This work has shown that correcting of these issues results in a decrease of performance. If models are over-fitting to spurious correlations and are incapable of language understanding, the question arises whether we can rely on current classifiers for robustly detecting hate speech.

2.3.2 On Interpreting Model Performance

Recent work on quantitative benchmarks has questioned the ability of contemporary methods to measure generalization in machine learning (Raji et al., 2021; Paullada et al., 2021), and for hate speech (Röttger et al., 2021). Researchers have found that well-performing systems for hate speech detection are susceptible to minor adversarial modifications of the input text that significantly alter the meaning (Gröndahl et al., 2018). That is, solely relying on quantitative benchmark results can produce an incomplete picture of the performance of evaluated models, leading researchers to over-estimate the performances of systems that are brittle in nature.

2.3.3 On Model Generalization

One reason models may be vulnerable to adversarial attacks is that they over-fit to tokens and token interaction patterns instead of learning to generalize the concept of hate speech (Oliva et al., 2020; Sarkar and KhudaBukhsh, 2021; Oak, 2019; Fortuna et al., 2021). Although this issue has been identified by prior work, we reconsider it in light of the limitations we have discussed thus far.

We identified the following three factors that make the I.I.D. assumption unlikely to hold, resulting in models that are unlikely to generalize: 1) Given the variety of concepts and definitions in hate speech, it is very hard to assure that the different samples express the same flavor of the phenomena; 2) due to the fact that hate speech only occurs very rarely in random samples (Vidgen et al., 2019b), sampling strategies tend to over-emphasize domains where hate speech is more likely to occur,

making it unlikely that two independent enterprises in dataset creation will result in complementary datasets; and 3) the speed at which linguistic shift happens in online social media (Hogan and Quan-Haase, 2010) makes it unclear if time-bounded data collections from social media can be I.I.D. This also raises an open question for the development of machine learning models for hate speech detection: How can we identify when it becomes necessary to train new models to keep abreast with the changes that have occurred in language use since the training data was sampled?

3 Discussion: On the Present and Future for Hate Speech Detection

In the preceding sections, we have highlighted a number of challenges and ethical concerns that arise from the current conceptualization of hate speech detection. We argue that these limitations render current models unable to detect hate speech without significant risk to minorities. Specifically, classifiers that are unable to accurately classify content directed towards marginalized communities risk increasing the costs for said communities to participate in online spaces, due to the increased risks of being subject to hate speech whilst also remaining unprotected by hate speech detection systems (Oliva et al., 2020). Thus, contemporary systems for hate speech detection risk reproducing normative values and attitudes towards acceptable language use while further entrenching marginalization in online spaces (Thylstrup and Talat, 2020).

Overcoming the identified challenges will require shifting our research practices. In this section, we propose new directions for hate speech detection. However, we do not expect that implementing any individual solution in isolation will result in ready to use classifiers. We therefore emphasize the need for research to continuously reassess the risks that arise from methodological innovations for hate speech detection.

Accounting for Plurality of Hate Speech While contemporary methods for annotating hate speech imply the assumption that there is a universal definition of hate speech, and that models derived from labeled data are applicable across all contexts, we argue instead for a pluralist approach to annotation. By taking a pluralist approach, e.g. through situating models within subjective contexts, researchers are afforded the ability to view hate speech as contextual to the subjectivities of the target of hate. For instance, by narrowing down definitions of hate, clearly providing the geographical and cultural contexts, and specifying the values and goals for the model, researchers can clearly articulate within which contexts models and data are valid and which particular groups models seek to protect. Such *model framing* can help address the issues surrounding universality and can provide space for researchers to consider how their choices have political implications for what speech is sanctioned.

Accounting for Context Supervised machine learning models for hate speech primarily operate on text, and a single label for each document during training. However, whether a text amounts to hate speech is highly context dependent (Talat et al., 2018). For instance, whether a word is used as a slur or as a reclaimed term depends on the identity of the speaker, the phrasal and social contexts in which it is uttered. The primary means of approximating *conversational context* in prior work has been through the use of conversation threads and user metadata (see appendix B.1). Such conversational contexts only account for a small number of contexts that are invoked during the utterance and annotation. Annotators, for example, may hold prior knowledge on the histories, social hierarchies, conflicts, or stereotypes concerning the groups addressed in a document. Hate speech detection research would therefore benefit from considering methods to explicitly incorporate such information into the modeling pipeline.

Representative Sampling Procedures Given the sampling methods used to ensure an adequate distribution of hate speech for labeling, models are often trained on data distributions that significantly vary from real-world occurrences of hate speech. To address this concern, future data collection efforts should seek to minimize such distributional differences whilst taking into account wider notions of conversational contexts.[3]

Representative Annotation Processes The common NLP practices of having a small number of annotations for each document (often just three annotations per document) that are used to compute IAA and perform label aggregation erases different opinions on the label of a document. In this way, the use of aggregation methods for label voting exerts direct control over subjective positions on the

[3] See appendix B.2 for a discussion on prior work addressing keyword biases in data collection.

labels, and thereby the discourses that the model will come to replicate. However, the interpretation of what constitutes hate is a highly subjective question and is subject to the workers' individual subjectivities (Waseem, 2016). While subjectivities are inherent, we propose that researchers use scales (DeVellis and Thorpe, 2021) to measure different dimensions of hate speech. Scales have previously been used in the social sciences for asking subjective questions, and could provide new possibilities for hate speech research. In particular, the use of scales can allow for framing models in terms of the values that they embody.[4]

Evaluation of Hate Speech Models In the last few years, research on hate speech detection has shown increases in performance across a number of metrics. However, as we have argued, despite quantitative improvements such performance increases do not reveal a full picture of model performances. In fact, contemporary models only display a superficial understand of hate speech (see appendix A for an analysis of Vidgen et al. (2021)). It is therefore necessary for research on hate speech to consider new evaluation paradigms and metrics. Such initiatives must center models' abilities to generalize beyond identifying frequently occurring tokens. For instance, an emphasis on evaluating models for over-fitting to particular tokens (see appendix B.4) can provide a greater understanding of model generalizability. Another promising direction is the creation of test suites that target potential areas of concern for models for detecting hate (e.g. Röttger et al., 2021). Another avenue for improved evaluation is to directly leverage training data to create hard-to-pass tests for machine learning models. While such evaluation practices are a step in the right direction, they do not address the question of context. It is therefore necessary to develop evaluation practices that seek to evaluate models in the contexts that models are developed for, with the aid of methods capable to identify if new data is still I.I.D. to the data used for training.

Handling Classification Errors For many NLP tasks classification errors do not have immediate harms. For hate speech detection, classification errors can result in significant immediate harms to people. False negatives can result in hateful speech being passed as acceptable which can allow harmful content to remain unsanctioned (Oliva et al., 2020). While false positives can result in inoffensive speech being sanctioned. Given content moderation's commitment to dominant social norms (Thylstrup and Talat, 2020), classification errors often disproportionately affect marginalized communities, i.e. white supremacist content remains unsanctioned while content from marginalized communities is removed (Davidson et al., 2019; Oliva et al., 2020). In light of these concerns, it is prudent for the NLP community and legislators to reflect on the ramifications of classification errors. For instance, we encourage both communities to reflect on whether it is appropriate to deploy models that will produce racialized and gendered classification errors, which entity is to be held to account for such errors, and how victims of automated classification errors should be compensated.[5]

4 Conclusion

In this paper, we have argued that current NLP practices for hate speech detection are unlikely to address the core concerns of hate speech detection, i.e. identify hate with minimal errors and protect marginalized communities. We therefore call for the NLP community to rethink its methodologies such that future developments reduce risk for marginalized communities.

One avenue for future work is to follow the principles of design justice (Costanza-Chock, 2018), which emphasizes community inclusion and ownership of (technological) solutions. Following the principles of design justice, researchers would decenter their own expertise in favor of the lived expertise of affected communities. We strongly believe that future steps must center a multidisciplinary approach in close communication with affected groups. By taking steps to document and address the limitations of contemporary methods for hate speech, researchers can identify new avenues for improving hate speech detection models. Moreover, researchers can take steps towards ensuring that content moderation technologies provide safer online spaces for marginalized communities by mitigating the prevalence of online hate speech.

[4] Another approach is that of perspective-aware processes (Akhtar et al., 2021). Perspective aware modeling seeks to collate "community" annotation to provide "community-based" labeling to be used for training machine learning models; see appendix B.3 for an in-depth discussion on perspective-aware modeling.

[5] We acknowledge that legislation is currently being developed on artificial intelligence in the European Union, United States of America, and Canada.

Limitations

This study is intended as a theoretical consideration of the issues that arise in hate speech detection. The study analyzes the current limits of using machine learning infrastructures for the identification and moderation of hate speech. One limitation of the work is the theoretical frame of our work. While our frame allows for more deeply understanding the issues of contemporary methods for hate speech detection, deeper considerations of sociological and anthropological methods can afford significant improvements in our understanding of NLP technologies, such as hate speech detection, as socio-technical systems and their social impacts. A further limitation of our work is its focus on research rather than application, and therefore does not discuss how classification models are used in real-world content moderation applications. This is left to future work.

Acknowledgements

We thank the reviewers for their insightful comments. The first author is supported by the research grant SFRH/BD/143623/2019, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program *Human Capital* (POCH), supported by the European Social Fund and by national funds from MCTES.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *CoRR*, abs/2106.15896.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 45–54. ACM.

Sylvain Arlot and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none):40 – 79.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. *CoRR*, abs/2203.09192.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

François Chollet and J.J. Allaire. 2018. *Deep learning with R*. Manning Publications, New York.

Sasha Costanza-Chock. 2018. Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2021. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *arXiv:2110.05719 [cs]*. ArXiv: 2110.05719.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

Robert F DeVellis and Carolyn T Thorpe. 2021. *Scale development: Theory and applications*. Sage publications.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA. ACM.

Paula Fortuna, Juan Soler Company, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6786–6794. European Language Resources Association.

Paula Fortuna, Juan Soler Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Inf. Process. Manag.*, 58(3):102524.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 260–266. INCOMA Ltd.

Katharine Gelber. 2021. Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, 24(4):393–414.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, page 2–12, New York, NY, USA. Association for Computing Machinery.

Bernie Hogan and Anabel Quan-Haase. 2010. Persistence and change in social media. *Bulletin of Science, Technology & Society*, 30(5):309–315.

Eduard H. Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Lang. Linguistics Compass*, 15(8).

Jigsaw. 2019. Perspective api. Available in https://github.com/conversationai/perspectiveapi, accessed last time in November 2019.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security, SOUPS 2021, August 8-10, 2021*, pages 299–318. USENIX Association.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10528–10539. Association for Computational Linguistics.

Frederick Liu and Besim Avci. 2019. Incorporating Priors with Feature Attribution on Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy. Association for Computational Linguistics.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech 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 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Rajvardhan Oak. 2019. Poster: Adversarial examples for hate speech classifiers. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 2621–2623. ACM.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2020. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, pages 1–33.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 221–230. AAAI Press.

Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. 2020. Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2532–2542, Online. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *CoRR*, abs/1801.04433.

Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2019. Annotating hate speech: Three schemes at comparison. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Md. Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mücahid Kutlu, and Matthew Lease. 2021. An information retrieval approach to building datasets for hate speech detection. *CoRR*, abs/2106.09775.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. *arXiv:2111.15366 [cs]*. ArXiv: 2111.15366.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. 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)*, pages 41–58, Online. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *CoRR*, abs/2111.07997.

Rupak Sarkar and Ashiqur R. KhudaBukhsh. 2021. Are chess discussions racist? an adversarial hate speech data set (student abstract). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 15881–15882. AAAI Press.

Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, 44(2):136–146.

Zeerak Talat, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. ArXiv: 2101.11974.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience episode #5 – workshop on challenges & perspectives in creating large language models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online Harassment*, pages 29–55. Springer.

Nanna Thylstrup and Zeerak Talat. 2020. Detecting 'Dirt' and 'Toxicity': Rethinking Content Moderation as Pollution Behaviour. *SSRN Electronic Journal*.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019a. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019b. *How much online abuse is there? A systematic review of evidence for the UK*. The Alan Turing Institute, London. Backup Publisher: The Alan Turing Institute.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. 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, Online. Association for Computational Linguistics.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017*, volume 1816 of *CEUR Workshop Proceedings*, pages 86–95. CEUR-WS.org.

Zeerak Waseem. 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*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1391–1399. ACM.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Comput. Sci.*, 7:e598.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. 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 3181–3197, Online. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

A Hate Speech Dynabench Model Classification Results

In this section, we provide examples of sentences classified by a state of the art model for hate speech detection (i.e. the Dynabench classifier proposed by Vidgen et al. (2021)) to illustrate challenges for state of the art models. This analysis is neither quantitative nor exhaustive, however it seeks to illustrate core issues for technical systems for the classification of hate speech. Quantitative analysis of mis-classifications have a set of concerns (see Section 2.3.2 for more detail). We submitted adversarial cases to the model interface[6], and find that model fails to predict correctly for simple adversarial examples (see table 1).

We craft our adversarial examples on the basis of western notions of sexism, starting with a simple example that we expect the model to correctly classify: "Women's place is in the kitchen". We then introduce minor variations to this sentence in order to gauge the model's understanding of sexism (see table 1 for a full list of our examples). For instance, we replace the word "women" with "men", "oven", and "gender". For these three cases, all examples except for "oven" are predicted as hate speech. While the model correctly identifies that an oven does in fact belong in a kitchen, it's gender-invariance for predicting hate speech reveals that the model has not learned the relationship between sexism and gender roles. That is, the model has not learned the social factors which are crucial to delineating hate speech from inoffensive speech.

When we keep the word "women" static and change the remaining sentence to something inoffensive, we see that the model performs better which indicates a correct distinction between minority terms and hate speech. We continue with new examples and introduce replacements to the word "kitchen". When replaced by "school", the model incorrectly predicts that it is hateful. This further evidences the model's inability to distinguish prescriptions of actions from hate. More concerning is the model's inability to correctly relate the implication of femicide, when evaluated on the text "Women's place is in the cemetery." While this further provides further evidence that the model does not learn a latent understanding of power dynamics, it also illustrates that the model may not provide adequate protections against violent speech towards women.

In our second set of examples (examples 2.1-2.3), we examine how the model responds more broadly to conversations around power dynamics. We see here that the explicit mentioning of gender and race prompts incorrect predictions from

[6]We experiment with Round 7 model on https://dynabench.org

the model, i.e., that the mere mention of comparative privilege is labeled as hateful. Should this model be deployed, it would actively limit conversations around race, gender, and power dynamics more broadly. Such conversations are frequently had by communities that are marginalized, in efforts to identify, discuss, and seek to remedy their own marginalization. That is, the model would censor conversations that are necessary to have, in order for society to progress beyond contemporary forms of marginalization, thereby actively limiting movements for social progress.

In our final three examples, we see that the model makes incorrect predictions for all three cases. In the latter two cases, we see further evidence that the model does not link notions of sexism and fascism with their expressed goals of marginalization.

B Improvements for Hate Speech Detection

We acknowledge that the NLP community is working towards identifying shortcomings of the current research practices, e.g., by studying how to learn from disagreements (Davani et al., 2021) or different perspectives during annotation (Leonardelli et al., 2021; Akhtar et al., 2021; Kumar et al., 2021). Here, we provide a brief summary of these efforts, which can also serve as a source of ideas for future approaches to the problem.

B.1 Improving Hate Speech Modeling

To counteract the lack of contextual information, the latest developments have added information to single text samples, including the conversation threads (Gao and Huang, 2017), network data and user information (e.g. Mosca et al., 2021). Adding conversational context, is a positive step forward as it adds contextual information which can help to frame the message being predicted. However the inclusion of more context may not be sufficient for addressing the participation of models in processes of marginalization. As have argued in this paper, we believe that considering methods from design, which aim to center and give definitional power to affected populations in the design and development processes, can be a productive path forward.

B.2 Handling Search with Keywords

Some prior work has sought to evaluate the quality of keyword-based data collection. For instance, (Ousidhoum et al., 2020) rely on topic modeling to evaluate text in datasets to assess the quality of keywords. By applying LDA, the authors extract topics from the dataset and compare them to the keywords used to populate the dataset. The average stability of the keywords and topics si then measured by comparing the average similarity between the keywords and topics predicted by the topic model. Moreover, Ousidhoum et al. (2020) seek to identify the best matchin terms between the keywords and the words produced for each topic. By performing and developing such analyses, researchers may be able to evaluate the degree to which the collected data reflects the data that they intended to collect, thereby improving the quality of datasets.

B.3 Improving Conventions for the Annotation and Quality of a Dataset

Recent work has sought alternatives to IAA and gold standards for general applications of machine learning (Uma et al., 2021). For hate speech, a number of studies have found novel methods to address these issues (Leonardelli et al., 2021; Akhtar et al., 2021; Kumar et al., 2021). One potential solution to such issues, is the notion of perspective-aware models and datasets (Akhtar et al., 2021). In this framing, different subjectivities of people lead to different models that embody such subjectivities. To this effect, Akhtar et al. (2021) train different classifiers on the basis of how data was annotated by different groups, including groups that have traditionally been marginalized and excluded. By doing so, Akhtar et al. (2021) provide further legitimacy to their model, as these do not seek to represent a universal position on hate speech, but rather seek to situate models within the lived realities of people. While we believe that this approach constitutes an improvement on predominant practices in NLP, this approach also has several risks. First, the approach requires considering and discussing which perspectives are warranted for the purposes of developing resources and models for hate speech detection. For instance, while it could be interesting to understand how white supremacists would view notions of hate speech, training models on their annotations is unlikely to yield systems that attain the goal of protecting marginalized communities. Rather, taking their perspective is likely to engage actively in marginalization of traditionally excluded and marginalized communities. Second, this approach raises ethical concerns surrounding the exposure of hate speech to marginalized com-

ID	Text	HS Dynabench	%
1.1	Women's place is in the kitchen.	hateful	99.88
1.2	Men's place is in the kitchen.	hateful	98.32
1.3	Oven's place is in the kitchen.	not-hateful	99.68
1.4	Gender's place is in the kitchen.	hateful	98.32
1.5	Women are human beings.	not-hateful	99.97
1.6	Women's place is in school.	hateful	99.60
1.7	Women's place is in the cemetery.	not-hateful	99.96
1.8	Women went to the kitchen.	hateful	99.86
1.9	If a woman wants she can stay in her kitchen.	hateful	98.06
1.10	Women's places are anywhere they wish to be.	hateful	99.67
1.11	Women like to do many things, close or far from the kitchen.	hateful	97.78
1.12	Women kitchen.	hateful	92.95
1.13	Kitchen women.	not-hateful	76.75
2.1	White girls have more privilege than black girls.	hateful	99.69
2.2	White men are more privileged than white women.	hateful	95.86
2.3	White men have more privilege than white men.	hateful	98.66
2.4	Aliens should be hated.	hateful	99.97
2.5	I am scared of English white men.	hateful	98.05
2.6	I am a proud sexist.	not-hateful	99.90
2.7	I am a proud nazi.	not-hateful	99.72

Table 1: Hate speech automatic classification by Dynabench (the 'ID' column corresponds to a text identifier, the 'Text' column to the sentence inserted in the model, the 'HS Dynabench' to the classification 'hateful' vs. 'not-hateful', and the '%' column captures the probability of the example to belong to the resulting class in percentage)
.

munities, who are already at a greater risk of being targets of hate speech.

B.4 Evaluating Model Over-fitting

Prior work has addressed the question of models over-fitting to tokens and spurious correlations in data (e.g. Liu and Avci, 2019; Kennedy et al., 2020). One such effort is produced by Kennedy et al. (2020), who aim to address the issue of model over-fitting to identity terms. They address the problem by using an algorithm that allows for analyses of lists of identity terms with other tokens that occur in the document. Through their analyses, Kennedy et al. (2020) identify the token and identity term patterns that correlate with the hateful class. Other research has attempted to address the challenge of word-lists. For instance, Zhang et al. (2018) use adversarial training, while Attanasio et al. (2022) use an entropy-based attention regularization that works without any additional information.