Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective[*]

Svetlana Kiritchenko, Isar Nejadgholi, Kathleen C. Fraser
National Research Council Canada
{svetlana.kiritchenko, isar.nejadgholi, kathleen.fraser}@nrc-cnrc.gc.ca

Abstract

We review a large body of Natural Language Processing research on automatic abuse detection with a focus on ethical challenges, organized around eight established ethical principles: privacy, accountability, safety and security, transparency and explainability, fairness, human control of technology, professional responsibility, and promotion of human values. We highlight the need to examine the broad social impacts of this technology, and to bring ethical and human rights considerations to every stage of the application life-cycle, from task formulation and dataset design, to model training and evaluation, to application deployment. Guided by these principles, we identify several opportunities for rights-respecting, socio-technical solutions to detect and confront online abuse. We also stress that online abuse is a social problem and recommend that these solutions are grounded in theories and findings from social sciences and centered around lived experiences of affected communities.

1 Introduction

The pervasiveness of abusive content on the internet can lead to psychological and physical harms. While social media platforms strive to monitor online content and remove abusive posts, the sheer volume of posts poses significant problems. Automatic detection of abusive content can provide assistance and partly alleviate the burden of manual inspection.

A wealth of research in Natural Language Processing (NLP) has been devoted to the problem of automatic abusive content detection. Here, we use the term *abusive* broadly, defining it as any language that could offend, demean, or marginalize another person. Abusive language detection has been studied under a plethora of names, such as *cyberbullying*, *hate speech*, *toxicity*, and others. Although current technologies achieve high classification performance in research studies, it has been observed that the real-life application of this technology can cause unintended harms, such as the silencing of under-represented groups.

We examine the task of automated abusive language detection from the *ethical* viewpoint, bringing together both technical and social issues under a single ethical and human rights framework. We begin by gathering all the related sub-fields, and briefly survey the past work with a focus on the different task formulations, common data collection and annotation techniques, and algorithms. We then discuss the challenges that the field faces from the ethical perspective, using the Harvard 'Principled Artificial Intelligence' framework as a scaffold [Fjeld *et al.*, 2020]. These challenges include fairness and mitigation of unintended biases, transparency, explainability, privacy, safety, and security. We discuss the trade-off between the right to free speech and the right to human dignity, and our professional responsibility to promote and protect all human rights in our work as AI researchers and practitioners. Finally, we discuss several promising directions for rights-respecting technology to confront online abuse.

Enumerating these ethical dilemmas is not simply an academic exercise; inattention to these issues can lead to human and economic harms in the real world. Through a better understanding of the ethical landscape, we hope to inspire new and creative solutions to effectively confront online abuse.

2 Overview of the Common Practices

We summarize the common practices in defining the task, collecting and annotating data, and training a predictive model.

Task Formulation. The abusive language detection task has typically been formulated as a supervised classification problem across various definitions of abusive language. In addition to the main task of determining whether a text is abusive or not, several other dimensions have been explored, including categories of abuse, implicit versus explicit abuse, target of abuse, legality of abuse, and the implied stereotypes in abusive language [Fišer *et al.*, 2017; Vidgen *et al.*, 2019; Sap *et al.*, 2020]. Multiple terms and definitions have been used to describe abusive content depending on how the abuse is expressed (e.g., hate speech, insult, physical threat, stereotyping). Focusing on slightly different aspects of abuse, these categories have obscure boundaries, and are often challenging for humans and machines to tell apart [Founta *et al.*, 2018]. Even the definitions of a single category (e.g., hate speech) can vary among researchers, and result in incompatible datasets [Fortuna *et al.*, 2020].

[*] This paper is an extended abstract of the article in Journal of Artificial Intelligence Research, 71: 431-478, July 2021.

Data Collection and Annotation. Abusive content is relatively infrequent, and random sampling results in datasets extremely skewed towards benign samples [Founta *et al.*, 2018]. Most existing sampling strategies mainly rely on using known abusive/profane lexicons to find abusive content. Although keyword search is simple and efficient, the choice of search terms for querying social media can lead to topic bias in trained classifiers [Wiegand *et al.*, 2019]. Data annotation, often performed through crowd-sourcing, presents a number of challenges as well. In addition to the inherent subjectivity of language, differing understandings of what to consider abusive language can lead to low inter-annotator agreement across and even within datasets.

Algorithms. The task of abusive language detection has been formulated as a supervised binary, multi-class, multi-label, or multi-task classification problem. Various machine learning algorithms, from dictionary-based and rule-based, to feature-based and deep neural networks, have been applied to build abuse detection systems, and high performances in cross-validation settings have been achieved as these algorithms improved. However, these models are not always robust when it comes to cross-dataset generalization [Arango *et al.*, 2019]. Ensemble and multi-task learning have been shown to improve generalization [Waseem *et al.*, 2018]. Further, multi-lingual multi-task learning as well as zero- and few-shot learning can improve the performance on tasks and languages with limited amount of annotated data.

3 Current Ethics-Related Challenges

We review current technological and sociological challenges in the field of automatic online abuse detection with respect to eight common ethical and human rights principles. These eight principles emerged as core thematic trends outlined in many ethical AI frameworks and guidelines as summarized in the recent study by the Berkman Klein Center for Internet & Society at Harvard University [Fjeld *et al.*, 2020]. Table 1 summarizes the ethical challenges in addressing online abuse, which we discuss below.

Promotion of Human Values. This principle is largely congruous with fundamental human rights, and includes the following three main concerns: supporting and promoting human values and human flourishing, benefiting society, and ensuring broad access to technology. Online abusive content detection brings forward two conflicting human values: freedom of speech and respect for equality and dignity [Maitra and McGowan, 2012]. This conflict can be viewed as two sides of the same coin: protection of equality and dignity is necessary to ensure that everybody has the right to free speech, and that the voices of minority groups and individuals are not silenced through threats and offensive behavior [Delgado and Stefancic, 1997]. Also, automation of content moderation can reinforce social hierarchies and amplify social inequalities by limiting access to technology to certain groups, as researchers and companies implement abusive language detection algorithms for some languages and not others. In addition to linguistic differences across languages, notions of what is 'offensive' may be culturally-specific, presenting further challenges to creating datasets in multiple languages and applying knowledge transfer and multi-lingual approaches.

Fairness and Non-Discrimination. Algorithmic decision-making can perpetuate social biases by discriminating against individuals because of their membership in certain social groups. Fairness research focuses on identifying and mitigating biases that can potentially be harmful to certain subpopulations. Unintended biases in NLP systems can originate from various sources. Pre-trained word embeddings and language models often encode social biases and are also prone to generating racist, sexist, or otherwise toxic language [Gehman *et al.*, 2020]. Data sampling techniques deployed to boost the number of abusive examples may result in a skewed distribution of concepts and entities related to targeted identity groups. These unintended entity misrepresentations often translate into biased abuse detection systems. Human annotators' demographic features, such as first language, age and education, as well as insensitivity to dialect or limited knowledge of various aspects of abusive behavior can also lead to biased annotations [Sap *et al.*, 2019]. Finally, learning algorithms can further amplify or mitigate certain biases. While it is impossible to completely eliminate all biases, AI researchers and developers can work towards quantifying unfairness in system outputs for various demographics, identifying the origins of different types of biases, and designing techniques to minimize the harmful outcomes.

Transparency and Explainability. These principles intend to shed light on the process of creating an automatic system and make it understandable by different stakeholders. Transparent documentation on the purpose, scope of use, performance, and provenance of automatic systems as well as on the data on which the models were trained can help in mitigating the ethical risks [Gebru *et al.*, 2018]. Effective explanations of the system's inner workings and its decisions, tailored to different stakeholders, are critical for the realization of accountability, human control of technology, safety and security, and fairness and non-discrimination.

Privacy. This principle encompasses both the use of user data to train machine learning models for online abuse detection in research and commercial settings, and individual users' agency to control their personal data. Another issue related to user privacy and online abuse is the practice of "doxxing", or publishing private information about a person online, typically in order to subject that person to harassment. Very little research so far has considered the automated detection of such behaviour.

Safety and Security. These measures are crucial elements for building reliable systems: systems that are safe, in that they perform as intended, and also secure, in that they are not vulnerable to being compromised by unauthorized third parties. One of the major safety risks is the mismatch between the training and test environments. To minimize this and other safety risks, models need to be systematically tested on a variety of inputs and language phenomena and continuously maintained and retrained as the language of the users evolves over time. Security of current abuse detection systems can also be easily compromised by adversaries through insertion of typos or addition of innocuous words to the original abusive texts [Gröndahl *et al.*, 2018].

Ethics Theme	Online Abuse Specific Issues	Related AI Challenges
Promotion of human values	Finding balance over two conflicting human rights, freedom of speech and respect for equality and dignity	Overcoming ambiguous and non-realistic task formulations; designing alternative applications to ensure safe communication environments for all
Fairness & non-discrimination	Striving for equal system performance on texts that are about or written by different demographics	Collecting representative datasets; identifying, quantifying and mitigating potentially unfair system outputs; optimizing measures of fairness besides overall accuracy
Transparency & explainability	Moving away from making critical decisions using black box models; providing developers with tools to inspect systems' behavior and identify risks; providing explanations for automated decisions	Producing and maintaining high-quality documentation (data sheets and model cards); designing and using interpretability tools to detect biases in models; providing accessible explanations to users
Privacy	Ensuring data privacy, personal privacy, and users' right to control their own data	De-identifying personal data; applying privacy-preserving computation (e.g., federated learning); allowing users to remove their data from training corpora
Safety & security	Considering consequences of false positive and false negative decisions; building systems that do not heavily rely on keywords, are not easy to deceive, and are robust against adversarial attacks	Measuring and minimising the risk of false decisions; identifying system vulnerabilities, including susceptibility to spurious correlations; improving the out-of-distribution robustness; testing systems in real-world scenarios
Accountability	Auditing systems and assessing their impact on individuals, society and environment; ensuring the ability to appeal; setting legal responsibilities	Auditing design decisions internally throughout all stages of application development and deployment; designing and employing interpretability and explainability tools
Human control of technology	Moving away from fully automated moderation due to inaccurate systems; enabling users to appeal automated decisions and request human review	Enabling human-in-the-loop technologies; providing rationale to users to enable appeals
Professional responsibility	Building accurate systems; considering potential long-term effects; refusing to work on harmful applications; engaging all stakeholders; upholding scientific integrity	Evaluating system performance in various settings; involving stakeholders in the design process; raising public awareness for long-term possible harms of technology (e.g., censorship)

Table 1: Ethics and human rights related issues in online abuse detection, and the associated NLP/AI challenges.

Accountability. This principle refers to the concerns about who is accountable for automatically made decisions as well as the potential impacts of the technology on the social and natural world. It is generally agreed that the organizations that develop and deploy AI systems should be responsible for the systems' outcomes and impacts. Audit for ethical compliance, both internal and external, is required for accountability at both the levels of development and deployment. Currently, the social media corporations have little accountability for either automatic or human decisions regarding content moderation. Civil society organizations, policy makers, and regular users call for better transparency to the public about the processes (including automatic decision making) and results of content moderation, meaningful opportunities for users to appeal any content or account suspension or removal, and justification for any content removal decisions.

Human Control of Technology. This principle refers to the ability of users to appeal automated decisions and request human review, or even opt out of automated decisions entirely. Given the ambiguity of language and the need to protect freedom of speech, human review of uncertain or contested decisions is often essential. For users to be able to challenge the system's outputs, the outcomes have to be presented in an easy-to-understand form with information on the factors and logic that influenced the decision.

Professional Responsibility. This principle encompasses tenets such as ensuring the accuracy of the systems we build, adopting principles of responsible design, considering the long-term effects of our work, engaging stakeholders who may be affected by our systems, and upholding scientific integrity. NLP researchers and developers need to situate their work within a broader societal and historical context, uncover the implicit assumptions and normative values being reinforced, and critically assess the potential impact of the technology on a global scale. This can include taking actions such as: choosing not to work on projects that do not support the social good or that have the potential for long-term harm, engaging with stakeholders and taking their feedback seriously, and being open and transparent about the limitations and failures of our technology, including publishing negative results.

4 Ways to Move Forward

In this section, we outline several emerging research themes where the AI community can contribute to developing ethics-aware technologies to tackle online abuse. Figure 1 shows some of the ways that the discussed approaches can be incorporated at different stages of the design, development, and deployment of AI solutions.

While the task of online abuse detection is commonly formulated as a binary or multi-class classification problem, it becomes increasingly evident that abusive language is much more nuanced, and such a task formulation significantly limits the applicability of the developed technology in real life. Online abusive content embodies a spectrum of practices that differ in motivation, expression, and consequences, and needs to be examined within a regional and historical context [Po-

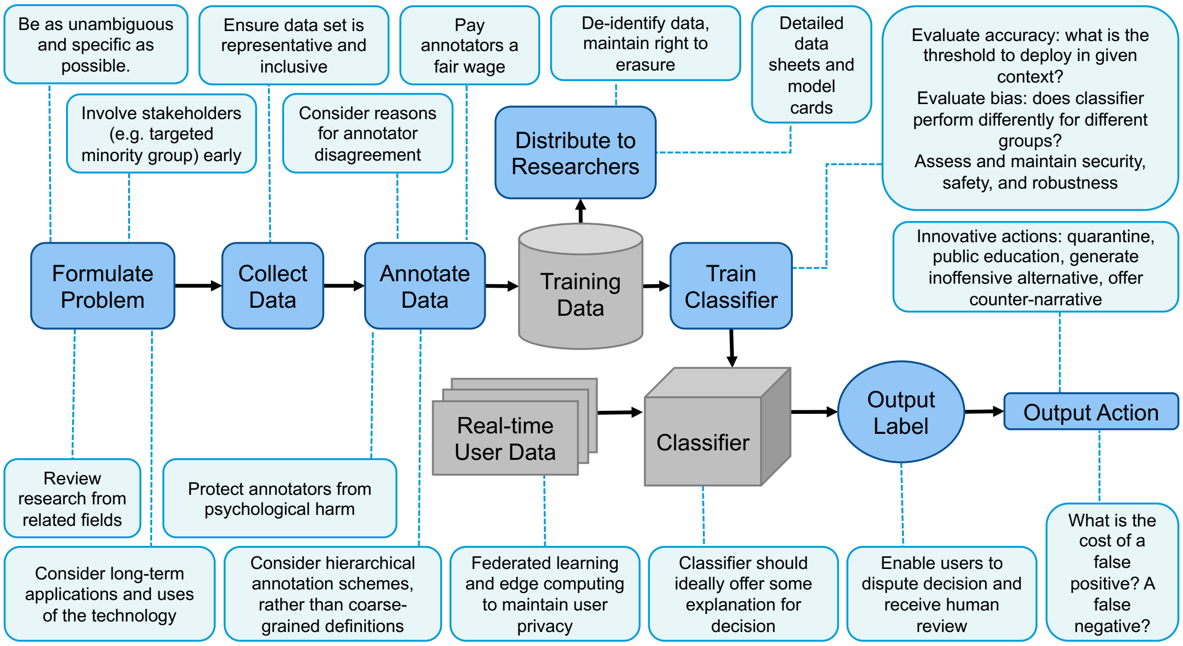
Be as unambiguous and specific as possible.
Ensure data set is representative and inclusive
Pay annotators a fair wage
De-identify data, maintain right to erasure
Detailed data sheets and model cards
Evaluate accuracy: what is the threshold to deploy in given context?
Evaluate bias: does classifier perform differently for different groups?
Assess and maintain security, safety, and robustness
Involve stakeholders (e.g. targeted minority group) early
Consider reasons for annotator disagreement
Distribute to Researchers
Formulate Problem
Collect Data
Annotate Data
Training Data
Train Classifier
Innovative actions: quarantine, public education, generate inoffensive alternative, offer counter-narrative
Real-time User Data
Classifier
Output Label
Output Action
Review research from related fields
Protect annotators from psychological harm
What is the cost of a false positive? A false negative?
Consider long-term applications and uses of the technology
Consider hierarchical annotation schemes, rather than coarse-grained definitions
Federated learning and edge computing to maintain user privacy
Classifier should ideally offer some explanation for decision
Enable users to dispute decision and receive human review

Figure 1: A high-level overview of some of the ways ethical considerations can be incorporated throughout the ML pipeline.

hjonen and Udupa, 2017]. This defies easy binary division into content that is acceptable and content that is not. Recently, the complexity of the task formulation has started to be recognized by the research community, and as a first step, several studies have proposed multi-dimensional, multi-level frameworks to address the task.

The current practices of dealing with abusive content by major social media platforms are also often binary: posts that are deemed to violate a platform's terms of use are permanently removed; all the other posts are shown to users. Such black-and-white decisions can lead to power abuse by the social media companies, restricting users' rights to freedom of speech and causing harm to individuals and businesses. Several alternative mechanisms, such as 'nudging' and value sensitive design [Vincent and Jane, 2017], 'quarantining' [Ullmann and Tomalin, 2020], and style transfer [Nogueira dos Santos *et al.*, 2018], have been proposed in the literature that provide a middle ground between permanent removal of some content and no content moderation at all. In addition to preventing or at least discouraging users from posting abusive content, other mechanisms of mitigating the harmful effects of online abuse have been proposed. These include counter-narrative (or counterspeech), which provides a non-aggressive response to abusive content that aims to deconstruct any stereotypes and misinformation with thoughtful reasoning and fact-bound arguments. Other approaches include the automatically detecting potentially viral posts and flagging them for manual inspection, as well as AI-driven applications of public education on evolving social norms and the potential harms of abusive language.

Further work on interpretability and explainability is critical in addressing several ethical principles mentioned above. Different stakeholders, including designers and developers of the systems, data scientists, regulators, and end users, can use explanations to help debug the system, validate its fairness, improve its security, or appeal its decisions. However, these different stakeholders and their different objectives require divergent, tailored solutions [Vaughan and Wallach, 2020].

The problem of online abuse cannot be solved by AI technology alone as, ultimately, online abuse is a social problem that can be either amplified or mitigated with the help of technology. Therefore, AI researchers should work together with social scientists, anthropologists, psychologists, criminologists, human rights activists, and ethicists to understand abusive online behavior, its motivations and expressions, and how it is propagated through social networks, and to design communication technologies that encourage ethical behavior and discourage unethical behavior [Prabhakaran *et al.*, 2020].

Finally, technological solutions should be centered around the lived experiences and the needs of the victims of online abuse. Involvement of the affected communities in the design decisions, including the decision of whether to build a particular system at all, would help better position the task in the social and political context, account for its many nuances, and identify and mitigate potential ethical issues [Katell *et al.*, 2020]. Successful community engagements are built around trust between researchers and the community members. Further, committing to a fair and transparent compensation instead of limiting the engagement to low income or volunteer work is another way forward to productive partnerships.

By identifying these future directions for research and critical thinking, we hope to better position NLP researchers to make useful and ethical contributions to the problem of abusive content online.

References

[Arango *et al.*, 2019] Aymé Arango, Jorge Pérez, and Barbara Poblete. 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*, pages 45–54, 2019.

[Delgado and Stefancic, 1997] Richard Delgado and Jean Stefancic. *Must we defend Nazis?: Hate speech, pornography, and the new first amendment*. NYU Press, 1997.

[Fišer *et al.*, 2017] Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*, pages 46–51, 2017.

[Fjeld *et al.*, 2020] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, 2020.

[Fortuna *et al.*, 2020] Paula Fortuna, Juan Soler, and Leo Wanner. 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*, pages 6786–6794, May 2020.

[Founta *et al.*, 2018] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2018.

[Gebru *et al.*, 2018] Timnit Gebru, J. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, H. Wallach, Hal Daumé, and K. Crawford. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden, 2018.

[Gehman *et al.*, 2020] Sam Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*, 2020.

[Gröndahl *et al.*, 2018] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. All you need is "love" evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 2–12, 2018.

[Katell *et al.*, 2020] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 45–55, 2020.

[Maitra and McGowan, 2012] Ishani Maitra and Mary Kate McGowan. *Speech and harm: Controversies over free speech*. Oxford University Press on Demand, 2012.

[Nogueira dos Santos *et al.*, 2018] Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 189–194, Melbourne, Australia, July 2018.

[Pohjonen and Udupa, 2017] Matti Pohjonen and Sahana Udupa. Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, 11:19, 2017.

[Prabhakaran *et al.*, 2020] Vinodkumar Prabhakaran, Zeerak Waseem, Seyi Akiwowo, and Bertie Vidgen. Online abuse and human rights: WOAH satellite session at RightsCon 2020. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 1–6, 2020.

[Sap *et al.*, 2019] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah 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.

[Sap *et al.*, 2020] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020.

[Ullmann and Tomalin, 2020] Stefanie Ullmann and Marcus Tomalin. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80, 2020.

[Vaughan and Wallach, 2020] Jennifer Wortman Vaughan and Hanna Wallach. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence*, 2020.

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

[Vincent and Jane, 2017] Nicole A. Vincent and Emma A. Jane. Beyond law: Protecting cyber victims through engineering and design. In Elena Martellozzo and Emma A. Jane, editors, *Cybercrime and its Victims: An International Perspective*, pages 209–223. Routledge, Oxon, 2017.

[Waseem *et al.*, 2018] Zeerak Waseem, James Thorne, and Joachim Bingel. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In J. Golbeck, editor, *Online harassment*, pages 29–55. Springer, 2018.

[Wiegand *et al.*, 2019] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of abusive language: the problem of biased datasets. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 602–608, 2019.