Technological Solutions to Online Toxicity: Potential and Pitfalls

Arezo Bodaghi [a], Benjamin C. M. Fung [b], and Ketra A. Schmitt [c]

a) Concordia Institute for Information Systems Engineering, Concordia University, Canada arezo.bodaghi@concordia.ca
b) School of Information Studies, McGill University, Canada ben.fung@mcgill.ca
c) Concordia Institute for Information Systems Engineering and the Centre for Engineering in Society, Concordia University, Canada ketra.schmitt@concordia.ca

Social media platforms present a perplexing duality, acting at once as sites to build community and a sense of belonging, while also giving rise to misinformation, facilitating and intensifying disinformation campaigns and perpetuating existing patterns of discrimination from the physical world. The first step platforms take in mitigating the harmful side of social media involves identifying and managing toxic content. Users produce an enormous volume of posts which must be evaluated very quickly. This is an application context that requires machine learning (ML) tools, but as we detail in this paper, ML approaches rely on human annotators, analysts, and moderators. Our review of existing methods and potential improvements indicates that neither humans nor ML can be removed from this process in the near future. However, we see room for improvement in the working conditions of these human workers.

Dealing with this problem is challenging due to the impracticality of manually removing toxic content, given the volume, velocity, variety of online material. Consequently, platforms have increasingly adopted moderation systems incorporating machine learning (ML) models. However, these models have limitations, including biases and discriminatory outcomes. As a result, some platforms have opted to engage human moderators in assessing content flagged as potentially toxic by ML models and making the final decisions. However, these systems still undeniably have limitations that require deep investigation. Therefore, this article aims to comprehensively review moderation systems for automatic toxicity detection in social media, emphasizing the need to understand their constraints.

1. Toxic Language on Social Media

Social Media platforms play a vital role in our daily lives by enabling users to stay in touch, express their views in real-time, and providing instant access to information. However, the smooth sharing of content, and the cover of anonymity on microblogging platforms along with the lack of normative cues in online interactions also contribute to the widespread dissemination of antisocial and toxic behaviours [1].

The pervasive use of toxic language on social media has serious social implications. Individuals may hesitate to express their opinions or participate in discussions due to the fear of being targeted with harmful content. In severe cases, this phenomenon can lead to mental health issues and social isolation, particularly impacting teenagers. Toxicity sometimes co-occurs with hate speech, violent threats and can cross-over into the physical realm through doxing, swatting, or stalking.

Manually removing toxic content is infeasible due to the sheer volume and multilingual diversity of online content. Human removal would require enormous staffing levels and expense, prompting the adoption of machine learning (ML) techniques as a viable solution. While ML models exhibit effectiveness in real-time content classification on a large scale, their performance is not without limitations, particularly in the context of toxic language [2].

Toxicity detection models may confidently make incorrect predictions based on spurious lexical features. In addition, subtle wording similarities can lead to inaccurate outcomes, highlighting a significant concern: bias in modeling, training, and usage [3], [4]. This bias may result in discrimination against specific social subgroups, including Black users, women, and LGBTQI+ communities, in automated decision-making systems [5].

Despite these technical limitations the urgent need to address anti-social online behavior has led platforms to frequently adopt moderation systems integrating ML-based models. These systems are used to identify potentially harmful content, and some platforms also involve human moderators to review flagged content and make the final decision. It is worth noting that human judgments are utilized for annotating training datasets to develop **ML models**. As a result, both

ML techniques and human individuals (moderators and annotators) remain crucial components for effective moderation systems in controlling online toxicity.

This collaboration between humans and ML has constraints that necessitate a deep dive into potential causes of misidentification and poorly managed outcomes. Our aim in this article is to comprehensively review ML techniques designed for automatic toxicity detection, emphasizing the need to examine and understand their limitations in the dynamic and varied landscape of social media content. We categorize these methods into social, policy, and technical approaches, with a particular focus on technical solutions. We explore the potential of strictly technical approaches to address these risks. We also explore the limitations inherent in a purely "technological fix". We close by considering alternate approaches to creating a less toxic social media experience.

2- Tools for Evaluating Antisocial behavior

Little scholarship exists that explores the limitations of innovative techniques for addressing antisocial behaviors. Some existing studies have examined Natural Language Processing (NLP) techniques for automatic hate speech detection while also recognizing and addressing their limitations [6]. Others have explored the reliability of pre-trained large language models (LLMs), assessing their effectiveness in decision-making tasks that involve aspects of uncertainty, robust generalization, and adaptation [7]. Furthermore, a comprehensive analysis of the risks associated with LLMs has been conducted to better guide responsible innovation [8]. This analysis draws from various fields, including computer science, linguistics, and social sciences. The examination and consideration of uncertainty in machine learning techniques, including the realm of online toxicity detection, have prompted the proposal and application of methods designed to address this issue [9][10].

Despite these efforts, uncertainty persists around the efficacy of ML techniques to detect and manage toxic speech. Notably, the presence of biases in both data and algorithms poses a significant issue, potentially resulting in increased discrimination [11]. Bias in datasets causes posts from minoritized groups to be over flagged (False Positives (FP) identifying toxicity when none existed) [12]. Furthermore, research has predominantly concentrated on textual content, and

the conclusions drawn about uncertainty in toxicity detection using ML techniques for textual posts may not necessarily be applicable to multimodal content. Consequently, it is necessary to gain a comprehensive understanding of uncertainty in identifying toxic multimodal content.

Deploying ML techniques at the platform level poses broader challenges. Some studies indicate that considering the prediction uncertainty of neural networks facilitates detecting complex text inputs. This includes short or lengthy texts with less informative tokens and potentially incorrect predictions, necessitating manual verification. Google's Jigsaw team [13] presented CoToMoD, a benchmark to assess the effectiveness of systems involving both ML models and human moderators. They introduced principled metrics like Oracle-Model Collaborative Accuracy (OC-ACC), OC-AUC, and Review Efficiency, which evaluate the system's performance in utilizing human attention and decisions, going beyond traditional predictive performance or uncertainty calibration measures.

3. Challenges in Online Toxicity Detection

Obstacles to effective ML detection occur at various stages of content moderation. We start by examining the challenges related to obtaining labeled datasets, which can introduce uncertainty and impact the accuracy of prediction results. Following that, we delve into the modeling approach, covering both the training and testing phases.

3.1 Challenges in Data Preparation

Machine learning techniques, particularly supervised learning, are valuable tools for detecting online toxicity [14]. The initial and crucial step in these techniques involves data collection, which, in the context of online toxicity, entails gathering data from various platforms. **Social media platforms have a global reach, empowering people worldwide to create profiles.** However, the pattern of use differs among regions, with certain platforms being popular in one region than another, and others being outright banned by federal governments. This regional variability in use patterns adds a layer of complexity to analysis but does not significantly impact the diverse user base across social media platforms. The diversity results in vast amounts of unstructured and

multilingual content being shared every second on social media platforms, rendering its management exceptionally complex. Once a dataset is collected and cleaned, it needs to be reviewed and labeled by annotators to be used for training classifiers. This task itself has many challenges. Many datasets are publicly available for toxicity detection, including the Jigsaw Toxicity Dataset [15], which involves Wikipedia comments labeled by human annotators, and ToxiGen [16], comprising a large-scale machine-generated dataset that addresses adversarial and implicit hate towards minority groups. However, precise guidelines on how these datasets were annotated are not available. In some cases, tools such as Google's Perspective API[1], HATECHECK[2] or pretrained models including HateBERT [17] are used to annotate the dataset. However, they may not perform perfectly, indicating that the labeled dataset may not be perfectly accurate and can potentially introduce bias [18]. Therefore, for each specific task, a rubric is required for annotation. A rubric is a set of guidelines on how specific words should be interpreted in different contexts. However, this guidance document is not enough to ensure unbiased annotation of potentially toxic posts. Annotators need an ethical framework to direct the task. Two additional challenges relate to the annotators themselves. Human annotators are exposed to sometimes violent and disturbing content, which has an often-hidden human cost. **These annotators are often paid far less than prevailing minimum wage, and below a living wage [19], [20].**

These poor working conditions compound the second challenge human annotators face, which is having sufficient identity knowledge of the target of anti-social behavior to accurately identify toxicity. For example, when detecting toxic comments against LGBTQIA2+ people, the annotators should ideally include people from that group to review the provided dataset and label each sample according to the rubric. However, detailed demographic information about the annotators is rarely available. Moreover, another significant challenge arises from the ambiguity that exists between various forms of toxic language, including hate speech and offensive language. **In addition, annotating multilingual datasets are also challenging. Many ML algorithms are trained primarily on English-language data, leading to potential shortcomings in their performance when applied to other languages. Furthermore, these models often struggle to accurately**

[1] https://perspectiveapi.com/
[2] https://hatecheck.ai/

identify subtle content in statements that carry multiple meanings. An illustrative example is the challenge posed by dog-whistle phrases in various languages. A dog whistle refers to a pejorative term deliberately crafted to be discernible only by individuals actively engaged in discrimination against a particular group, while remaining undetected by the general population—those who neither experience discrimination in this manner nor partake in discriminatory behavior [21].

Deep Learning (DL) and transformer-based models have demonstrated effective results in identifying toxic content [22]. However, they may also worsen the problem of data bias since these models rely heavily on large amounts of training data that may not be inclusive of all user groups. **While Large Language Models (LLMs) have recently shown impressive capabilities in a wide range of applications and tasks, including natural language understanding, generation, and translation, it is important to note that they can also inherit biases from the training data, as this data can mirror the societal biases present during its collection**. One potential remedy to inherited data bias is to create models with logic awareness. Adam et al. [23] conducted a study to determine if language models that incorporate logic-awareness could successfully mitigate the presence of harmful biases. Their findings revealed that when a language model lacks explicit logic acquisition, it often displays a significant degree of biased reasoning, while, integrating logic learning into the language model can decrease.

Rule enforcement is another approach to reducing anti-social behavior. Platforms work to keep users safe by enforcing rules like removing toxic posts or suspending accounts from users who frequently engage in hateful conduct. But it is not clear if suspending users who behave badly reduces the overall prevalence of toxic speech online. Suspending an account may encourage bad actors to migrate to other platforms with fewer rules. Ali et al. [24] analyzed posting behavior on Twitter (now called X) and Reddit and compared the same individuals' content on Gab, a site known for promoting hateful conduct and not enforcing behavioral rules. The results of this comparison revealed that users exhibited increased toxicity when they experience suspension on one platform and are compelled to migrate to another. Additionally, their level of activity rises, leading to a higher frequency of posts.

In terms of the process of managing online toxicity, removing toxic content, and suspending users are critical moments because these actions have important consequences. Content moderation involves both people and ML models working together. For instance, the model can pick outposts that probably break the rules, and then human moderators can take a closer look at them. Accurate classification is crucial in this situation. If toxic content is mistakenly identified as non-toxic (False Negative (FN)), users will see harmful content (even if they have settings to prevent this). Moreover, having a high number of FP or FN can lead users to disengage from discussions and become less active on these platforms. In addition, over-flagging content as toxic can also drive users away, as it is likely that many of their posts will be marked as toxic, discouraging them from posting or commenting due to the high chance of their content getting blocked. Jhaver et al. [25] examined how Reddit users responded to the platform's moderation process. Their findings showed that 18% of the participants agreed that their posts were correctly removed, 37% were uncertain about the reasons behind their post removal, and 29% expressed frustration regarding the removal of their posts.

3.2 Challenges in Model Construction

After preparing the dataset, the subsequent step involves training and testing the model, where the importance of a well-prepared dataset cannot be overstated. Quality and quantity both play vital roles in this phase. To ensure the reliability of results, it is imperative to maintain a near balance in the number of samples across different classes, creating a balanced dataset. Datasets with imbalanced class distributions pose a frequent challenge across various classification tasks [26]. This issue is particularly challenging in the domain of toxic language detection because toxic language is typically less frequent when compared to non-toxic language. As an example, consider a dataset of 100,000 tweets gathered from Twitter (now called X), where the proportion of toxic tweets was relatively low, accounting for approximately 5% (around 5,000 instances), while the majority of the data consisted of non-toxic content [27]. Consequently, this data imbalance can result in models excelling at recognizing nontoxic language but struggling with identifying toxic content. This issue can be compounded by the use of self-reported data or crowd-sourcing for annotation, which may not accurately reflect the diversity of toxic language used online. Various solutions have been proposed to address class imbalance issues at both the data and algorithmic

levels. At the data level, these solutions involve different types of re-sampling, such as random oversampling, random under sampling, directed oversampling, directed under sampling, oversampling with informed generation, and their combinations. At the algorithmic level, solutions include adjusting class costs, adjusting the probabilistic estimate at tree leaf, adjusting the decision threshold, and recognition-based learning. While these techniques have shown promising results in improving the performance of models on imbalanced datasets, they also have limitations. Oversampling and under sampling can lead to overfitting and underfitting, respectively, and may not work well when the dataset is extremely imbalanced [28]. Cost-sensitive learning requires accurate estimation of the misclassification costs, which may be difficult in practice. These oversampling techniques are particularly effective in image data but may not perform optimally with textual content. Therefore, data augmentation methods, like synonym replacement, back-translation and text generation, have been introduced, primarily tailored to address these limitations in textual data. While employing back-translation techniques can enhance accuracy by balancing the dataset, it is crucial to acknowledge that this approach may still lead to a significant number of false detections, which can be prohibitively costly in real-world applications [29]. The occurrence of misclassifications highlights an additional challenge, which is closely tied to the existing evaluation metrics. The most frequently used evaluation metrics encompass Accuracy, Precision, Recall, and the Area Under the ROC Curve (AUC-ROC). These metrics effectively measure the model's performance on training and testing data but may not accurately reflect how the model will perform in real-world applications. In the work done by [18], they provided evidence that an adversary can make subtle changes to a highly toxic phrase, causing the system to assign a significantly lower toxicity score. Their experiment involved applying this method to the sample phrases provided on the Perspective website, consistently reducing the toxicity scores to the level of non-toxic phrases. This finding underscores the detrimental effect of adversarial examples on the usability of toxic content detection systems.

Another challenge to address is the requirement for high throughput, particularly in the context of toxicity detection due to the enormous data flow received every second. While throughput is a critical aspect for ML techniques in general, it holds even greater significance in the realm of toxicity detection. This implies that the models we develop should not only be highly accurate but also exceptionally fast, as users expect their posts and comments to appear for their followers

within seconds. The most proposed techniques in the literature are tested in laboratory conditions on relatively small datasets and timescales, where processing speed is not an urgent consideration. In order to implement these models at a platform level their performance must be characterized at large scales and in real-time. At present, no/only one/few published articles engage with platform-level performance, and only one considers processing speed. Ensuring that toxicity detection techniques perform adequately at scale requires constructing measures of uncertainty along with the development of reliable performance evaluation metrics. that fully consider uncertainty, continues to be an ongoing concern.

4- Humans and Machine Learning: a team approach to detection

Our exploration of the methods available to limit toxic speech online revealed that both ML techniques and human intervention are necessary through the process of data collection, annotation, analysis, and action.

While data collection processes are easily automated, human annotators are still vital in enabling any ML technique to operate on data. While automated toxicity detectors exist, our analysis and others have shown that these have very low reliability. A human annotator must interact with content to understand its meaning within a context and community. But while these human annotators are needed to enable this process, their jobs are not well remunerated, nor are they pleasant for the person who must sift through toxic content. Once trained, several approaches can do a relatively effective job in identifying toxic posts. However, certain techniques are necessary to limit data bias that can inadvertently over flag posts from marginalized communities. Human intervention is again needed to verify automated detection approaches, and to make more permanent decisions like content removal or user suspension (or to verify automated decisions on these topics).

Given the sheer volume of posts that are made within and across platforms, the consequences of FP and FN from automated detection methods seem unacceptably high. But so do the social and ethical implications of hiring human annotators.

What other solutions exist beyond these two poles? Individual users can take actions to protect themselves from toxic speech. This can include settings to prevent inappropriate or offensive images from showing, as well as muting or blocking problematic posters. Groups of people also create block lists to create safe online experiences for communities with common values or identity. Some platforms also allow different levels of sharing for identified groups.

We then have three possible approaches: individual protective measures, human based measures by the platform, and automated approaches. All these approaches involve some level of tradeoff - time and effort for individual protective measures, harm and low pay for human annotators, and risks of bias and false positives and negatives for automated approaches.

What can be done to mitigate these risks?

Improve working conditions for annotators and human moderators

Gig-based approaches, like those monetized through Amazon Mechanical Turk program, have found a way to provide a technological version of piecework, a classic way to underpay workers by paying them a low rate for a unit of work, rather than paying a fair wage per hour. Because online work can cross international boundaries, they are able to find workers willing to take poor wages to annotate toxic posts. But contract work plays other critical roles for social media platforms. Contract workers are the first line in content moderation and face serious mental health consequences due to the violent and disturbing content that they remove [30]. Workers are forced to sign non-disclosure agreements (NDAs) which prevent them from sharing the horrifying nature of their working conditions. Again, the global reach of the internet allows for platforms to hire contracting companies to pay workers in the global south less and provide them with fewer protections [31].

Develop social media platform cultures that prioritize care and respect

Platforms, like other communities, have cultures. We recognize certain platforms for very high levels of toxicity (Gab, 4Chan, 8Chan, TruthSocial), but online spaces also exist that prize

kindness and good conduct. Reddit is an example of collectively managed communities that manage to enforce their own norms of behavior. One insight into these communities is that they may be connected through common interests, and therefore have stronger potential bonds and common values. For example, wattpad [32], a networking site where authors can share work in progress or reddit sub-communities like r/fountainpens which bring together people with niche interests. At the same time, community-based platforms can clearly have other cultures and prioritize other types of behavior. But these examples may provide some insights into how platforms can become kinder places.

Improve algorithmic approaches

Large language models have improved at an astonishing pace. While detection methods are improving, they are still beset by a number of challenges including misidentification, and difficulty in differentiating sentiment and toxicity in specific contexts. The ways to address these shortcomings are not immediately clear.

Hire adequate staff at the platform level and treat them decently

Recent cuts across a swathe of social media companies have given rise to concerns that platforms will lack the resources to combat serious threats like disinformation, false information and online toxicity [33]. X, in particular, has seen a rapid rise in disinformation and toxicity [34], [35]. The cultures of the social media platform organizations themselves have influences on their employees. Just as poor working conditions and pay for gig workers who annotate posts for ML influence the quality of annotation and raise ethical concerns, the toxicity of the corporate culture at X under the new leadership influences the performance of workers employed by the platform. Just like anyone else, the human annotators, programmers and staff tasked with removal and suspension decisions need safe and stable work environments.

5-Conclusion

A common narrative of AI processes states that the human element can be removed from work, including decision-making as well as boring, repetitive tasks. But this review of the process of

online toxicity detection demonstrates how much this narrative leaves out. Toxicity detection begins with dataset annotation that is typically carried out by human annotators (a boring, repetitive task). Although pretrained tools are employed in some instances, human analysis and intervention is often necessary due to the imperfect accuracy of these tools, reaffirming the central role of human annotation in the process. The ML tasks only begin once annotation is complete, when the classifier can be trained and tested on a small dataset. These techniques are then applied to user-generated content to distinguish between toxic and nontoxic content. In some cases, potentially incorrect classifications are forwarded to human moderators for content removal decisions. In other words, a human worker retains decision-making in the moderation process. The classification results are of utmost importance, with both false positives and, in particular, false negatives incurring significant costs. Both human annotators and the ML algorithms produce errors. Moreover, results from our lab and others indicate that ML techniques for annotation without a human user are prone to much higher error rates. This means that for the near future, human involvement remains integral to the moderation process to tackle toxicity. However, it is crucial to emphasize the importance of prioritizing human safety within the moderation framework.

That means that for the near future, humans remain a critical part of the moderation process used to control online toxicity, but these humans need protection: for all the talk about AI Safety, Human Safety should be a priority.

References

[1] A. Vasalou, A. N. Joinson, and J. Pitt, "Constructing my online self: avatars that increase self-focused attention," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, in CHI '07. New York, NY, USA: Association for Computing Machinery, Apr. 2007, pp. 445–448. doi: 10.1145/1240624.1240696.

[2] M. S. Jahan and M. Oussalah, "A systematic review of Hate Speech automatic detection using Natural Language Processing," *ArXiv210600742 Cs*, May 2021, Accessed: Apr. 27, 2022. [Online]. Available: http://arxiv.org/abs/2106.00742

[3] Z. Wang and A. Culotta, "Identifying Spurious Correlations for Robust Text Classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 3431–3440. doi: 10.18653/v1/2020.findings-emnlp.308.

[4] X. Ferrer, T. van Nuenen, J. M. Such, M. Coté, and N. Criado, "Bias and Discrimination in AI: A Cross-Disciplinary Perspective," *IEEE Technol. Soc. Mag.*, vol. 40, no. 2, pp. 72–80, Jun. 2021, doi: 10.1109/MTS.2021.3056293.

[5] F. Faal, J. Yu, and K. Schmitt, "Domain Adaptation Multi-task Deep Neural Network for Mitigating Unintended Bias in Toxic Language Detection:," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, Online Streaming, --- Select a Country ---: SCITEPRESS - Science and Technology Publications, 2021, pp. 932–940. doi: 10.5220/0010266109320940.

[6] A. S. Parihar, S. Thapa, and S. Mishra, "Hate Speech Detection Using Natural Language Processing: Applications and Challenges," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Jun. 2021, pp. 1302–1308. doi: 10.1109/ICOEI51242.2021.9452882.

[7] D. Tran *et al.*, "Plex: Towards Reliability using Pretrained Large Model Extensions." arXiv, Jul. 15, 2022. Accessed: Apr. 26, 2023. [Online]. Available: http://arxiv.org/abs/2207.07411

[8] L. Weidinger *et al.*, "Ethical and social risks of harm from Language Models." arXiv, Dec. 08, 2021. Accessed: Oct. 29, 2023. [Online]. Available: http://arxiv.org/abs/2112.04359

[9] M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, Dec. 2021, doi: 10.1016/j.inffus.2021.05.008.

[10] N. A. Palomares and V. S. Wingate, "Victims' Goal Understanding, Uncertainty Reduction, and Perceptions in Cyberbullying: Theoretical Evidence From Three Experiments," *J. Comput.-Mediat. Commun.*, vol. 25, no. 4, pp. 253–273, Jul. 2020, doi: 10.1093/jcmc/zmaa005.

[11] A. Luccioni and Y. Bengio, "On the Morality of Artificial Intelligence [Commentary]," *IEEE Technol. Soc. Mag.*, vol. 39, no. 1, pp. 16–25, Mar. 2020, doi: 10.1109/MTS.2020.2967486.

[12] 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*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1668–1678. doi: 10.18653/v1/P19-1163.

[13] I. D. Kivlichan, Z. Lin, J. Liu, and L. Vasserman, "Measuring and Improving Model-Moderator Collaboration using Uncertainty Estimation." arXiv, Jul. 09, 2021. Accessed: Mar. 31, 2023. [Online]. Available: http://arxiv.org/abs/2107.04212

[14] F. Museng, A. Jessica, N. Wijaya, A. Anderies, and I. A. Iswanto, "Systematic Literature Review: Toxic Comment Classification," in *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, Nov. 2022, pp. 1–7. doi: 10.1109/ICITDA55840.2022.9971338.

[15] Google Jigsaw, "Toxic Comment Classification Challenge." Accessed: Apr. 08, 2023. [Online]. Available: https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

[16] 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." arXiv, Jul. 14, 2022. Accessed: Apr. 28, 2023. [Online]. Available: http://arxiv.org/abs/2203.09509

[17] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "HateBERT: Retraining BERT for Abusive Language Detection in English," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 17–25. doi: 10.18653/v1/2021.woah-1.3.

[18] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving Google's Perspective API Built for Detecting Toxic Comments." arXiv, Feb. 26, 2017. doi: 10.48550/arXiv.1702.08138.

[19] M. Díaz *et al.*, "CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul Republic of Korea: ACM, Jun. 2022, pp. 2342–2351. doi: 10.1145/3531146.3534647.

[20] J. Dzieza, "AI is a lot of work," New York Magazine: Intelligencer. Accessed: Nov. 27, 2023. [Online]. Available: https://nymag.com/intelligencer/article/ai-artificial-intelligence-humans-technology-business-factory.html

[21] I. Olasov, “Offensive political dog whistles: you know them when you hear them. Or do you?,” Vox. Accessed: Nov. 27, 2023. [Online]. Available: https://www.vox.com/the-big-idea/2016/11/7/13549154/dog-whistles-campaign-racism
[22] D. Dessì, D. R. Recupero, and H. Sack, “An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection within Online Textual Comments,” *Electronics*, vol. 10, no. 7, Art. no. 7, Jan. 2021, doi: 10.3390/electronics10070779.
[23] H. Adam, A. Balagopalan, E. Alsentzer, F. Christia, and M. Ghassemi, “Mitigating the impact of biased artificial intelligence in emergency decision-making,” *Commun. Med.*, vol. 2, no. 1, Art. no. 1, Nov. 2022, doi: 10.1038/s43856-022-00214-4.
[24] S. Ali *et al.*, “Understanding the Effect of Deplatforming on Social Networks,” in *13th ACM Web Science Conference 2021*, in WebSci ’21. New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 187–195. doi: 10.1145/3447535.3462637.
[25] S. Jhaver, D. S. Appling, E. Gilbert, and A. Bruckman, “‘Did You Suspect the Post Would be Removed?’: Understanding User Reactions to Content Removals on Reddit,” *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, p. 192:1-192:33, Nov. 2019, doi: 10.1145/3359294.
[26] B. Vidgen and L. Derczynski, “Directions in abusive language training data, a systematic review: Garbage in, garbage out,” *PLOS ONE*, vol. 15, no. 12, p. e0243300, Dec. 2020, doi: 10.1371/journal.pone.0243300.
[27] A.-M. Founta *et al.*, “Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior.” arXiv, Apr. 15, 2018. Accessed: Sep. 20, 2022. [Online]. Available: http://arxiv.org/abs/1802.00393
[28] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, “Training deep neural networks on imbalanced data sets,” in *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada: IEEE, Jul. 2016, pp. 4368–4374. doi: 10.1109/IJCNN.2016.7727770.
[29] D. R. Beddiar, M. S. Jahan, and M. Oussalah, “Data expansion using back translation and paraphrasing for hate speech detection,” *Online Soc. Netw. Media*, vol. 24, p. 100153, Jul. 2021, doi: 10.1016/j.osnem.2021.100153.
[30] C. Newton, “The secret lives of Facebook moderators in America,” The Verge. Accessed: Oct. 29, 2023. [Online]. Available: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona
[31] V. Elliott and T. Parmar, “The despair and darkness of people will get to you,” Rest of World. Accessed: Oct. 29, 2023. [Online]. Available: https://restofworld.org/2020/facebook-international-content-moderators/
[32] S. Moscone, “The Beauty of Wattpad,” Journal. Accessed: Oct. 29, 2023. [Online]. Available: https://vocal.media/journal/the-beauty-of-wattpad
[33] L. Aratani, “Concern as US media hit with wave of layoffs amid rise of disinformation,” *The Guardian*, Dec. 10, 2022. Accessed: Oct. 29, 2023. [Online]. Available: https://www.theguardian.com/media/2022/dec/10/media-layoffs-cnn-buzzfeed-gannett-recount-protocol
[34] P. Suciu, “X Is The Biggest Source Of Fake News And Disinformation, EU Warns,” Forbes. Accessed: Oct. 29, 2023. [Online]. Available: https://www.forbes.com/sites/petersuciu/2023/09/26/x-is-the-biggest-source-of-fake-news-and-disinformation-eu-warns/
[35] J. Cohen and U. of S. California, “Analysis finds hate speech has significantly increased on Twitter.” Accessed: Oct. 29, 2023. [Online]. Available: https://phys.org/news/2023-04-analysis-speech-significantly-twitter.html