Debiasing Hate Speech Classification Models for Queer Language Through Keyword Analysis

D.S. Yahathugoda*, Rupika Wijesinghe, Ruvan Weerasinghe
University of Colombo School of Computing, Sri Lanka

***Abstract*—This article uses words or language that is considered profane, vulgar, or offensive by some readers.**

Detecting hate speech is critical for moderating harmful content on social media and the Internet. However, existing models often struggle to accurately identify hate speech targeting queer communities due to inherent biases in training data and language usage. This research explores debiasing techniques for hate speech classification models, with a focus on queer language via keyword analysis. By analyzing established hate speech datasets and queer-specific linguistic traits, this study aims to identify words and phrases the models pay attention to the most and apply different debiasing approaches such as reweighting and adversarial debiasing to enhance the efficacy and equity of hate speech aimed at queer communities, without unfairly silencing queer voices. We found that these methods improved the accuracy of queer-specific datasets but showed a decrease in performance on more general datasets. These findings suggest that we must develop more community-specific models to safeguard them from harmful content. This research contributes to advancing the understanding of bias in hate speech detection models and provides practical guidance for devising more inclusive and fair classification systems for online content moderation.

***Index Terms*—Hate Speech Detection, Algorithmic Bias, LGBT, Queer.**

I. Introduction

Historically, marginalized people have faced oppression and discrimination in many aspects of their lives, such as education, employment, health care and family. These experiences can cause stress and trauma, which can affect their mental health and well-being [1]. This particularly applies to the queer (also referred to as LGBTQ+) people.

With the rise of the internet and social media, they have offered a safe space for the people in these groups to be themselves and share their thoughts. However, online harassment is a common issue for queer people, as evident by a study conducted in 2022, which found that 66% of LGBTQ+ users experienced harassment online [2] and another study which [3] found that 74% of queer youth in the UK reported being bullied online because of their sexual orientation or gender identity in 2017. Online abuse and violence also have serious consequences for queer people's mental health and well-being. For instance, a study found that online harassment was associated with lower levels of self-esteem and life satisfaction among queer adults[4].

Given the prevalence and impact of online abuse and violence against queer people, it is important to develop effective strategies to prevent and combat this problem. One of the possible strategies is to use automated systems that can detect and moderate toxic content online. Toxicity detection is a natural language processing task that aims to classify text as toxic or non-toxic based on predefined criteria. Toxicity detection systems can help online platforms flag and remove harmful content, warn or ban abusive users, and promote healthy and respectful online interactions. However, some of their voices are silenced by the prevalent biases in automatic content moderation tools themselves, due to them not being fairly represented in the process of the making of these systems[5]. Even though there have been many efforts to alleviate these biases, particularly when it comes to racial identities and women, debiasing in regards to the queer community largely remains unexplored[6]. In this research, we aim to identify how we can reduce the biases in these tools against the queer community.

II. Related work

One of the domains where queer people encounter significant stressors is the online environment. The internet can offer many benefits for queer people, such as access to information, support networks, self-expression, identity exploration, community building, activism, etc. [7]. However, it can also expose them to various forms of online abuse and violence that target their sexual orientation and/or gender identity[2][3].

A study found that cyberbullying victimization was associated with higher levels of depression and suicidal ideation among queer youth[8]. Another study [9] found that online harassment was associated with lower levels of self-esteem and life satisfaction among queer adults.

Given the prevalence and impact of online abuse and violence against queer people, it is imperative to develop effective strategies to prevent and combat this problem. One of the possible strategies is to use automated systems that can detect and moderate toxic content online. Toxicity detection is a natural language processing task that aims to classify text as toxic or non-toxic based on predefined criteria. Toxicity detection systems can help online platforms flag and remove harmful content, warn or ban abusive users, and promote healthy and respectful online interactions.

Correspondence: Dilhara Yahathugoda (E-mail: dilharasavinday@gmail.com)
Received: 16-06-2024 **Revised:** 12-08-2024 **Accepted:** 09-09-2024
Dilhara Yahathugoda, Rupika Wijesinghe and Ruvan Weerasinghe are from University of Colombo School of Computing (dilharasavinday@gmail.com, crw@ucsc.cmb.ac.lk, arw@ucsc.cmb.ac.lk)
DOI: https://doi.org/10.4038/icter.v18i2.7296
The 2025 Special Issue contains the full papers of the abstracts published at the 24th ICTer International Conference.

A. Biases in Online Toxicity Detection Models

Toxicity detection is not a straightforward task, as it involves many nuances, complexities, and challenges. One of the main challenges is the presence of bias in toxicity detection systems, which can result in unfair or inaccurate outcomes for certain groups of users or topics. Bias can arise from various sources, such as the data used to train the systems, the algorithms used to process the data, or the human judgments used to evaluate the systems. Bias can also manifest in different ways, such as underrepresentation, overrepresentation, misrepresentation, or misclassification of certain groups or topics.

A study found that Perspective API, a widely used toxicity detection system developed by Google's Jigsaw, assigned higher toxicity scores to comments containing identity terms related to sexual orientation or gender identity (e.g., gay, lesbian, transgender) than to comments containing neutral terms (e.g., tall, happy, American) [10].

B. Keyword analysis

The keyword extraction method we have chosen for this task is the Harmonic mean of relative frequencies (HMRF) introduced in the recent paper, *Systematic keyword and bias analyses in hate speech detection*[11]. We closely follow the keyword extraction method and debiasing techniques explored in that paper. However, our paper explores bias related to queer community, where the authors explored bias with a focus on racial bias.

The score of *w* is determined by computing the harmonic mean of two relative frequencies within a set of texts *S*. Then, it utilizes the cumulative distribution function (CDF)[12] on the relative frequencies. This function, denoted as $FX(x)$, represents the probability that the random variable X is less than or equal to x. Therefor, $CDF(f_1^S)$ shows the proportion of words with a value of f_1^S that is less than or equal to $f_1^S(w)$ and $CDF(f_2^S)$ signifies the ratio of words with a value of f_2^Sthat is equal to or less than $f_2^S(w)$. Through the CDF, one can observe the position of either $f_1^S(w)$ or $f_2^S(w)$ in the distribution of words cumulatively. Equation: 1 represents the complete formula.

$$HMRF_S(w) = \frac{2 \times CDF(f_1^S(w)) \times CDF(f_2^S(w))}{CDF(f_1^S(w)) + CDF(f_2^S(w))} \quad (1)$$

In datasets used for Hate Speech Detection (HSD), we categorize texts into hateful (H) and non-hateful (N) sets. So, H and N represent each of these sets individually, with the combined set denoted as $C = H \cup N$. Therefore, C is depicted as a tuple $(HMRF_N(w), HMRF_H(w))$, where $HMRF_N(w)$ and $HMRF_H(w)$ denote the HMRFs for w within the non-hateful and hateful sets, respectively.

C. Reducing Bias

Although there aren't many examples of debiasing and evaluating hate speech detection models specifically towards the queer community, multiple papers have suggested racial and gender debiasing as well as methods to generalise debiasing as well.

A study proposed a framework to measure and mitigate the bias of toxicity detection models towards different demographic groups, such as race, gender, religion, etc[13]. They used a large-scale dataset of Wikipedia comments annotated with toxicity and demographic information and applied various debiasing techniques, such as reweighting, oversampling, and adversarial learning. They found that their framework improved the fairness and accuracy of toxicity detection models for minority groups.

AAEBERT, a pre-trained language model tailored for African American English (AAE) through the retraining of BERT-base on AAE tweets was introduced for racial debiasing[14]. Utilizing AAEBERT, they extract tweet representations from diverse hate speech datasets and conduct classification into two classes: AAE dialect and non-AAE dialect. To address biases, they employ a three-layer feedforward neural network, utilizing the representation from AAEBERT and a dialect label as inputs for adversarial debiasing.[15] also suggests an adversarial debiasing approach to effectively separate the two classes, demonstrating its effectiveness in English, Arabic, German, and Hindi. Furthermore, their method have shown improved performance over baselines, even in a multilingual setting. Our paper closely follows the adversarial method employed here, the extension being that our model looks into queer language.

III. Methodology

The Methodology chapter provides a detailed description of the data collection methods, debiasing and analytical techniques employed in the study.

A. Data Collection and Preproccesing

CivilComments dataset[16] is a large-scale dataset containing comments from the Civil Comments platform, annotated for toxicity and other attributes, including identity-based hate speech. It was released by Jigsaw (a technology incubator within Alphabet) and the Conversation AI team at Google in collaboration with the Wikimedia Foundation. The dataset aims to facilitate research on hate speech detection and toxic comment moderation.

We have chosen the CivilComments dataset particularly due to its subset of CivilCommentsIdentities[16] which contains identity labels indicating whether a comment relates to a certain identity, whether it be racial, religious, sexual or gender. This dataset also contains queer identity labels such as homosexual, bisexual and transgender. This is important in the phase where we try to measure the debiasing against these identity labels.

In this dataset, the label 'toxicity' is indicative of the overall offensiveness or hatefulness of the particular comment. However, columns such as 'identity attack', 'insult', 'obscene', 'sexual_explicit' and 'severe toxicity' also give us insight into the nature of the comment. We took in all these columns when we were encoding where more weight is given to identity attacks and threats as they are direct types of hate speech. Moreover, insult and severe toxicity are also given more weight but not as much since it is not as severe as threats

and identity attacks. Sexual explicit is given the lowest weight and obscene is ignored due to the nature of queer comments which may contain comments that are categorized under them. Finally, these values are normalized and binary encoded based on whether the final value is over 0.5 or not. Table: I represents the distribution of values in the final dataset containing the sensitive keywords.

TABLE I: Hate speech dataset

Dataset	Hate	Non-Hate	Total
Train	12470	62150	74620
Test	3076	15580	18656

Datasets relating to queer comments are relatively sparse on the internet. However, Google BigQuery hosts a dataset containing comments extracted from Reddit containing around 1.7 billion comments. This dataset also contains comments from the r/lgbt subreddit, which is the main subreddit dedicated to queer discourse, and other queer-related subreddits. It contains the number of upvotes so that we can use these values to get comments that are accepted by the users. We chose comments that had at least 5 upvotes so that we can be confident in our assumption that these comments are mostly non-hateful. We also removed comments that weren't at least three words in length. The final dataset contained 124871 comments, with 60566 queer comments and 64305 casual comments.

1) Exploring Models: BERT-based models and transformer models in general have shown significant performance in various NLP tasks including hate speech detection[17]. So the hate speech classifier taken to be debiased is a RoBERTa-based classifier introduced in 2021[18]. RoBERTa is an extension and optimization of the BERT model[19] introduced in 2019[20].

They introduce a process that combines human expertise with machine learning to dynamically create datasets and enhance the performance and resilience of hate speech detection models. This process led to a dataset comprising approximately 40,000 entries, crafted and labelled by trained annotators across four iterative phases of dynamic dataset creation. This dataset includes around 15,000 challenging variations, with each instance of hate speech meticulously annotated to specify the type and target of hate.

2) Analyse Hate Speech Classification Model: This section describes the processes followed in analysing the hate speech classification model regarding the queer-related comments retrieved in the data collection section.

The collected comments from queer subreddit were sent as input to the HSD model and relevant labels and confidence scores were added to the data frame. Table: II shows the overall results.

Total	Hate	Non-Hate	Hate Percentage
208381	16714	191667	8.02%

TABLE II: Queer comments results from the HSD model

This shows that the RoBERTa model is relatively fair to queer comments in this instance. However, we need to further analyze why the comments labelled as hateful are seen as hateful by the HSD model. For this task, we employ Transformers-Interpret[1], a Python package that is built using Captum[2], specifically tailored to interpreting transformer-based models from huggingface for various tasks.

B. Extract Keywords

With the usage of the python package released alongside the paper that introduced HMRF[3], we calculate the top-ranked words in our queer comments dataset to understand the words and phrases that were relatively more prevalent in the hateful class. Figure: 1 contains the unigrams that were extracted with HMRF.

gays, homo, bitch, ass, gay, agenda, laugh, straights, kink, queers, fuck, gaysper, christians, lesbians, queer, pussy, lesbian, cishets, dick, shit, gayest, lgbt, het, kinks, bitches, homie, hell, suddenlygay, butch, transphobes, christian, straight, parade, religion, cis, bi, jesus, frogs, homos, gaymers, chuds, fil, gayness, black, fuckin, bisexual, white, stop, chud, homosexual

Fig. 1: Extracted Unigrams

Figure: 1 shows that there are queer identity terms such as gay, lesbian, bi and lgbt are common among the comments decided to be toxic by the system, which is on par with the previous studies that have carried out similar analysis. They also contain words such as religion, transphobes and cishets (short for cisgender heterosexual) which are often found in the discussions of stressful and discriminatory experiences among queer people. Words such as gayest, gaymers and bitch are commonly used in humorous or sarcastic ways, are also represented in these results. This further points to the system's lack of understanding of context and humour, which is a commonly raised issue regarding HSD models.

C. Train a Model to Identify Queer Language and Contexts

Using the queer comments we collected, a BERT model is trained to identify queer language. The model was trained on Google Colab with a V100 GPU with a learning rate of 2e-5 and batch size of 16. The model was able to achieve an accuracy of 84% after two epochs and we have hosted it in a huggingface repository[4] for further use. It should be noted that the validation set was roughly balanced for both classes so that the accuracy metric is appropriate. However, there is room for improvement in this model by using more data as well as looking into one-class classification due to the nature of language[21].

[1] https://pypi.org/project/transformers-interpret/
[2] https://captum.ai/
[3] https://pypi.org/project/hmrf/
[4] https://huggingface.co/savinda99/queer-bert

D. Perform Debiasing

1) Debias Using Hateful Comments Without Keywords: The dataset with hateful comments that didn't contain keywords was trained on the hate speech model. The idea is that, since one of the highest contributors to bias in HSD is the class imbalances, so by training the model on the new dataset that doesn't contain sensitive words but is hateful, we are performing a reweighting method on the existing model.

2) Debias Using Hate and Non-Hate Comments With Keywords: Using the dataset containing sensitive keywords, we train the base HSD model to see the results. The idea of giving this dataset is that, since the model had associated neutral and identity-related words with the hateful class, we are training the model to further learn how to differentiate the usage of these words across the two classes

3) Adversarial Debiasing: The adversary aims to optimize the equation $L = L_C - \alpha L_A$ as its objective. Here, L_C represents the classifier loss, which is the HSD models loss, L_A represents the adversary loss, which is the queer language model trained in Section III-C, and α is a hyperparameter that regulates the balance between maximizing the adversary and minimizing the classifier. The objective is for the classifier to effectively predict hate speech while minimizing the adversary's ability to predict queer language from the inputs.

The adversary network is composed of three layers of a feedforward neural network, illustrated by Figure: 2, employing a Leaky ReLU activation function[22]. The first layer consists of 256 neurons, followed by a layer with 100 neurons, and finally, an output layer with two neurons utilizing a softmax function. Its objective is to classify the comments into queer or non-queer attributes, employing the cross-entropy loss. We used 0.03 for the α values for the final model since we observed that a higher value led to high attacks on the model, which gave us predictions that were heavily skewed to the non-hateful label. We have given two types of inputs for the adversary and they are described in below subsections.

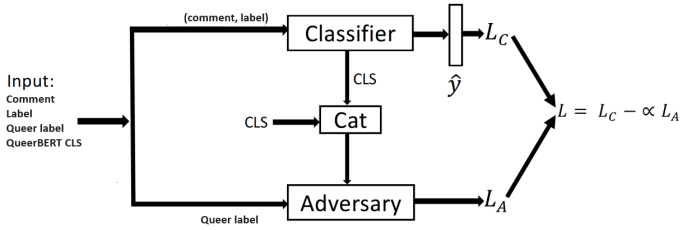

Fig. 2: Adversary Network with QueerBERT

- The adversary is given the outputs from the last layer of the HSD model for each comment, which includes the contextualized embeddings for each token in the input sequence. Then the adversary tries to predict if it's queer-related or not.
- Instead of providing the output from the last layer of the classification model, here the adversary is given a fused representation. The representation from the HSD model is combined with the representation from the queer-language model before sending it to the adversary.

IV. Results & Discussion

To measure each of these models against the identified metrics, we employ the test set created with hate speech data. For the bias calculation, we use the implementation provided in the Jigsaw classification challenge[5]. The final model score is determined by combining the overall AUC with the generalized mean of the Bias AUCs using Equation: 2

$$\text{score} = w_0 \times \text{AUC}_{\text{overall}} + \sum_{a=1}^{A} w_a M_p(m_{s,a}) \qquad (2)$$

- A is the number of sub metrics (3).
- $m_{s,a}$ represents the bias metric for identity subgroup s using submetric a.
- w_a denotes a weighting for the relative importance of each sub metric, with all four w values set to 0.25.

Table: III illustrates the overall measures for each of the models we have trained.

Trained method	F1	Accuracy	Bias score
Base	0.330	0.739	0.592
Without keywords	0.344	0.645	0.606
With keywords	0.820	0.942	0.881
Adversarial without dialect	0.492	0.885	0.698
Adversarial with dialect	0.518	0.885	0.715

TABLE III: Overall results for different debiased models

We can observe that all of the debiasing methods have improved the base model in regards to its F1, accuracy and bias metric. The low F1 score for the base model is expected since we are especially focusing on a dataset containing queer comments, which is a weak point in hate speech detection models as we discussed in Section II. We can also observe that training the model with comments that contain keywords which were sensitive to the model, has the most improvement out of the four methods. This is expected since the model gained new information on the correct usage of these words in actual hate speech and queer-language usage. Even though the values for the model trained on hateful comments without keywords have slightly better scores than the base model, part of the gap between the keyword-related model can be explained since the keyword-related model was trained on a similar subset to the test set which contained these keywords.

Both of the adversarial debiasing methods have shown significant improvement over the base model, which is on par with findings on other studies[14]. It should be noted that the adversarial without dialect model has better scores than the adversarial with dialect model, this can have several reasons. One is that the BERT models already contain the needed information regarding each token and that combining the

[5]https://www.kaggle.com/code/dborkan/benchmark-kernel

last layer representations of the queer-language classification model adds more complexity to it, resulting in a slightly lower score. However, as we observed, the queer-language model was not as accurate in its classification, this can also manifest in the final scores being lower due to inaccurate embeddings.

To measure the out-of-domain performance of our models, we employ the OLEA (Offensive Language Error Analysis)[6] python package, which implements bias metrics for the COLD (Complex Offensive Language Dataset) dataset[23]. The results show that adversarial learning was more effective in identifying the correct usage of reclaimed slurs, showing a significant improvement over the base models and reweighted methods. This is expected because the adversary specifically targets these words within a certain context during the debiasing process. However, it was at the cost of identifying hateful comments that contained these slurs. This is consistent with the earlier bias values where the adversary was more polarizing compared to other methods.

We also inferred the models with the benchmarking hate speech dataset HateCheck, a collection of practical assessments designed for evaluating hate speech detection systems[24]. The results are shown in Table: IV

Model	F1 Score
Base	0.968
Without Keywords	0.965
With Keywords	0.879
Adversarial without Dialect	0.632
Adversary with Dialect	0.632

TABLE IV: F1 scores for HateCheck

The HateCheck dataset scored the without-keywords model as best based on their F1 values. This means that it could have more generalizability than others. However, the without keywords model also performed adequately on this dataset which we can take as a compromise of the two. Since all the F1 scores were lower than the base model, it can suggest that fine-tuning for a specific subset, here the queer community, could lead to lower generalizability. There is also a measurable variance between test sets between different datasets identified in other studies that may factor into this [25].

V. Future Work

Even though we only looked into reweighting and adversarial learning using sensitive keywords, there are a multitude of various debiasing methods are were carried out in studies. We can further look into how those methods compare against the ones we evaluated.

As we measured, the queer language model is not as accurate as it could be. So, further exploring how we can identify queer languages and how we can differentiate them from regular English can further enhance our understanding of them as well as potentially provide more specific embedding during adversarial learning.

As we only performed the analysis on a queer comments dataset that was posted over five months, there can be differences between datasets due to the evolving nature of language. Conducting longitudinal studies to track changes in hate speech patterns and model performance over time can yield insights into the effectiveness and durability of debiasing interventions. Understanding how societal attitudes and norms evolve can inform the development of more resilient hate speech detection systems.

[6]https://pypi.org/project/olea/

VI. Conclusion

The analysis indicates that all debiasing methods enhanced the base model's performance in terms of F1 score, accuracy, and bias metrics. Training the model with comments containing sensitive keywords demonstrated the most improvement, While models trained on hateful comments without keywords showed slightly better scores than the base model.

Both adversarial debiasing methods exhibited significant improvement over the base model. However, the adversarial without dialect model outperformed the model with dialect, possibly due to the added complexity of combining representations of the queer-language classification model.

In conclusion, addressing bias in hate speech detection models targeting the queer community requires a multifaceted approach that combines technical innovation, interdisciplinary collaboration, and ethical considerations. By continuing to refine debiasing techniques, expand dataset representation, and engage stakeholders, we can work towards more equitable and effective solutions for combating online hate speech.

References

[1] I. H. Meyer, "Resilience in the study of minority stress and health of sexual and gender minorities." *Psychology of Sexual Orientation and Gender Diversity*, vol. 2, no. 3, p. 209, 2015.

[2] "Social Media Safety Program — GLAAD — glaad.org," https://glaad.org/smsi/lgbtq-social-media-safety-program/, [Accessed 16-09-2023].

[3] I. Matters, "Ditch the Label Annual bullying survey 2017 — Internet Matters — internetmatters.org," https://www.internetmatters.org/hub/research/ditch-the-label-annual-bullying-survey-2017/, [Accessed 16-09-2023].

[4] A. Lenhart, M. Ybarra, K. Zickuhr, and M. Price-Feeney, "Online harassment, digital stalking and cyberabuse in america," 2016.

[5] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on bert model," *PloS one*, vol. 15, no. 8, p. e0237861, 2020.

[6] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling bias in toxic speech detection: A survey," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–32, 2023.

[7] S. L. Craig, L. McInroy, L. T. McCready, and R. Alaggia, "Media: A catalyst for resilience in lesbian, gay, bisexual, transgender, and queer youth," *Journal of LGBT Youth*, vol. 12, no. 3, pp. 254–275, 2015.

[8] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth." *Psychological bulletin*, vol. 140, no. 4, p. 1073, 2014.

[9] F. Stevens, J. R. Nurse, and B. Arief, "Cyber stalking, cyber harassment, and adult mental health: A systematic review," *Cyberpsychology, Behavior, and Social Networking*, vol. 24, no. 6, pp. 367–376, 2021.

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

[11] G. L. De la Peña Sarracén and P. Rosso, "Systematic keyword and bias analyses in hate speech detection," *Information Processing & Management*, vol. 60, no. 5, p. 103433, 2023.

[12] F. GABBIANI and S. J. COX, "Chapter 11 - probability and random variables," in *Mathematics for Neuroscientists*, F. GABBIANI and S. J. COX, Eds. London: Academic Press, 2010, pp. 155–173. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123748829000113

[13] J. H. Park, J. Shin, and P. Fung, "Reducing gender bias in abusive language detection," *arXiv preprint arXiv:1808.07231*, 2018.

[14] E. Okpala, L. Cheng, N. Mbwambo, and F. Luo, "Aaebert: Debiasing bert-based hate speech detection models via adversarial learning," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2022, pp. 1606–1612.

[15] S. Yuan, A. Maronikolakis, and H. Schütze, "Separating hate speech and offensive language classes via adversarial debiasing," in *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 2022, pp. 1–10.

[16] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, "Nuanced metrics for measuring unintended bias with real data for text classification," *CoRR*, vol. abs/1903.04561, 2019. [Online]. Available: http://arxiv.org/abs/1903.04561

[17] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*. Springer, 2020, pp. 928–940.

[18] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the worst: Dynamically generated datasets to improve online hate detection," in *ACL*, 2021.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[21] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

[22] C. Banerjee, T. Mukherjee, and E. Pasiliao Jr, "An empirical study on generalizations of the relu activation function," in *Proceedings of the 2019 ACM Southeast Conference*, 2019, pp. 164–167.

[23] A. Palmer, C. Carr, M. Robinson, and J. Sanders, "Cold: Annotation scheme and evaluation data set for complex offensive language in english," *Journal for Language Technology and Computational Linguistics*, vol. 34, no. 1, pp. 1–28, 2020.

[24] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. B. Pierrehumbert, "Hatecheck: Functional tests for hate speech detection models," *arXiv preprint arXiv:2012.15606*, 2020.

[25] S. D. Swamy, A. Jamatia, and B. Gambäck, "Studying generalisability across abusive language detection datasets," in *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, 2019, pp. 940–950.