Demoting Racial Bias in Hate Speech Detection

Mengzhou Xia Anjalie Field Yulia Tsvetkov
Language Technologies Institute
Carnegie Mellon University
{mengzhox,anjalief,ytsvetko}@cs.cmu.edu

Abstract

In current hate speech datasets, there exists a high correlation between annotators' perceptions of toxicity and signals of African American English (AAE). This bias in annotated training data and the tendency of machine learning models to amplify it cause AAE text to often be mislabeled as abusive/offensive/hate speech with a high false positive rate by current hate speech classifiers. In this paper, we use adversarial training to mitigate this bias, introducing a hate speech classifier that learns to detect toxic sentences while demoting confounds corresponding to AAE texts. Experimental results on a hate speech dataset and an AAE dataset suggest that our method is able to substantially reduce the false positive rate for AAE text while only minimally affecting the performance of hate speech classification.

1 Introduction

The prevalence of toxic comments on social media and the mental toll on human moderators has generated much interest in automated systems for detecting hate speech and abusive language (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018), especially language that targets particular social groups (Silva et al., 2016; Mondal et al., 2017; Mathew et al., 2019). However, deploying these systems without careful consideration of social context can increase bias, marginalization, and exclusion (Bender and Friedman, 2018; Waseem and Hovy, 2016).

Most datasets currently used to train hate speech classifiers were collected through crowdsourced annotations (Davidson et al., 2017; Founta et al., 2018), despite the risk of annotator bias. Waseem (2016) show that non-experts are more likely to label text as abusive than expert annotators, and Sap et al. (2019) show how lack of social context in annotation tasks further increases the risk of annotator bias, which can in turn lead to the marginalization of racial minorities. More specifically, annotators are more likely to label comments as abusive if they are written in African American English (AAE). These comments are assumed to be incorrectly labelled, as annotators do not mark them as abusive if they are properly primed with dialect and race information (Sap et al., 2019).

These biases in annotations are absorbed and amplified by automated classifiers. Classifiers trained on biased annotations are more likely to incorrectly label AAE text as abusive than non-AAE text: the false positive rate (FPR) is higher for AAE text, which risks further suppressing an already marginalized community. More formally, the disparity in FPR between groups is a violation of the Equality of Opportunity criterion, a commonly used metric of algorithmic fairness whose violation indicates discrimination (Hardt et al., 2016). According to Sap et al. (2019), the false positive rate for hate speech/abusive language of the AAE dialect can reach as high as 46%.

Thus, Sap et al. (2019) reveal two related issues in the task of hate speech classification: the first is biases in existing annotations, and the second is model tendencies to absorb and even amplify biases from spurious correlations present in datasets (Zhao et al., 2017; Lloyd, 2018). While current datasets can be re-annotated, this process is time-consuming and expensive. Furthermore, even with perfect annotations, current hate speech detection models may still learn and amplify spurious correlations between AAE and abusive language (Zhao et al., 2017; Lloyd, 2018).

In this work, we present an adversarial approach to mitigating the risk of racial bias in hate speech classifiers, even when there might be annotation bias in the underlying training data. In §2, we describe our methodology in general terms, as it can be useful in any text classification task that seeks

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 7–14
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17

to predict a target attribute (here, toxicity) without basing predictions on a protected attribute (here, AAE). Although we aim at preserving the utility of classification models, our primary goal is not to improve the raw performance over predicting the target attribute (hate speech detection), but rather to reduce the influence of the protected attribute.

In §3 and §4, we evaluate how well our approach reduces the risk of racial bias in hate speech classification by measuring the FPR of AAE text, i.e., how often the model incorrectly labels AAE text as abusive. We evaluate our methodology using two types of data: (1) a dataset inferred to be AAE using demographic information (Blodgett et al., 2016), and (2) datasets annotated for hate speech (Davidson et al., 2017; Founta et al., 2018) where we automatically infer AAE dialect and then demote indicators of AAE in corresponding hate speech classifiers. Overall, our approach decreases the dialectal information encoded by the hate speech model, leading to a 2.2–3.2 percent reduction in FPR for AAE text, without sacrificing the utility of hate speech classification.

2 Methodology

Our goal is to train a model that can predict a target attribute (abusive or not abusive language), but that does not base decisions off of confounds in data that result from protected attributes (e.g., AAE dialect). In order to achieve this, we use an adversarial objective, which discourages the model from encoding information about the protected attribute. Adversarial training is widely known for successfully adapting models to learn representations that are invariant to undesired attributes, such as demographics and topics, though they rarely disentangle attributes completely (Li et al., 2018; Elazar and Goldberg, 2018; Kumar et al., 2019; Lample et al., 2019; Landeiro et al., 2019).

Model Architecture Our demotion model consists of three parts: 1) An encoder H that encodes the text into a high dimensional space; 2) A binary classifier C that predicts the target attribute from the input text; 3) An adversary D that predicts the protected attribute from the input text. We used a single-layer bidirectional LSTM encoder with an attention mechanism. Both classifiers are two-layer MLPs with a $\tanh$ activation function.

Training Procedure Each data point in our training set is a triplet $\{(x_i, y_i, z_i); i \in 1 \dots N\}$, where x_i is the input text, y_i is the label for the target attribute and z_i is label of the protected attribute. The (x_i, y_i) tuples are used to train the classifier C, and the (x_i, z_i) tuple is used to train the adversary D.

We adapt a two-phase training procedure from Kumar et al. (2019). We use this procedure because Kumar et al. (2019) show that their model is more effective than alternatives in a setting similar to ours, where the lexical indicators of the target and protected attributes are closely connected (e.g., words that are common in non-abusive AAE and are also common in abusive language datasets). In the first phase (pre-training), we use the standard supervised training objective to update encoder H and classifier C:

$$\min_{C,H} \sum_{i=1}^{N} \mathcal{L}(C(H(x_i)), y_i) \tag{1}$$

After pre-training, the encoder should encode all relevant information that is useful for predicting the target attribute, including information predictive of the protected attribute.

In the second phase, starting from the best-performing checkpoint in the pre-training phase, we alternate training the adversary D with Equation 2 and the other two models (H and C) with Equation 3:

$$\min_{D} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(D(H(x_i)), z_i) \tag{2}$$

$$\min_{H,C} \frac{1}{N} \sum_{i=1}^{N} \alpha \cdot \mathcal{L}(C(H(x_i)), y_i) + (1-\alpha) \cdot \mathcal{L}(D(H(x_i)), 0.5) \tag{3}$$

Unlike Kumar et al. (2019), we introduce a hyper-parameter α, which controls the balance between the two loss terms in Equation 3. We find that α is crucial for correctly training the model (we detail this in §3).

We first train the adversary to predict the protected attribute from the text representations outputted by the encoder. We then train the encoder to "fool" the adversary by generating representations that will cause the adversary to output random guesses, rather than accurate predictions. At the same time, we train the classifier to predict the target attribute from the encoder output.

Dataset	Example
Founta et al. (2018)	I am hungry and I am dirty as hell bruh, need dat shower and dem calories
Blodgett et al. (2016)	so much energy and time wasted hatin on someone when alla that coulda been put towards makin yourself better.... a... https://t.co/awCg1nCt8t

Table 1: Example from Founta et al. (2018) and Blodgett et al. (2016) where the state-of-the-art model misclassifies innocuous tweets (inferred to be AAE) as abusive language. Our model correctly classifies these tweets as non-toxic.

3 Experiments

3.1 Dataset

To the best of our knowledge, there are no datasets that are annotated both for toxicity and for AAE dialect. Instead, we use two toxicity datasets and one English dialect dataset that are all from the same domain (Twitter):

DWMW17 (Davidson et al., 2017) A Twitter dataset that contains 25K tweets annotated as *hate speech*, *offensive*, or *none*. The authors define hate speech as language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group, and offensive language as language that contains offensive terms which are not necessarily inappropriate.

FDCL18 (Founta et al., 2018) A Twitter dataset that contains 100K tweets annotated as *hateful*, *abusive*, *spam* or *none*. This labeling scheme was determined by conducting multiple rounds of crowdsourcing to understand how crowdworkers use different labels. Strongly impolite, rude, or hurtful language is considered abusive, and the definition of hate speech is the same as in DWMW17.

BROD16 (Blodgett et al., 2016) A 20K sample out of a 1.15M English tweet corpus that is demographically associated with African American twitter users. Further analysis shows that the dataset contains significant linguistic features of African American English.

In order to obtain dialect labels for the DWMW17 and FDCL18, we use an off-the-shelf demographically-aligned ensemble model (Blodgett et al., 2016) which learns a posterior topic distribution (topics corresponding to African American, Hispanic, White and Other) at a user, message, and word level. Blodgett et al. (2016) generate a AAE-aligned corpus comprising tweets from users labelled with at least 80% posterior probability as using AAE-associated terms. Similarly, following Sap et al. (2019), we assign AAE label to tweets with at least 80% posterior probability of containing AAE-associated terms at the message level and consider all other tweets as Non-AAE.

In order to obtain toxicity labels for the BROD16 dataset, we consider all tweets in this dataset to be non-toxic. This is a reasonable assumption since hate speech is relatively rare compared to the large amount of non-abusive language on social media (Founta et al., 2018).[1]

3.2 Training Parameters

In the pre-training phase, we train the model until convergence and pick the best-performing checkpoint for fine-tuning. In the fine-tuning phase, we alternate training one single adversary and the classification model each for two epochs in one round and train for 10 rounds in total.

We additionally tuned the α parameter used to weight the loss terms in Equation 3 over validation sets. We found that the value of α is important for obtaining text representations containing less dialectal information. A large α easily leads to over-fitting and a drastic drop in validation accuracy for hate speech classification. However, a near zero α severely reduces both training and validation accuracy. We ultimately set $\alpha = 0.05$.

We use the same architecture as Sap et al. (2019) as a baseline model, which does not contain an adversarial objective. For both of this baseline model and our model, because of the goal of demoting the influence of AAE markers, we select the model with the lowest false positive rate on validation set. We train models on both DWMW17 and FDCL18 datasets, which we split into train/dev/test subsets following Sap et al. (2019).

[1] We additionally did a simple check for abusive terms using a list of 20 hate speech words, randomly selected from `Hatebase.org`. We found that the percentage of sentences containing these words is much lower in AAE dataset ($\approx 2\%$) than hate speech datasets ($\approx 20\%$).

Dataset	Accuracy		F1	
	base	ours	base	ours
DWMW17	**91.90**	90.68	75.15	**76.05**
FDCL18	**81.18**	80.27	66.15	**66.80**

Table 2: Accuracy and F1 scores for detecting abusive language. F1 values are macro-averaged across all classification categories (e.g. hate, offensive, none for DWMW17). Our model achieves an accuracy and F1 on par with the baseline model.

	Offensive		Hate	
	base	ours	base	ours
FDCL18-AAE	20.94	**17.69**	3.23	**2.60**
BROD16	16.44	**14.29**	5.03	**4.52**

Table 3: False positive rates (FPR), indicating how often AAE text is incorrectly classified as hateful or abusive, when training with the FDCL18 dataset. Our model consistently improves FPR for offensiveness, and performs slightly better than the baseline for hate speech detection.

4 Results and Analysis

Table 2 reports accuracy and F1 scores over the hate speech classification task. Despite the adversarial component in our model, which makes this task more difficult, our model achieves comparable accuracy as the baseline and even improves F1 score. Furthermore, the results of our baseline model are on par with those reported in Sap et al. (2019), which verifies the validity of our implementation.

Next, we assess how well our demotion model reduces the false positive rate in AAE text in two ways: (1) we use our trained hate speech detection model to classify text inferred as AAE in BROD16 dataset, in which we assume there is no hateful or offensive speech and (2) we use our trained hate speech detection model to classify the test partitions of the DWMW17 and FDCL18 datasets, which are annotated for hateful and offensive speech and for which we use an off-the-shelf model to infer dialect, as described in §3. Thus, for both evaluation criteria, we have or infer AAE labels and toxicity labels, and we can compute how often text inferred as AAE is misclassified as hateful, abusive, or offensive.

Notably, Sap et al. (2019) show that datasets that annotate text for hate speech without sufficient context—like DWMW17 and FDCL18—may suffer from inaccurate annotations, in that annotators

	Offensive		Hate	
	base	ours	base	ours
DWMW17-AAE	**38.27**	42.59	**0.70**	2.06
BROD16	**23.68**	24.34	**0.28**	0.83

Table 4: False positive rates (FPR), indicating how often AAE text is incorrectly classified as hateful or offensive, when training with DWMW17 dataset. Our model fails to improve FPR over the baseline, since 97% of AAE-labeled instances in the dataset are also labeled as toxic.

are more likely to label non-abusive AAE text as abusive. However, despite the risk of inaccurate annotations, we can still use these datasets to evaluate racial bias in toxicity detection because of our focus on FPR. In particular, to analyze false positives, we need to analyze the classifier's predictions of the text as toxic, when annotators labeled it as non-toxic. Sap et al. (2019) suggest that annotators over-estimate the toxicity in AAE text, meaning FPRs over the DWMW17 and FDCL18 test sets are actually lower-bounds, and the true FPR is could be even higher. Furthermore, if we assume that the DWMW17 and FDCL18 training sets contain biased annotations, as suggested by Sap et al. (2019), then a high FPR over the corresponding test sets suggests that the classification model amplifies bias in the training data, and labels non-toxic AAE text as toxic even when annotators did not.

Table 3 reports results for both evaluation criteria when we train the model on the FDCL18 data. In both cases, our model successfully reduces FPR. For abusive language detection in the FDCL18 test set, the reduction in FPR is > 3; for hate speech detection, the FPR of our model is also reduced by 0.6 compared to the baseline model. We can also observe a 2.2 and 0.5 reduction in FPR for abusive speech and hate speech respectively when evaluating on BROD16 data.

Table 4 reports results when we train the model on the DWMW17 dataset. Unlike Table 3, unfortunately, our model fails to reduce the FPR rate for both offensive and hate speech of DWMW17 data. We also notice that our model trained with DWMW17 performs much worse than the model trained with FDCL18 data.

To understand the poor performance of our model when trained and evaluated on DWMW17 data, we investigated the data distribution in the test set and found that the vast majority of tweets

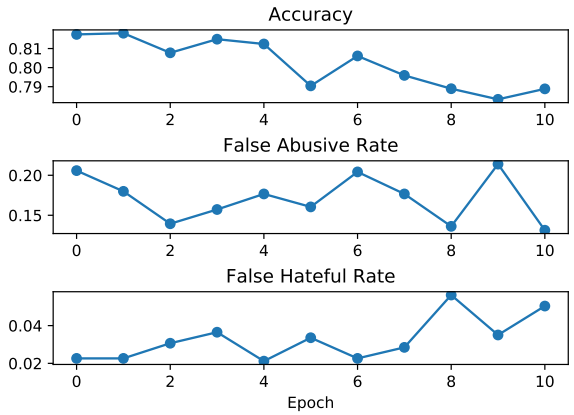

Figure 1: Accuracy of the entire development set of FDCL18 (top), and FPR rate for abusive (middle) and hate (bottom) speech detection for tweets inferred as AAE in the development set. X axis denotes the number of epochs. 0th epoch is the best checkpoint for pre-training step, which is also the baseline model.

labeled as AAE by the dialect classifier were also annotated as toxic (97%). Thus, the subset of the data over which our model might improve FPR consists of merely $< 3\%$ of the AAE portion of the test set (49 tweets). In comparison, 70.98% of the tweets in the FDCL18 test set that were labeled as AAE were also annotated as toxic. Thus, we hypothesize that the performance of our model over the DWMW17 test set is not a representative estimate of how well our model reduces bias, because the improvable set in the DWMW17 is too small.

In Table 1, we provide two examples of tweets that the baseline classifier misclassifies abusive/offensive, but our model, correctly classifies as non-toxic. Both examples are drawn from a toxicity dataset and are classified as AAE by the dialectal prediction model.

Trade-off between FPR and Accuracy In order to better understand model performance, we explored the accuracy and FPR of our model throughout the entire training process. We evaluate the best checkpoint of the pre-trained model (0^{th} epoch) and checkpoints of each epoch during adversarial training and show the results in Figure 1. While the baseline model (0^{th} epoch, before any adversarial training) achieves high accuracy, it also has a high FPR rate, particularly over abusive language. After adversarial training, the FPR rate decreases with only minor changes in accuracy. However, checkpoints with lower FPR rates also often have lower accuracy. While Tables 2 and 3 suggest that our model does achieve a balance between these metrics, Figure 1 shows the difficulty of this task; that is, it is difficult to disentangle these attributes completely.

Eliminatation of protected attribute In Figure 2, we plot the validation accuracy of the adversary through the entire training process in order to verify that our model does learn a text representation at least partially free of dialectal information. Further, we compare using one adversary during training with using multiple adversaries (Kumar et al., 2019). Through the course of training, the validation accuracy of AAE prediction decreases by about 6–10 and 2–5 points for both datasets, indicating that dialectal information is gradually removed from the encoded representation. However, after a certain training threshold (6 epochs for DWMW17 and 8 epochs for FDCL18), the accuracy of the classifier (not shown) also drops drastically, indicating that dialectal information cannot be completely eliminated from the text representation without also decreasing the accuracy of hate-speech classification. Multiple adversaries generally cause a greater decrease in AAE prediction than a single adversary, but do not necessarily lead to a lower FPR and a higher classification accuracy. We attribute this to the difference in experimental setups: in our settings, we focus on one attribute to demote, whereas Kumar et al. (2019) had to demote ten latent attributes and thus required multiple adversaries to stabilize the demotion model. Thus, unlike in (Kumar et al., 2019), our settings do not require multiple adversaries, and indeed, we do not see improvements from using multiple adversaries.

5 Related Work

Preventing neural models from absorbing or even amplifying unwanted artifacts present in datasets is indispensable towards building machine learning systems without unwanted biases.

One thread of work focuses on removing bias at the data level, through reducing annotator bias (Sap et al., 2019) and augmenting imbalanced datasets (Jurgens et al., 2017). Dixon et al. (2018) propose an unsupervised method based on balancing the training set and employing a proposed measurement for mitigating unintended bias in text classification models. Webster et al. (2018) present a gender-balanced dataset with ambiguous name-pair pronouns to provide diversity coverage for real-world data. In addition to annotator bias, sampling

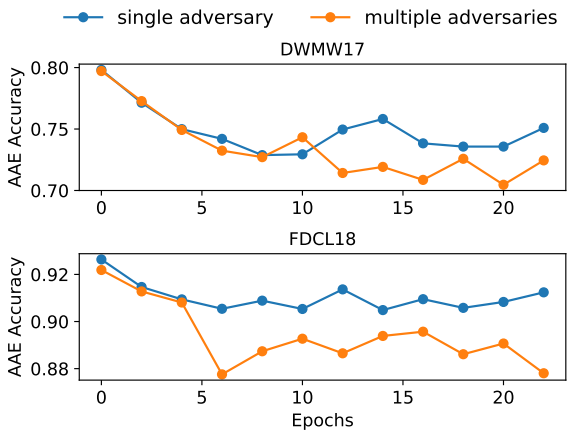

Figure 2: Validation accuracy on AAE prediction of the adversary in the whole training process. The green line denotes the training setting of one adversary and the orange line denotes the training setting of multiple adversaries.

strategies also result in topic and author bias in datasets of abusive language detection, leading to decreased classification performance when testing in more realistic settings, necessitating the adoption of cross-domain evaluation for fairness (Wiegand et al., 2019).

A related thread of work on debiasing focuses at the model level (Zhao et al., 2019). Adversarial training has been used to remove protected features from word embeddings (Xie et al., 2017; Zhang et al., 2018) and intermediate representations for both texts (Elazar and Goldberg, 2018; Zhang et al., 2018) and images (Edwards and Storkey, 2015; Wang et al., 2018). Though previous works have documented that adversarial training fails to obliterate protected features, Kumar et al. (2019) show that using multiple adversaries more effectively forces the removal.

Along similar lines, multitask learning has been adopted for learning task-invariant representations. Vaidya et al. (2019) show that multitask training on a related task e.g., identity prediction, allows the model to shift focus to toxic-related elements in hate speech detection.

6 Conclusion

In this work, we use adversarial training to demote a protected attribute (AAE dialect) when training a classifier to predict a target attribute (toxicity). While we focus on AAE dialect and toxicity, our methodology readily generalizes to other settings, such as reducing bias related to age, gender, or income-level in any other text classification task. Overall, our approach has the potential to improve fairness and reduce bias in NLP models.

7 Acknowledgements

We gratefully thank anonymous reviewers, Maarten Sap, and Dallas Card for their help with this work. The second author of this work is supported by the NSF Graduate Research Fellowship Program under Grant No. DGE1745016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We also gratefully acknowledge Public Interest Technology University Network Grant No. NVF-PITU-Carnegie Mellon University-Subgrant-009246-2019-10-01 for supporting this research.

References

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

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

Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

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

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 51–57.

Sachin Kumar, Shuly Wintner, Noah A Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4144–4154.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Virgile Landeiro, Tuan Tran, and Aron Culotta. 2019. Discovering and controlling for latent confounds in text classification using adversarial domain adaptation. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 298–305. SIAM.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Kirsten Lloyd. 2018. Bias amplification in artificial intelligence systems. *CoRR*, abs/1809.07842.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, pages 173–182. ACM.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94. ACM.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Tenth International AAAI Conference on Web and Social Media*.

Ameya Vaidya, Feng Mai, and Yue Ning. 2019. Empirical analysis of multi-task learning for reducing model bias in toxic comment detection. *arXiv preprint arXiv:1909.09758*.

Tianlu Wang, Jieyu Zhao, Kai-Wei Chang, Mark Yatskar, and Vicente Ordonez. 2018. Adversarial removal of gender from deep image representations. *arXiv preprint arXiv:1811.08489*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

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

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 585–596.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 629–634.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.