Towards Generalized Bias Mitigation

Shakil Sumon Amreeta Chatterjee Maha Jinadoss Anita Ruangrotsakun
Oregon State University
{sumons, chattera, jinadosm, ruangroc}@oregonstate.edu

Abstract

Debiasing in Natural Language Processing has primarily focused on removing bias from word vector representations generated by models such as Glove, BERT, and ElMo, which takes a significant amount of time and computing resources. The limited research on debiasing in downstream tasks has mainly focused on a single bias dimension, which is frequently not transferable to other bias dimensions. This is the gap we aim to fill, by introducing a generalized adversarial technique for debiasing downstream tasks. We have worked with the task of hate speech detection and experimented with three bias dimensions namely gender, race and religion. Based on our preliminary investigation, it appears that adversarial training has the potential to serve as a generalized method for mitigating bias in downstream NLP tasks. Moreover, our experience has revealed that the currently available fairness metrics lack consistency with each other, which raises the need for the development of more generalized fairness metrics to advance research in this field.

1. Introduction

Deep learning models achieved tremendous success in various natural language processing (NLP) tasks such as text classification [17], question answering [15], machine translation [26] and, natural language understanding [30]. Due to the impressive performance gain over the years, deep learning based NLP models are now becoming more and more commonplace in critical applications like healtcare, finance and hate speech detection [4,28]. Additionally, given the recent surge of Large Language Models (LLMs) [3, 7] and their monumental accomplishment on an array of tasks across the NLP domain and beyond, deep learning based models will be increasingly more influential in such critical applications. However, earlier research shows that deep learning models are biased towards certain underrepresented and minority groups of the society [10, 19, 20]. Given that the society we live in is inherently biased and the models we build mirror our collective mental state, it should not appear surprising [1]. Even so, the landscape became darker as we discovered that the bias present in the data is not just preserved, sometimes it gets amplified [14]. Deployment of these models in the wild will create numerous problems including representational and allocational harms for the minority groups [19]. Therefore, the deep learning community, particularly the NLP community, has taken steps to mitigate bias in deep learning models utilized for NLP tasks [6, 8, 16, 28]. The majority of research in this field concentrates on removing biases from pretrained word embeddings [2, 23, 27], which is an extensive undertaking in terms of computation, finances, and time. A different group of researchers concentrate on mitigating biases in downstream tasks, which alleviates the need for retraining large word embedding models but requires crafting distinct models for each new task and bias dimension. Therefore, this paper introduces a generalized debiasing framework for downstream NLP tasks that preserves the word embedding models and allows for expansion to multiple bias dimensions. Our approach employs an adversarial debiasing technique which proved to be effective for single bias dimension in prior research [18, 29].

We have picked Hate Speech Detection as our task as it is a prevalent problem on online platforms and can have tangible consequences such as discrimination, harassment, and violence in the real world. The bias dimensions we included in our study are race, gender, and religion, which we believe are the three predominant dimensions in the realm of hate speech detection. Most of the research we have reviewed in this area focus solely on a single bias dimension, but we hope to experiment with multiple bias dimensions.

Our experimentation involved two kinds of adversarial debiasing architectures that we refer to as A1 and A2. A1 was based on a conventional min-max style adversarial learning whereas A2 introduced an extra hyperparameter α to regulate the equilibrium between the loss of the classifier and the adversary. Although we could not carry out exhaustive experimentation to compare the performance of the two techniques, our limited exploration suggests that A2 outperforms A1 in some fairness metrics. Moreover, we have noticed that the fairness metrics we utilized to evaluate the models do not exhibit consistency with one another. Specifically, an increase in one metric leads to a decrease in another, despite the fact that a lower score indicates better debiasing for both metrics.

2. Related Work

Because natural language is a proxy for human behavior, our society's biases are reflected in the datasets and models we create. These biases can be translated into broader social impacts as NLP systems see more real-world applications [13]. Hate speech is a widespread issue on online platforms, and it can lead to real-world consequences such as discrimination, harassment, and violence. Manual filtering of such content is difficult given the volume of internet communications, necessitating automated filtering methods – toxicity detectors. However, reducing racial and gender bias is one of the most difficult challenges in developing toxicity detectors.

Several studies have discovered racial and gender bias among modern toxicity detectors, as certain trigger words (e.g., 'gay,' 'black') are more likely to be associated with false positives (that is, nontoxic text marked as toxic). They also have a higher tendency to classify African-American English posts as offensive or hate than "white" English [6], but also more often predict false negatives on "white" than African-American English [24].

To mitigate unintended bias in machine learning models, Dixon et al. [8] proposed an unsupervised approach based on balancing the training dataset in a text classification task to detect toxicity. They counteracted bias against 51 common identity terms by strategically adding non-toxic examples. Park et al. [21] also addressed biases towards gender identity terms in abusive language datasets by fine-tuning the model with a larger and less biased dataset. They also reduced gender bias by (1) debiasing word embeddings by removing gender stereotypical information, and (2) augmenting the training data by identifying male entities and swapping them with equivalent female entities to remove correlation between gender and classification decision.

Debiasing could also be incorporated during the model training process. In order to increase model sensitivity to the context surrounding group identifiers, Kennedy et al. [16] proposed a novel approach. They used regularization during training to increase the explanation-based importance of group identifiers, forcing models to take into account the context around them. This lessens the emphasis on group identifiers and emphasizes the more generalizable characteristics of hate speech, such as dehumanizing and insulting language.

Vaidya et al. [25] proposed an attention-based multi-task learning approach to reduce unintended model biases towards commonly-attacked identities in hate speech detection. They discovered that learning multiple related tasks at the same time can help reduce biases toward certain identities while also improving the health of online conversations. However, the proposed method had limitations in terms of semantic encoding capabilities, and the model is not adaptable when new or hidden identities emerge.

Closest to our work are Xia et al. [29] and Zhang et al.'s work [31]. Xia et al. [29] introduced a hate speech classifier that learns to recognize toxic sentences while demoting confounds related to AAE texts (AAE) to reduce bias towards signals of African American English. Zhang et al. [31] proposed adversarial technique for mitigating gender bias during training. This study concentrated on supervised deep learning tasks, where the goal is to predict an outcome variable Y given an input variable X while remaining unbiased with regard to a variable Z (protected variable). For the UCI Adult dataset, they used this for income prediction and analogy completion. According to Zhang et al., the technique significantly decreased bias in the task of predicting revenue.

3. Methodology

3.1. Data Preprocessing

We preprocessed the data and prepared it for text classification. The preprocessing steps included cleaning the data by removing empty comments and replacing usernames and URLs with generic tokens in the comments. We then tokenized the comments. The target labels were converted to binary classes, and the protected attribute labels were extracted from the input data.

3.2. Performance Metrics

The accuracy and F1-Score were used as the primary measures of model performance. We also provided precision (macro, unweighted) and recall data (macro, unweighted).

3.3. Fairness Metrics

For each of the protected and nonprotected groups, we calculated the toxicity rate, true positive rate, false positive rate, demographic parity, true positive parity, false positive parity, and equalized odds. The definition provided below are based on [9, 11, 22].

- **Toxicity Rate**: Percentage of texts labelled by the classifier as toxic is defined as the toxicity rate. For a fair model, the rate of toxicity should be equal for both protected and non protected attributes
- **True positive rate** :The percentage of positive data points that are correctly classified as positive.
- **False positive rate** :The percentage of negative data points that are incorrectly classified as positive. For a

fair classifier the FPR rate will be same for different demographic groups.

- **Demographic parity** :The rate of toxicity prediction is equal for different groups.

- **Equalized odds** :When both True Positive Parity and False Positive Parity conditions are met, the equality of odds is satisfied.

- **True positive parity**:The rate at which the model predicts that toxic comments are toxic is equal for protected attributes and non protected attributes.

- **False positive parity**:The rate at which the model predicts non toxic comments as non toxic is equal for all groups

3.4. Baselines – Naive Classifiers

Our first baseline model, B1, uses a pretrained BERT model and two linear layers. Our second baseline model, B2, has an embedding layer, a bidirectional LSTM encoder, an attention mechanism, and one linear layer. The attention mechanism used is based on the Bahdanau. The baseline model's architecture and attention mechanism class implemented code provided in [18].

3.5. Adversarial Debiasing – Implementation 1

We came across two styles of architecting adversarial debiasing systems, the first of which we explored we will refer to as A1. The general architecture is based on Zhang et al [31] but we referenced the implementation of [5] and [12].

The classifier uses a pretrained BERT model to encode the raw text and uses two linear layers to do the toxicity classification task. The adversary consists of two linear layers. The first linear layer in both the classifier and adversary use PyTorch's xavier_normal_ function to initialize the parameters. The classifier and adversary are each pretrained for three epochs. They are then alternately trained for 30 iterations where each iteration includes one epoch to train the adversary and one mini-batch to train the classifier. The classifier and adversary are set up such that the classifier receives the text input and produces a predicted toxicity classification, then the adversary tries to predict the protected attribute label from the predicted toxicity classification, similar to how it's done in [31].

3.6. Adversarial Debiasing – Implementation 2

The second adversarial debiasing architecture we explored was based on [29] and we will refer to it as A2. The model is composed of three components: Firstly, an encoder H which transforms the text into a high-dimensional space; Secondly, a binary classifier C that forecasts the target attribute based on the input text; And finally, an adversary D that predicts the protected attribute from the input text. The encoder we used is a single-layer bidirectional LSTM with attention mechanism, while both classifiers are two-layer MLPs featuring a tanh activation function.

In our dataset, we represented each data point as a triplet, denoted by $(x_i, y_i, z_i); i \in 1...N$, where $\mathbf{x}_i$ denotes the text we aimed to classify, $\mathbf{y}_i$ represents the corresponding label, and $\mathbf{z}_i$ represents the protected attribute such as race, gender, or religion in our experimental setting. The classifier C is trained using the tuples (x_i, y_i), while the adversary D is trained using the tuples (x_i, z_i).

We adopted the training process from [18], which suggests a two-step procedure for adversarial debiasing. The first step involves the pretraining of the encoder H and classifier C with standard supervised objective.

$$\min_{C,H} \sum_{i=1}^{N} \mathbf{L}(\mathbf{C}(\mathbf{H}(\mathbf{x}_i), \mathbf{y}_i) \tag{1}$$

The pretraining was carried out with the expectation and assumption that the encoder would encode the necessary information to predict both the target and protected attributes.

In the second step, we took the best performing encoder in the task of predicting the target attribute and start alternate training the adversary and the encoder shown in equations 2 and 3.

$$\min_{D} \frac{1}{N} \sum_{i=1}^{N} \mathbf{L}(\mathbf{D}(\mathbf{H}(\mathbf{x}_i), \mathbf{z}_i) \tag{2}$$

$$\min_{H,C} \frac{1}{N} \sum_{i=1}^{N} \alpha \mathbf{L}(\mathbf{C}(\mathbf{H}(\mathbf{x}_i), \mathbf{y}_i) + (1-\alpha)\mathbf{L}(\mathbf{D}(\mathbf{H}(\mathbf{x}_i), 0.5) \tag{3}$$

The hyperparameter α determines to what extent which loss term should be prioritized. Equation 3 is expected to trick the adversary by generating random representations by forcing it to spit out random guesses. Since our protected output is binary, we are using 0.5 at equation 3.

4. Results

To evaluate the two adversarial debiasing approaches against the baseline methods, we created a held-out test set from the Jigsaw dataset so that each model's results are comparable. Each of the models are trained on the same training set and do not see the test set until evaluation time. We think this experimental setup best mimics the way real-world classification models are trained on known data and then deployed on unseen data. Each model is evaluated against typical classification evaluation metrics as well as our chosen fairness metrics. Our chosen evaluation metrics include accuracy, precision, recall, and F1 score, for all of which, a higher score means better performance. Our fairness metrics include demographic parity, equality of opportunity, and equality of odds, and for each of these, a lower score means better performance.

4.1. Results for the Gender Dimension

We first explored the gender bias dimension with ”female” as the protected attribute and ”male” as the nonprotected attribute. From the results in Table 1, we observe that A1 achieved the highest accuracy and precision across all four metrics when trained and tested on the dataset with gender as the bias dimension. However, the baseline B1 outperformed both adversarial models and the other baseline in terms of recall and F1 score.

From Table 2, we see that all four models achieved a demographic parity value of 0.0225. This corresponds to the slight difference in toxicity rates for the protected and nonprotected groups. An ideal outcome would mean both toxicity rates are equal and demographic parity equals 0. Equality of opportunity can be thought of as true positive parity, and again a score of 0 means parity has been achieved. From Table 2, we see that A2 and the baseline had at least double the equality of opportunity score that A1 achieved. A1 clearly performed better in this category. Equality of odds is achieved when both true positive parity and false positive parity are achieved. From the results in Table 2, we observe that while A2 achieved a much lower false positive parity rate than A1 and the baselines, but A1 still achieved a better equality of odds score across all three models.

Table 1. Evaluation metric results

Metric	B1	B2	A1	A2
Accuracy	0.8981	0.8659	0.8994	0.8708
Precision	0.8139	0.7402	0.8341	0.7516
Recall	0.7663	0.7163	0.7387	0.7143
F1 Score	0.7869	0.7271	0.7742	0.7304

Table 2. Fairness metric results with P denoting ”protected” (female) and NP denoting ”nonprotected” (male) demographic attributes

Metric	B1	B2	A1	A2
Toxicity rate (P)	0.1429	0.1429	0.1429	0.1429
Toxicity rate (NP)	0.1654	0.1654	0.1654	0.1654
Demographic parity	0.0225	0.0225	0.0225	0.0225
True positive rate (P)	0.5487	0.4824	0.5005	0.4754
True positive rate (NP)	0.6097	0.5231	0.5160	0.5054
Equality of opportunity	0.0610	0.0407	0.0155	0.0300
False positive rate (P)	0.0417	0.0642	0.0320	0.0605
False positive rate (NP)	0.0475	0.0745	0.0275	0.0607
False positive parity	0.0058	0.0103	0.0045	0.0002
Equality of odds	0.0667	0.0511	0.0200	0.0301

4.2. Results for the Religion Dimension

We investigated the religion dimension with ”Muslim” as the protected attribute and ”Christian” as the nonprotected attribute. From the results in Table 3, it appears that baseline B1 achieved the highest performance on all standard classification evaluation metrics except for recall, which was achieved by A1. However, Table 4 shows that once again, A1 achieved the best performance on the fairness metrics of equality of opportunity and equality of odds.

Table 3. Evaluation metric results for religion

Metric	B1	B2	A1	A2
Accuracy	0.8954	0.8706	0.8815	0.8686
Precision	0.7815	0.7145	0.7435	0.7096
Recall	0.7277	0.6662	0.7458	0.6701
F1 Score	0.7502	0.6854	0.7446	0.6865

Table 4. Fairness metric results with P denoting ”protected” (Muslim) and NP denoting ”nonprotected” (Christian) attributes

Metric	B1	B2	A1	A2
Toxicity rate (P)	0.2454	0.2454	0.2454	0.2454
Toxicity rate (NP)	0.0937	0.0937	0.0937	0.0937
Demographic parity	0.1517	0.1517	0.1517	0.1517
True positive rate (P)	0.5562	0.4526	0.5125	0.4667
True positive rate (NP)	0.4461	0.3273	0.6052	0.3374
Equality of opportunity	0.1100	0.1253	0.0927	0.1294
False positive rate (P)	0.0947	0.1206	0.0816	0.1288
False positive rate (NP)	0.0286	0.0358	0.0656	0.0389
False positive parity	0.0661	0.0848	0.0160	0.0898
Equality of odds	0.1761	0.2101	0.1087	0.2192

4.3. Results for the Race Dimension

Finally, we investigated the bias dimension of race with ”Black” as the protected attribute and ”White” as the nonprotected attribute. Once again, baseline B1 achieved the best performance on the typical classification evaluation metrics but A1 achieved the best performance on the fairness metrics.

Overall, our experimental results align with our expectation that adversarial debiasing training would impact typical evaluation metrics negatively, but would improve performance with regard to the fairness metrics.

Table 5. Evaluation metric results for race

Metric	B1	B2	A1	A2
Accuracy	0.7897	0.7054	0.7784	0.7042
Precision	0.7396	0.6258	0.7244	0.6262
Recall	0.7219	0.6155	0.7133	0.6185
F1 Score	0.7295	0.6195	0.7183	0.6217

Table 6. Fairness metric results with P denoting ”protected” (Black) and NP denoting ”nonprotected” (White) attributes

Metric	B1	B2	A1	A2
Toxicity rate (P)	0.3148	0.3148	0.3148	0.3148
Toxicity rate (NP)	0.2644	0.2644	0.2644	0.2644
Demographic parity	0.0504	0.0504	0.0504	0.0504
True positive rate (P)	0.5657	0.4326	0.5680	0.4137
True positive rate (NP)	0.5692	0.4002	0.5642	0.4296
Equality of opportunity	0.0036	0.0323	0.0038	0.0159
False positive rate (P)	0.1131	0.1729	0.1294	0.1724
False positive rate (NP)	0.1287	0.1834	0.1427	0.1931
False positive parity	0.0156	0.0105	0.0133	0.0207
Equality of odds	0.0192	0.0428	0.0172	0.0366

5. Conclusion

In this work, we attempted to design a generalized framework for mitigating bias from downstream NLP tasks. To achieve this, we explored two approaches of adversarial debiasing, with one method involving the tuning of a hyperparameter to balance the loss between the classifier and adversary. We assessed the outcomes generated by each approach utilizing a variety of fairness metrics. Based on our limited experimentation with the two adversarial methods, it appears that the technique which does not demand hyperparameter tuning outperforms the one that does. However, we acknowledge that the outcomes we observed is not conclusive due to the inconsistency of the fairness metrics applied during the experiment. Moreover, we believe that we may not have dedicated sufficient efforts to fine-tune the hyperparameters for our A2 implementation.

In future, we aim to explore more fairness metrics that maintain consistency and potentially commit efforts to design a universal fairness metrics for debiasing across tasks and bias dimensions. Furthermore, we aspire to conduct experiment on intersectional debiasing, which we could not explore due to time constraints.

References

[1] R. Richard Banks, Jennifer L. Eberhardt, and Lee Ross. Discrimination and implicit bias in a racially unequal society. *California Law Review*, 94(4):1169–1190, 2006. 1

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Quantifying and reducing stereotypes in word embeddings, 2016. 1

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[4] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’15, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery. 1

[5] Shweta Chopra, Nupur Baghel, Shubham Annadate, and Rajalakshmi Dorairaj Shanmugasundaram, 2020. 3

[6] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019. 1, 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. 1

[8] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. 2018. 1, 2

[9] Aaron Fraenkel, 2020. 2

[10] Ismael Garrido-Muñoz , Arturo Montejo-Ráez , Fernando Martínez-Santiago , and L. Alfonso Ureña-López . A survey on bias in deep nlp. *Applied Sciences*, 11(7), 2021. 1

[11] Google, 2022. 2

[12] Henk Griffioen, 2018. 3

[13] Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 2

[14] Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. Mitigating gender bias amplification in distribution by posterior regularization. *CoRR*, abs/2005.06251, 2020. 1

[15] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans, 2019. 1

[16] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*, 2020. 1, 2

[17] Yoon Kim. Convolutional neural networks for sentence classification, 2014. 1

[18] Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. Topics to avoid: Demoting latent confounds in text classification, 2021. 1, 3

[19] John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States, July 2022. Association for Computational Linguistics. 1

[20] Peter A. Noseworthy, Zachi I. Attia, LaPrincess C. Brewer, Sharonne N. Hayes, Xiaoxi Yao, Suraj Kapa, Paul A. Fried-

man, and Francisco Lopez-Jimenez. Assessing and mitigating bias in medical artificial intelligence. *Circulation: Arrhythmia and Electrophysiology*, 13(3):e007988, 2020. 1

[21] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 2

[22] Otávio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Debiasing methods for fairer neural models in vision and language research: A survey. *arXiv preprint arXiv:2211.05617*, 2022. 2

[23] Archit Rathore, Sunipa Dev, Jeff M. Phillips, Vivek Srikumar, Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang, and Bei Wang. Verb: Visualizing and interpreting bias mitigation techniques for word representations, 2021. 1

[24] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*, 2021. 2

[25] Ameya Vaidya, Feng Mai, and Yue Ning. Empirical analysis of multi-task learning for reducing model bias in toxic comment detection. *arXiv preprint arXiv:1909.09758*, 2019. 2

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 1

[27] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Vicente Ordonez, and Caimng Xiong. Double-hard debiasing: Tailoring word embeddings for gender bias mitigation. 1

[28] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online, July 2020. Association for Computational Linguistics. 1

[29] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*, 2020. 1, 2, 3

[30] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019. 1

[31] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 2, 3