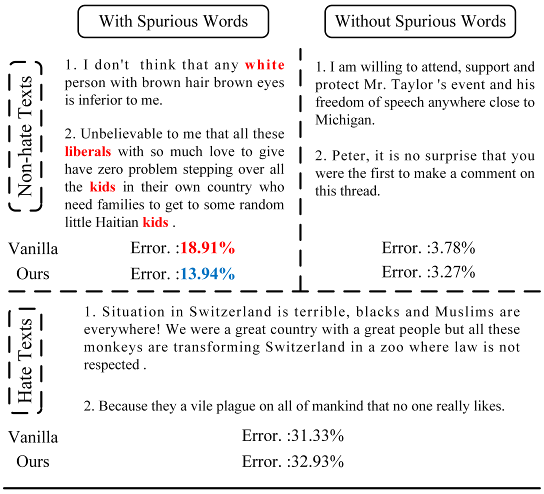
# Robust Hate Speech Detection Without Predefined Spurious Words

**Anonymous ACL submission**

## Abstract

Hate speech detection classifiers suffer from spurious correlations between specific words and the hate class. The spurious words can be either the identity words (e.g., "black", "female", "gay") or non-identity words (e.g., "sport", "football"). The current studies mainly focus on removing spurious correlations based on predefined identity words. In this paper, we develop a novel spurious correlation mitigating strategy, called ARLHAD, without any prior knowledge of spurious words. ARLHAD leverages a minimax game for optimization between a classifier and an adversary, in which the classifier aims to improve the hate speech detection performance by minimizing the classification loss while the adversary aims to maximize the loss mainly caused by spurious words. After training, ARLHAD improves the overall performance and more importantly, alleviates the spurious correlations. Experimental results on three hate speech detection datasets show the effectiveness of ARLHAD.


With Spurious Words
Without Spurious Words
Non-hate Texts
Hate Texts
Vanilla
Ours
Vanilla
Ours

1. I don't think that any **white** person with brown hair brown eyes is inferior to me.

2. Unbelievable to me that all these **liberals** with so much love to give have zero problem stepping over all the **kids** in their own country who need families to get to some random little Haitian **kids** .

Error. :**18.91%**
Error. :**13.94%**

1. I am willing to attend, support and protect Mr. Taylor 's event and his freedom of speech anywhere close to Michigan.

2. Peter, it is no surprise that you were the first to make a comment on this thread.

Error. :3.78%
Error. :3.27%

1. Situation in Switzerland is terrible, blacks and Muslims are everywhere! We were a great country with a great people but all these monkeys are transforming Switzerland in a zoo where law is not respected .

2. Because they a vile plague on all of mankind that no one really likes.

Error. :31.33%
Error. :32.93%

Figure 1: Classification results of the hate speech dataset STORMFRONT. The vanilla fine-tuned BERT model (Vanilla) learns spurious correlations between certain words and the hate class, leading to a high error rate for non-hate texts containing spurious words. Our model lowers the error rate of non-hate texts containing spurious words by mitigating spurious correlations and improving overall performance.

## 1 Introduction

Recent studies have demonstrated that hate speech detection classifiers suffer from spurious correlations between spurious words (e.g., "black", "white", "sport", "liberals") and hate class (Dixon et al., 2018; Kennedy et al., 2020; Wiegand et al., 2020; Ramponi et al., 2022). This is because spurious words frequently occur in hateful texts, which makes the classifier trained on such data have a high false positive rate for non-hate texts containing spurious words due to capturing the spurious correlations. A recent study shows that spurious words can be either the identity words (e.g., "black", "white") or the non-identity words (e.g., "kids", "liberals") (Ramponi et al., 2022). For example, as shown in Figure 1, "*I don't think that any white...*" and "*Unbelievable to me that all these liberals...*" are predicted as hateful by a vanilla fine-tuned BERT classifier. Hence, both identity and non-identity words severely undermine the generalization and robustness of hate speech detectors.

However, most existing approaches focus on mitigating spurious correlations between identity words and the hate class, such as adversarial training (Xia et al., 2020; Subramanian et al., 2021), instance re-weighting (Zhang et al., 2020; Subramanian et al., 2021), data re-balancing (Dixon et al., 2018; Park et al., 2018; Garg et al., 2019), and identity words masking or removing (Ramponi et al., 2022). Since these approaches highly rely on prior knowledge of spurious words, creating a pre-defined word list that includes many common identity words is a prerequisite. However, due to the massive number of non-identity words, creating a comprehensive pre-defined non-identity word list can be impractical. Without prior knowledge of spurious words, the current approaches cannot mitigate spurious correlations between non-identity words and the hateful class.

In this work, we aim to mitigate spurious correlations in hate speech detection without defining the spurious words. As shown in Figure 1, we can notice that the main error of a vanilla fine-tuned BERT model on non-hate class comes from the texts containing spurious words (18.91% error rate) while having a low error in texts without spurious words (3.85% error rate). Based on this observation, if the classifier has good accuracy on non-hateful texts with spurious words, the spurious correlation can be mitigated. To this end, inspired by adversarially reweighted learning (Lahoti et al., 2020), we propose an adversarially-reweighting-learning-based hate speech detection approach (ARLHAD) that can mitigate the spurious correlation without predefined spurious words. The idea is to enhance the losses due to the errors from non-hateful texts with spurious words so that the classifier can further optimize for these non-hateful texts. Concretely, we introduce a minimax game between a *classifier* and an *adversary* into the optimization, in which the classifier aims to mainly improve the classification accuracy and the adversary learns to identify and enhance the loss due to the misclassifications affected by spurious words in the non-hate class. The alternating training between the classifier and the adversary enables the model to mitigate spurious correlations while improving the overall performance. We show the effectiveness of our approach for spurious correlation mitigation on three hate speech detection datasets.

## 2 ARLHAD

### 2.1 Problem Statement

We consider the binary classification of hate speech detection. Let $\mathcal{X}$ denote texts and $\mathcal{Y}$ denote their labels (hate: $\mathcal{Y} = 1$ and non-hate: $\mathcal{Y} = 0$). Due to the imbalanced nature of hate speech datasets, the non-hate class is the majority class. We consider a set of words $\mathcal{S}$ spuriously correlated with the hate class. Given a set of training texts $\mathcal{D}_{train} = \{(x_i, y_i)\}$, $i = 1, \dots, n$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y} = \{0, 1\}$, but no observed spurious words $\mathcal{S}$, our goal is to learn a function (neural network parameterized by $\theta$) $f(\cdot) : \mathcal{X} \to \mathbb{R}^2$ to accurately classify testing samples in $\mathcal{D}_{test}$, especially, lower the false positive rate caused by spurious correlations between $\mathcal{S}$ and the hate class.

### 2.2 Approach

In order to mitigate spurious correlations and improve the overall performance, ARLHAD introduces a minimax game between two players: the classifier $f(x_i; \theta)$ and the adversary $\lambda(x_i; \phi)$ defined as an adversarial neural network: $g(x_i; \phi) \to \mathbb{R}$. The classifier and adversary play the adversarial games as below:

$$\min_{\theta} \max_{\phi} \sum_{i=1}^{n} \lambda(x_i; \phi) \cdot L(f(x_i; \theta), y_i),$$

where $L(f(x_i; \theta), y_i)$ indicates the objective function to train the classifier.

In particular, the classifier $f(x_i; \theta)$ aims to learn the optimal parameters $\theta$ by minimizing the expected loss $L$. The adversary $\lambda(x_i; \phi)$ learns to assign high weights $\lambda$ for misclassified texts of the classifier $f(x_i; \theta)$ to maximize the loss $L$. Because the classifier makes significant errors on non-hate texts with spurious words, the adversary would assign high weights to these texts. The classifier then adjusts itself to further minimize the loss, leading to a low error on non-hate texts with spurious words.

**Classifier:** The classifier $f(x_i; \theta)$ aims to learn the optimal parameters $\theta$ for hate speech detection. Due to the lack of sufficient hateful texts for training, hate speech classifiers usually perform poorly in the hate class. Hence, the classifier $f(x_i; \theta)$ is trained to minimize the label-distribution-aware margin (LDAM) loss (Cao et al., 2019), which is a state-of-the-art approach for imbalanced learning to enhance the performance of the minority class, defined as follows:

$$L(f(x_i; \theta), y_i) = -\log \frac{e^{f_{y_i}(x_i;\theta) - \Delta_{y_i}}}{e^{f_{y_i}(x_i;\theta) - \Delta_{y_i}} + e^{f_{1-y_i}(x_i;\theta)}}$$
$$\Delta_{y_i} = \frac{C}{n_{y_i}^{1/4}}, y_i \in \{0, 1\}$$

where $C$ is a hyperparameter, $n_{y_i}$ is the number of texts with the label $y_i$, and $f_{y_i}(x_i; \theta)$ is the logit output targeted label $y_i$. The key idea of LDAM is that the minority class is associated with a larger $\Delta_{y_i}$. Then, $\Delta_{y_i}$ is subtracted from the logit output $f_{y_i}(x_i; \theta)$, which increases the LDAM loss and thus encourages a larger margin for the minority class.

**Adversary:** Since LDAM primarily focuses on improving the performance of hateful texts, the model's errors are mainly attributed to the non-hateful texts containing spurious words that are easily misclassified. The adversary $g(\cdot) : \mathcal{X} \to [0, 1]$

parameterized by $\phi$ is to identify such texts. In particular, in order to maximize the loss of the objective function, the adversary $g(x_i; \phi)$ learns to assign large weights for the misclassified texts. In this way, when the learner $f(x_i; \theta)$ is trained to minimize the loss of the objective function, the learner will improve the performance of misclassified texts that mainly consist of non-hateful texts with spurious words.

In practice, to prevent exploding gradients during optimization, we perform normalization of assigned weights. Besides, we don't want the $\lambda(x_i; \phi)$ to be small and add 1 to make all the training examples contribute to the training loss:

$$\lambda(x_i; \phi) = 1 + n \cdot \frac{g(x_i; \phi)}{\sum_{i=1}^{n} g(x_i; \phi)}$$

The alternating training between the learner $f(x_i; \theta)$ and the adversary $g(x_i; \phi)$ enables the model to refine its ability to enhance its overall performance and address misclassifications of non-hateful texts affected by spurious correlations. In the experiments, we fine-tune the uncased BERT-base model (Devlin et al., 2019) for the learner $f(x; \theta)$ and the adversary $g(x; \phi)$. The output layer of the learner is a linear function with two logits, while the output layer of the adversary is a sigmoid activation mapping $g(x; \phi)$ into $[0, 1]$.

## 3 Experiments

### 3.1 Datasets

We conduct experiments on three hate speech detection datasets, including **GAB** (Kennedy et al., 2020), **STORMFRONT** (De Gibert et al., 2018), and **REDDIT** (Röttger et al., 2021). All three datasets provide the spurious words labeled by Ramponi et al., 2022. [1]

We split each dataset into train, validation, and test same as Ramponi et al., 2022. Dataset statistics are summarized in Table 1. Besides, we ensure that the ratio of hateful texts and non-hateful texts in training, validation, and testing sets is consistent for each dataset.

### 3.2 Baselines

Little work exists on mitigating spurious correlations without access to spurious words for hate speech detection. At the same time, most of the hate speech datasets are class imbalanced with a majority proportion of texts belonging to the negative or non-hate class, so we compare our method (ARLHAD) with the combinations of imbalanced learning methods and two spurious correlation mitigation techniques that need prior information about spurious words.

Table 1: Statistics of the datasets. Train, Val, and Test denote the number of training data, validation data, and testing data. Class Ratio denotes the ratio of hateful texts and non-hateful texts. Avg Len denotes the average words per post of each dataset.

| Dataset | Train | Val | Test | Class Ratio | Avg Len |
|---|---|---|---|---|---|
| GAB | 19881 | 2466 | 2591 | 1:14 | 24 |
| STORMFRONT | 8360 | 997 | 1044 | 1:8 | 17 |
| REDDIT | 17262 | 2157 | 2157 | 1:12 | 31 |

We adopt Re-Weight (RW) and LDAM for imbalanced learning. Specifically, we minimize the cross-entropy loss and re-weight each text with the balanced class weight for RW while we minimize the LDAM loss for the LDAM approach.

Mask and Remove are effective techniques for mitigating spurious correlations when the prior information of spurious words is available (Ramponi et al., 2022). For Mask, we replace all the spurious words with a unique token [MASK] only in the training set, which encourages the model to blend all the spurious words into the same contextualized representation. As for Remove, we just remove all the spurious words from the training set. By masking and removing the spurious words, we can prevent the model from learning the spurious correlations between spurious words and hate class.

### 3.3 Evaluation Metrics

We evaluate models' performance using *macro-$F1$* (F1) score and use False Negative Rate (FNR) to show models' performance in the minority class. We further evaluate the effectiveness of spurious correlations mitigation using the False Positive Rate (FPR) following Ramponi et al., 2022. However, we think FPR is not sufficient to evaluate whether the model is able to mitigate spurious correlations. Obviously, if the model only improves the accuracy of non-hateful texts without spurious words, it also helps to lower the FPR. Consequently, we propose False Positive Equality Gap (FPEG), as defined below:

$$\text{FPEG} = |\text{FPR}_{/\mathcal{S}} - \text{FPR}_{\mathcal{S}}|$$

where $\text{FPR}_{/\mathcal{S}}$ denotes the false positive rate on the non-hateful texts without spurious words and $\text{FPR}_{\mathcal{S}}$

[1] The list of spurious words is available at https://github.com/dhfbk/hate-speech-artifacts

Table 2: Experiment results of all methods on three benchmark datasets. For each dataset, the best-performing result of each metric is highlighted in boldface. Scores are averages of 5 runs with different seeds and subscriptions indicate standard deviation. ↑ denotes higher scores are better, whereas ↓ denotes lower scores are better.

| | | GAB | | | STORMFRONT | | | REDDIT | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F1$_\uparrow$ | FPR$_\downarrow$ | FNR$_\downarrow$ | F1$_\uparrow$ | FPR$_\downarrow$ | FNR$_\downarrow$ | F1$_\uparrow$ | FPR$_\downarrow$ | FNR$_\downarrow$ |
| RW | Vanilla | $0.708_{0.01}$ | $0.090_{0.01}$ | $0.284_{0.05}$ | $0.752_{0.01}$ | $0.105_{0.03}$ | $0.262_{0.06}$ | $0.663_{0.05}$ | $0.180_{0.07}$ | $\mathbf{0.158_{0.07}}$ |
| | Mask | $\mathbf{0.714_{0.01}}$ | $\mathbf{0.034_{0.01}}$ | $0.545_{0.03}$ | $0.729_{0.02}$ | $\mathbf{0.055_{0.02}}$ | $0.497_{0.10}$ | $0.643_{0.04}$ | $0.193_{0.05}$ | $0.194_{0.02}$ |
| | Remove | $0.682_{0.02}$ | $0.121_{0.03}$ | $0.234_{0.05}$ | $0.740_{0.01}$ | $0.092_{0.02}$ | $0.344_{0.09}$ | $0.611_{0.05}$ | $0.239_{0.07}$ | $0.173_{0.06}$ |
| LDAM | Vanilla | $0.682_{0.01}$ | $0.131_{0.02}$ | $\mathbf{0.179_{0.04}}$ | $0.737_{0.02}$ | $0.136_{0.02}$ | $\mathbf{0.203_{0.04}}$ | $0.704_{0.04}$ | $0.130_{0.05}$ | $0.205_{0.08}$ |
| | Mask | $0.700_{0.02}$ | $0.083_{0.02}$ | $0.353_{0.05}$ | $0.728_{0.03}$ | $0.113_{0.05}$ | $0.318_{0.10}$ | $0.682_{0.03}$ | $0.140_{0.04}$ | $0.239_{0.06}$ |
| | Remove | $0.655_{0.01}$ | $0.153_{0.01}$ | $0.192_{0.02}$ | $0.734_{0.02}$ | $0.119_{0.02}$ | $0.272_{0.06}$ | $0.695_{0.02}$ | $0.119_{0.04}$ | $0.280_{0.08}$ |
| ARLHAD | | $\underline{0.707_{0.01}}$ | $\underline{0.077_{0.02}}$ | $\underline{0.355_{0.07}}$ | $\underline{\mathbf{0.762_{0.02}}}$ | $0.100_{0.04}$ | $\underline{0.271_{0.07}}$ | $\underline{\mathbf{0.765_{0.01}}}$ | $\underline{\mathbf{0.042_{0.02}}}$ | $0.408_{0.02}$ |
| | | FPR/s | FPRs | FPEG$_\downarrow$ | FPR/s | FPRs | FPEG$_\downarrow$ | FPR/s | FPRs | FPEG$_\downarrow$ |
| RW | Vanilla | $0.018_{0.01}$ | $0.167_{0.02}$ | $0.149_{0.02}$ | $0.044_{0.01}$ | $0.197_{0.03}$ | $0.153_{0.02}$ | $0.096_{0.04}$ | $0.263_{0.09}$ | $0.167_{0.06}$ |
| | Mask | $0.009_{0.01}$ | $0.063_{0.01}$ | $\mathbf{0.054_{0.01}}$ | $0.022_{0.01}$ | $0.102_{0.04}$ | $\mathbf{0.080_{0.03}}$ | $0.110_{0.03}$ | $0.272_{0.06}$ | $0.162_{0.04}$ |
| | Remove | $0.045_{0.01}$ | $0.204_{0.04}$ | $0.159_{0.04}$ | $0.039_{0.01}$ | $0.170_{0.04}$ | $0.131_{0.03}$ | $0.151_{0.05}$ | $0.325_{0.09}$ | $0.174_{0.04}$ |
| LDAM | Vanilla | $0.037_{0.01}$ | $0.235_{0.03}$ | $0.198_{0.02}$ | $0.059_{0.02}$ | $0.249_{0.04}$ | $0.190_{0.03}$ | $0.058_{0.03}$ | $0.201_{0.07}$ | $0.143_{0.05}$ |
| | Mask | $0.029_{0.01}$ | $0.142_{0.04}$ | $0.113_{0.03}$ | $0.058_{0.04}$ | $0.196_{0.07}$ | $0.138_{0.03}$ | $0.074_{0.03}$ | $0.205_{0.05}$ | $0.131_{0.04}$ |
| | Remove | $0.059_{0.01}$ | $0.256_{0.02}$ | $0.197_{0.02}$ | $0.056_{0.01}$ | $0.214_{0.05}$ | $0.158_{0.04}$ | $0.063_{0.03}$ | $0.174_{0.06}$ | $0.111_{0.05}$ |
| ARLHAD | | $0.020_{0.01}$ | $0.139_{0.03}$ | $\underline{0.119_{0.03}}$ | $0.048_{0.02}$ | $0.167_{0.05}$ | $\underline{0.119_{0.03}}$ | $0.015_{0.01}$ | $0.066_{0.01}$ | $\underline{\mathbf{0.051_{0.01}}}$ |

are the false positive rate on non-hateful texts with spurious words. Intuitively, if the model is able to mitigate spurious correlations between $\mathcal{S}$ and hate class, the spurious words $\mathcal{S}$ should have less impact on the classification which leads to similar performance between non-hateful texts with spurious words and without spurious words. Therefore, a lower FPEG means the model is stronger in mitigating spurious correlations.

### 3.4 Implementation Details

We fine-tune the uncased BERT-base model (Devlin et al., 2019) for all experiments. We train each baseline for 10 epochs with the Adam optimizer (Kingma et al., 2014), a mini-batch size of 64, and a learning rate of $2e^{-5}$ on a machine equipped with two NVIDIA GeForce RTX 3090. The code is available online.[2]

### 3.5 Results and Discussion

We report the mean and standard deviation over 5 runs with different seeds in Table 2. We can observe that ARLHAD achieves the best F1 over all the baselines on STORMFRONT and REDDIT datasets. Compared with the vanilla methods of RW and LDAM, we lower the False Positive Rate and achieve better FPEG over three different hate speech datasets, which validates the effectiveness of our method in mitigating spurious correlations and improving overall performance.

With the prior information about the spurious words, the Remove or Mask methods show lower FPEG than the Vanilla methods, which means applying these two methods can remove spurious correlations between spurious words and the hate class. However, the Remove and Mask cannot achieve better overall performance in most cases. Without prior knowledge of the spurious words, ARLHAD not only enhances the overall performance but also achieves FPR and FPEG scores comparable to those of the Remove and Mask methods. Importantly, our method can achieve the best FPEG scores on the REDDIT dataset.

Noting that mitigating the spurious correlations can potentially compromise the model's performance in the hate class, as we can see the FNR increases when we apply mitigating strategies. Intuitively, since most hate speech includes spurious words, using spurious words as a feature to make classifications can indeed help to improve the performance of the hate class. In contrast, mitigating the correlations between spurious words and the hate class can hurt the model's utility.

## 4 Conclusions

We have designed a novel method called ARLHAD to mitigate spurious correlations without prior information about spurious words. ARLHAD forces the classifier to optimize the misclassified non-hateful texts due to spurious correlation via an adversarial game. Experiments demonstrated the effectiveness of our method in mitigating spurious correlations compared with two effective approaches (Mask and Remove) that rely on prior information about spurious words. In future work, we aim to explore mitigating spurious correlations without compromising the performance of the hate class.

[2] https://tinyurl.com/ARLHAD-code-2024

## Limitations

Our work represents a step forward toward mitigating spurious correlation in the absence of spurious words. However, we acknowledge a limitation. Across the experimental results, we can observe a common thing among spurious correlation mitigation methods: while mitigating spurious correlations can enhance the performance of the non-hate class, it leads to a decline in the performance of the hate class. Ideally, we desire a hate speech detector that is both immune to spurious correlations and achieves high accuracy in detecting hate speech. Although our method does not achieve both simultaneously, we believe it can serve as an inspiration for future studies in this area.

## References

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465-476, Online. Association for Computational Linguistics.

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.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. 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. Association for Computational Linguistics.

Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Fairness-aware class imbalanced learning In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2045–2051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: mitigating discrimination in text classifications with instance weighting In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. 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. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019*, pages 219–226, Honolulu, HI, USA. ACM.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1565–1576.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, pages 728–740.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11-20, Brussels, Belgium. Association for Computational Linguistics.

Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027-3040. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida M. Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth PortilloWightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Gabriel Cardenas, Alyzeh Hussain, Austin Lara, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2020. The Gab hate corpus: A collection of 27k posts annotated for hate speech.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P.Kingma and Jummy Ba. 2014. Adam: A Method for Stochastic Optimization.