Mitigating Dialect Bias in Toxicity Detection

Noman Vadsariya Alan Liang Katie Hancock

Abstract

Automated toxicity detection systems face a key challenge in maintaining fairness across different dialects. A major source of bias in these models comes from the underlying training data, where dialects such as African American English (AAE) are often associated with toxic labels. In this work, we conduct an experimental study of different dialect fairness methods proposed in prior research. We evaluate both surface-level and model-level approaches to understand their effectiveness. Our results show that model-level methods outperform surface-level methods. In particular, using XGBoost with vector-scaling loss and adversarial debiasing, we are able to reduce the False Positive Rate gap between AAE and Standard American English (SAE) to nearly negligible levels, while only slightly reducing the overall F1 score and keeping False Negative Rates under control.

1. Introduction

Automated toxicity detection systems are widely deployed across online platforms to identify and moderate harmful content. In this work, we study the problem of building a toxicity classifier that is both accurate and fair across dialects, specifically between African American English (AAE) and Standard American English (SAE). Formally, we consider labels $y \in \{\text{non-toxic, toxic}\}$ and groups $g \in \{\text{AAE, SAE}\}$, with the goal of minimizing classification error while reducing the disparities in error rates between groups.

The motivation for this work comes from evidence that toxicity detection systems exhibit systematic bias against certain linguistic communities. Prior studies show that models often assign higher toxicity scores to AAE text, even when it is non-toxic (Sap et al., 2019). This bias largely stems from dataset and annotation imbalances, where AAE is more frequently associated with offensive labels. As a result, models learn spurious correlations between dialect and toxicity rather than true semantic signals. This leads to higher false positive rates for AAE and raises concerns about fairness and reliability in deployed moderation systems (Blodgett et al., 2020).

Several approaches have been proposed to mitigate such biases. Adversarial debiasing methods aim to remove sensitive information by training a predictor alongside an adversary that attempts to infer protected attributes, encouraging the model to learn representations that are invariant to group membership (Zhang et al., 2018). More recent work has explored multi-task learning approaches that explicitly model dialect as an auxiliary task. For example, (Spliethöver et al., 2024) propose a framework that jointly learns dialect identification and biased language detection, allowing the model to better capture linguistic variation while reducing bias. Their results show that incorporating dialect modeling can improve fairness and lead to more reliable predictions across dialect groups. In addition, recent work has investigated modifications to the loss function to better handle group-sensitive or imbalanced data distributions, such as approaches that adjust class margins during training (Kini et al., 2021).

In this project, we study and mitigate dialect bias through controlled experiments. Our main goal is to reduce the False Positive Rate for AAE group while maintaining overall model performance. We first analyze the impact of dataset imbalance on bias, and then evaluate both surface-level and model-level methods. Through these experiments, we examine the trade-offs between fairness, accuracy, and stability in toxicity detection systems.

2. Methods

2.1. Dataset Preprocessing

We begin by describing the primary dataset and preprocessing steps used to support our analysis of dialect bias in toxicity detection. Our study is based on the widely used dataset introduced by (Davidson et al., 2017), which contains English tweets annotated into three categories: hate speech, offensive language, and none. Following prior work in toxicity detection, we consolidate the hate speech and offensive language labels into a single toxic class, while none category is treated as non-toxic. This binary formulation aligns with the objective of measuring disparities in harmful content classification while simplifying the evaluation of fairness metrics such as false positive rates.

A key limitation of the Davidson dataset is the absence of

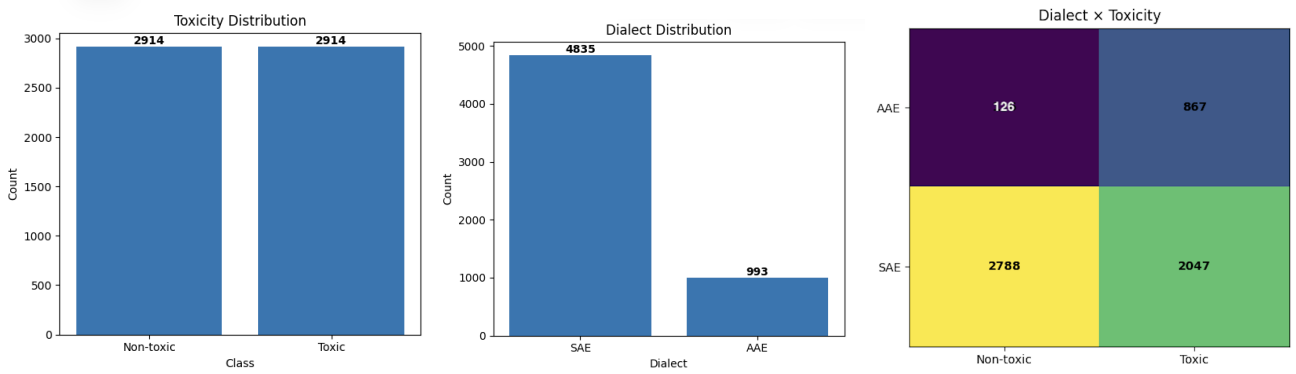


Figure 1. Distribution of class labels and dialect groups in the balanced dataset (strict threshold 0.8). The left plot shows balanced toxic and non-toxic classes, the middle plot shows overall dialect distribution with SAE more prevalent than AAE, and the right heatmap shows the joint distribution, highlighting underrepresentation of AAE in the non-toxic class.

dialect annotations. To address this, we augment the dataset using the TwitterAAE model (Blodgett et al., 2016), which estimates the probability of a tweet belonging to AAE or SAE. Since predictions are probabilistic, we consider two thresholds: a relaxed threshold of 0.6 and a stricter threshold of 0.8. The relaxed setting increases coverage but introduces noise, while the stricter threshold provides fewer but more reliable labels. We use the stricter threshold (0.8) in our experiments to ensure high-confidence dialect assignments.

Following preprocessing, we observe two distinct forms of imbalance in the dataset. First, there is a label imbalance, with substantially more toxic instances than non-toxic ones. Second, there is a dialect imbalance, wherein AAE-associated tweets appear more frequently within the toxic class. These two imbalances are not independent; rather, they are correlated in a way that complicates analysis. Training directly on this data risks fusing the effects of class imbalance with dialect bias.

To isolate dialect bias, we perform label balancing as a controlled preprocessing step. Specifically, we construct a balanced dataset by applying random undersampling to the majority class (toxic) so that it matches the number of non-toxic instances.

2.1.1. Distribution in Balanced Dataset

After balancing the class labels, dialect imbalance remains pronounced, particularly within the non-toxic class (Figure 1). AAE is strongly underrepresented in non-toxic examples compared to SAE. This skew can influence model behavior: with limited exposure to non-toxic AAE, the model may learn a spurious association between AAE and toxicity. Overall, this highlights that bias in toxicity detection is closely tied to the underlying data distribution. By controlling for label imbalance and examining dialect distribution separately, we create a more reliable setting to study and mitigate fairness issues in later stages.

2.2. Baseline Methods

We implemented a set of baseline models to evaluate how different modeling approaches behave with respect to dialect bias. These baselines span both traditional machine learning and deep learning methods, allowing us to compare how bias manifests across model types.

2.2.1. XGBoost with BERT Embeddings

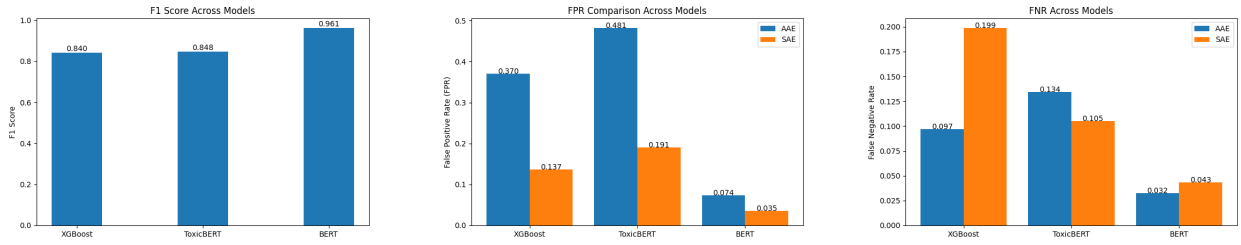
We first employ an XGBoost model, a gradient-boosted decision tree framework known for its strong performance on structured data tasks (Chen & Guestrin, 2016). We use contextualized text representations derived from BERT (Devlin et al., 2019) as input features. One key motivation for including XGBoost is its compatibility with custom objective functions, which makes it particularly suitable for later experimentation with fairness-aware loss formulations. In addition, tree-based models have been widely used in prior work on toxicity detection and bias analysis, providing a useful point of comparison with neural models.

2.2.2. ToxicBert

We use ToxicBERT, a pretrained transformer-based model that has been fine-tuned specifically for toxicity detection tasks. Because it has already been optimized for toxicity classification, it serves as a strong, task-specific baseline. Using ToxicBERT without further fine-tuning allows us to evaluate how a widely deployed toxicity model behaves on our dataset, particularly with respect to dialect bias.

2.2.3. Fine-tuned BERT

We implement a standard BERT model (Devlin et al., 2019) and fine-tune it directly on our processed dataset. Fine-



(a) F1 Score (b) False Positive Rate (c) False Negative Rate

Figure 2. Comparison of baseline models across performance and fairness metrics. (a) Fine-tuned BERT achieves the highest F1 score, while XGBoost and ToxicBERT perform similarly. (b) XGBoost and ToxicBERT show significantly higher FPR for AAE compared to SAE. (c) FNR chart shows an opposite trend, with higher values for SAE in simpler models and more balanced behavior in BERT.

tuning allows the model to adapt its contextual representations to the task at hand while learning decision boundaries tailored to our data distribution. BERT's deep contextual encoding enables it to capture semantic and syntactic nuances in text, which can reduce reliance on superficial correlations such as dialectal cues.

2.3. Experiments

2.3.1. Reweighting with XGBoost

The first approach we explore is reweighting, where we assign a higher weight α to non-toxic AAE samples and a weight of 1 to all others. This weight scales the loss during training, increasing the penalty for misclassifying non-toxic AAE examples. The goal is to reduce false positives for AAE. This approach is simple to implement and does not require changes to the model, but indirectly influences fairness by modifying the training signal rather than explicitly optimizing a fairness objective.

2.3.2. Adversarial Debiasing

We train an XGBoost predictor f on BERT embeddings jointly with an adversary a that recovers the dialect group g from margin $z = f(x)$ and label y (Zhang et al., 2018); conditioning on y targets Equalized Odds (Hardt et al., 2016) rather than demographic parity:

$$a(z, y) = \sigma\big(\beta^\top [z,\ y,\ zy]\big)\,, \quad \mathcal{L}_{\text{adv}} = \text{BCE}(a,\ g).$$

At each boosting round we refit a and apply the gradient-reversal projection update of (Zhang et al., 2018),

$$\tilde{g}_z = g_z - \text{proj}_{g_z^{\text{adv}}} g_z - \lambda\, g_z^{\text{adv}},$$

with g_z, g_z^{adv} the supervised and adversarial gradients w.r.t. z; the Hessian is unchanged. After training we tune per-group thresholds (Hardt et al., 2016) as

$$\arg\min |\text{FPR}_{\text{AAE}} - \text{FPR}_{\text{SAE}}| \text{ s.t. } \text{F1} \geq \rho \cdot \text{F1}^\star,$$

and grid-search (λ, c, ρ), where c is the adversary's inverse-ℓ_2 regularization.

2.3.3. Vector-Scaling Loss with XGBoost

We apply post-hoc vector scaling to XGBoost toxicity scores to reduce false positives for AAE. Starting from the predicted probability p, we convert to logit space,

$$z = \log \frac{p}{1-p},$$

and apply a group-specific affine transformation

$$s = \alpha_g z + \beta_g,$$

where α_g rescales the margin and β_g shifts the decision boundary. The adjusted score is mapped back through a sigmoid, and the final prediction uses threshold 0.5.

Because our main fairness concern is over-flagging non-toxic AAE as toxic, we calibrate only AAE scores. We fix $\alpha_{\text{SAE}} = 1$ and $\beta_{\text{SAE}} = 0$, tune only α_{AAE} and β_{AAE} on the validation set, and select the best trade-off using accuracy, F1, and group-specific FPR and FNR.

3. Results

In addition to overall F1 score, we focus on two group-based error metrics: False Positive Rate (FPR) and False Negative Rate (FNR) across dialect groups. FPR is important as it captures how often non-toxic content is incorrectly flagged as toxic, while FNR ensures that improvements in FPR do not come at the cost of missing toxic content.

3.1. Baseline Methods

Across all baseline models, we observe a clear difference in error rates across dialect groups (Figure 2). XGBoost and ToxicBERT achieve similar overall performance (F1 $\approx$ 0.84), but show large gaps in False Positive Rate (FPR). For XGBoost, FPR for AAE is 0.370 compared to 0.137 for

SAE, while ToxicBERT shows an even larger gap (0.481 vs 0.191). This means both models often misclassify non-toxic AAE tweets as toxic, with ToxicBERT showing the strongest bias. In contrast, the fine-tuned BERT model achieves higher performance (F1 = 0.961) and reduces this gap, with FPR of 0.074 for AAE and 0.035 for SAE. Although a gap still exists, it is much smaller, suggesting that better language representations help reduce bias.

Looking at False Negative Rate (FNR), we see the opposite pattern. For XGBoost, FNR is higher for SAE (0.199) than for AAE (0.097), meaning the model misses more toxic SAE content while predicting "toxic" more often for AAE. ToxicBERT shows a smaller difference (0.134 for AAE vs 0.105 for SAE), while BERT has low and more balanced FNR (0.032 for AAE vs 0.043 for SAE). This occurs because the model is less likely to predict toxicity for SAE, leading to more missed toxic SAE samples. In contrast, AAE has a lower FNR, as the model is more likely to predict toxicity for AAE due to the learned bias.

Based on these observations, our goal is to reduce the False Positive Rate for AAE while maintaining overall performance and controlling False Negative Rates, focusing on improving fairness in XGBoost while keeping F1 close to the baseline (≈ 0.84).

3.2. Experiments

3.2.1. Reweighting

As we vary the reweighting parameter α, we observe a consistent decrease in the FPR for AAE (Figure 3). This leads to a reduction in the FPR gap between AAE and SAE, indicating improved fairness. The FPR for SAE remains relatively stable across different values of α. At the same time, the FNR for AAE increases as α grows, reflecting a more conservative model that predicts "toxic" less often. The FNR for SAE shows only minor changes. Overall, this demonstrates a clear trade-off: reducing false positives for AAE comes at the cost of increased false negatives and a slight drop in overall performance. The best trade-off is observed at α = 6.75, where the FPR for AAE is reduced to approximately 0.23, with SAE at 0.16, resulting in a small gap of $\approx$ 0.06. This comes with a moderate increase in FNR for AAE (0.14) and a slight decrease in F1 score to around 0.81. These results show that reweighting effectively improves fairness, while maintaining reasonably strong overall performance.

3.2.2. Adversarial Debiasing

The EO-pressure parameter λ has a sharply non-monotonic effect: a narrow basin at $\lambda = 0.1$ nearly closes the FPR gap without harming F1, while $\lambda \in [0.15, 0.25]$ destabilizes training as the gradient-reversal term overwhelms the supervised signal (Figure 4). Unlike reweighting (which reshapes the loss surface) and vector scaling (which rescales scores post-hoc), the adversary reshapes the score distribution itself so that small per-group threshold shifts (Hardt et al., 2016) suffice for fairness. The selected configuration cuts $\mathrm{FPR}_{\mathrm{AAE}}$ by roughly 40% and shrinks the FPR gap by an order of magnitude relative to the XGBoost baseline at a cost of ≈ 0.011 F1 (Table 1).

3.2.3. Vector-Scaling Loss

Vector scaling substantially improved fairness for the balanced XGBoost baseline. The best setting was $\alpha_{\mathrm{AAE}} = 1.0$ and $\beta_{\mathrm{AAE}} = -1.0$, showing that shifting the AAE decision boundary mattered more than changing the slope. Under this setting, $\mathrm{FPR}_{\mathrm{AAE}}$ decreased from 0.37 to 0.15, while the FPR gap dropped from 0.233 to 0.011.

This came with the expected trade-off: $\mathrm{FNR}_{\mathrm{AAE}}$ increased from 0.097 to 0.161, while overall performance remained relatively stable, with F1 0.83 versus a baseline F1 of 0.84.

Figure 5 shows the effect of varying β_{AAE} while fixing $\alpha_{\mathrm{AAE}} = 1.0$. As β_{AAE} becomes more negative, $\mathrm{FPR}_{\mathrm{AAE}}$ decreases but $\mathrm{FNR}_{\mathrm{AAE}}$ increases, illustrating a clear fairness–performance trade-off. The circled point marks the selected setting, which gave the best overall balance.

Vector scaling gave one of the strongest fairness–performance trade-offs on the balanced dataset, suggesting that much of the dialect bias appears at the final scoring stage.

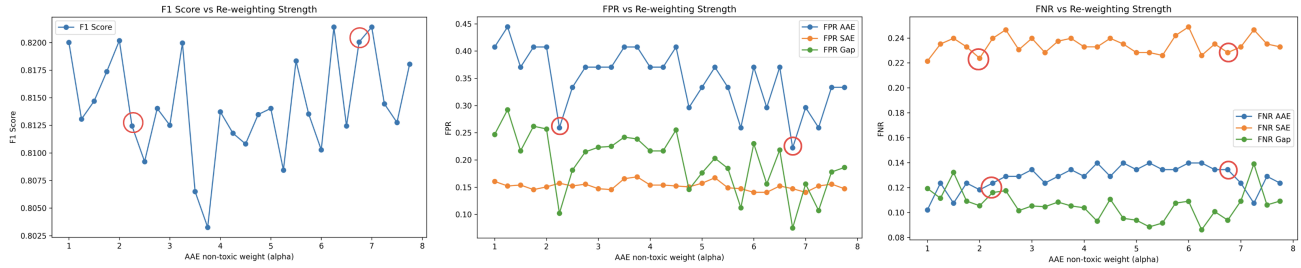
3.3. Adversarial Debiasing using BERT

We replace the linear adversary with an MLP that consumes the BERT embedding $e \in \mathbb{R}^{768}$ alongside $[z, y, zy]$:

$$a_\theta([e; z; y; zy]) : \mathbb{R}^{771} \to \mathbb{R}^{256} \to \mathbb{R}^{128} \to (0, 1),$$

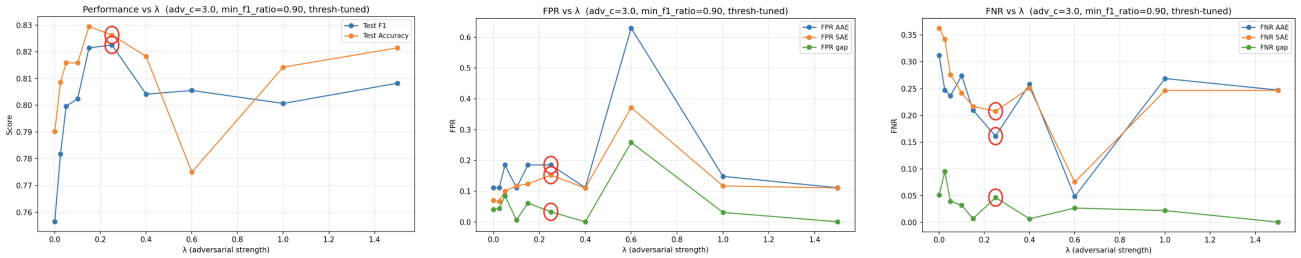
warm-started across boosting rounds; the gradient-reversal coupling (Zhang et al., 2018) and threshold tuning (Hardt et al., 2016) are unchanged. A BERT-conditioned adversary should detect dialect cues invisible to a margin-only adversary, in principle closing the FPR gap at lower λ.

Two findings stand out (Table 1). The optimal λ shrinks by a factor of 60, confirming the minimax prediction that more expressive adversaries deliver stronger fairness pressure per unit of λ. But the BERT adversary does not yield a uniformly better operating point: it reaches a higher F1 and a comparable FPR gap at a substantially higher mean FPR. Equalized Odds is a parity criterion, so flagging both groups equally aggressively satisfies it just as well as flagging both conservatively; given more capacity, the adversary closes the gap by raising SAE FPR to match AAE rather than lowering AAE. The practical implication is that adversary capacity is not a free lever under EO, when the goal is to reduce unfair



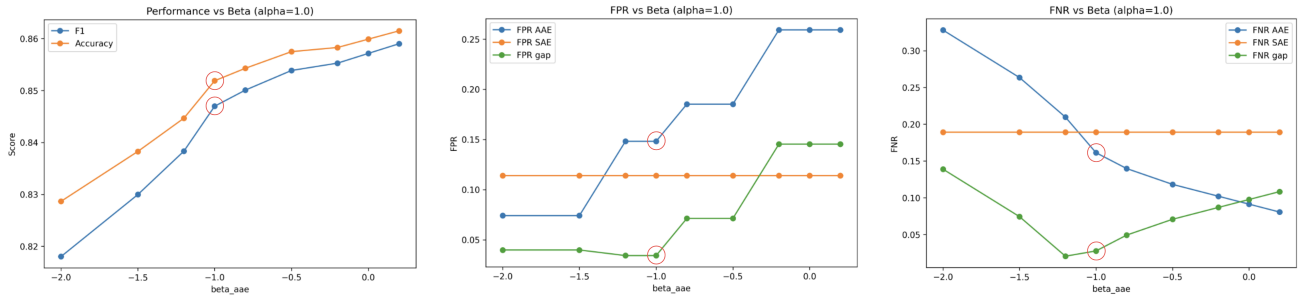
(a) F1 Score (b) False Positive Rate (c) False Negative Rate

Figure 3. Effect of reweighting strength α on performance and fairness



(a) F1 Score (b) False Positive Rate (c) False Negative Rate

Figure 4. Effect of adversary strength c on performance and fairness in adversarial debiasing



(a) F1 Score (b) False Positive Rate (c) False Negative Rate

Figure 5. Effect of vector-scaling parameter β on performance and fairness in XGBoost.

Experiments	F1	FPR_{AAE}	FPR_{SAE}	Gap	FNR_{AAE}	FNR_{SAE}	Gap
Re-weighting	0.81	0.23	0.16	0.06	0.14	0.23	0.09
Vector-Scaling Loss	0.83	0.15	0.14	0.01	0.16	0.18	0.02
Adv. Debiasing	0.82	0.19	0.16	0.03	0.16	0.21	0.05
Adv. Debiasing using BERT	0.90	0.29	0.31	0.02	0.17	0.22	0.05

Table 1. Comparative results of experiments on TwitterAAE

false positives on AAE rather than merely equalize them, a weaker margin-only adversary is preferable.

3.4. Generalization Testing with HateXplain

To evaluate generalization, we tested two of our best-performing methods on the HateXplain dataset (Mathew et al., 2022) to assess whether the fairness improvements transfer to an unseen data distribution.

3.4.1. Adversarial Debiasing

We apply the EO-trained models with their TwitterAAE-tuned thresholds directly to the HateXplain dataset (Table 2). While the in-domain FPR gaps are small, they increase significantly under distribution shift, and overall performance drops. Using default thresholds performs even worse, showing that group-specific calibration only partially transfers across datasets. The performance drop is expected due to differences between Twitter data and HateXplain's distribution. However, the fairness degradation is more important: FPR for AAE remains much higher than SAE, indicating that bias reappears under shift. This suggests that the bias reduction learned during training is not robust and depends on the data distribution.

Experiments	F1	FPR_{AAE}	FPR_{SAE}	Gap	FNR_{AAE}	FNR_{SAE}	Gap
Vector-Scaling Loss	0.64	0.90	0.72	0.18	0.07	0.19	0.13
Adv. Debiasing (balanced)	0.52	0.80	0.37	0.43	0.12	0.51	0.39
Adv. Debiasing (unbalanced)	0.53	0.72	0.41	0.31	0.20	0.49	0.29

Table 2. Comparative results of experiments on HateXplain dataset

Overall, these results show that fairness improvements do not generalize well across datasets. Threshold tuning helps, but is not sufficient. More robust approaches, such as domain-adaptive training or invariant fairness objectives, are needed for better generalization.

3.4.2. VECTOR-SCALING LOSS

We applied the calibrated VS-XGBoost model directly to HateXplain. As a result, the performance dropped substantially relative to the in-domain TwitterAAE experiments, with F1 $= 0.638$, indicating poor generalization under distribution shift. Fairness also deteriorated sharply. On HateXplain, the model produced very high false positive rates for both groups, with $\text{FPR}_{\text{AAE}} = 0.90$ and $\text{FPR}_{\text{SAE}} = 0.72$, yielding an FPR gap of 0.18. The false negative rates showed a similar group asymmetry, with $\text{FNR}_{\text{AAE}} = 0.07$ and $\text{FNR}_{\text{SAE}} = 0.19$, giving an FNR gap of 0.13 (Table 2).

These results suggest that vector scaling is effective in-domain but highly dataset-specific, and that post-hoc fairness corrections may need to be re-tuned on the target domain under distribution shift.

4. Process

4.1. Experiments with Unbalanced Dataset

On the unbalanced dataset, bias is much stronger, with the baseline heavily over-predicting toxicity, especially for AAE. With adversarial debiasing, reducing the FPR gap required much stronger regularization, and fairness could only be achieved after adjusting thresholds, indicating that bias remains embedded in the model. With vector-scaling loss, the FPR gap is reduced significantly (from 0.815 vs 0.530 to 0.482 vs 0.530), but this comes with an increase in FNR_AAE. Despite this trade-off, overall performance remains high (F1 ≈ 0.93). Overall, bias is harder to correct in the unbalanced setting, and improvements require stronger adjustments with noticeable trade-offs.

4.2. Group Thresholding

In addition to reweighting, we experimented with group thresholding, where different decision thresholds are used for AAE and SAE at inference time. Specifically, we set a higher threshold for AAE to require stronger confidence before predicting "toxic," aiming to reduce false positives for that group. While this approach reduced FPR_{AAE}, it also increased FNR_{AAE}, meaning the model missed more toxic AAE examples. It also led to a drop in overall performance (F1 ≈ 0.81). These results suggest that group thresholding mainly shifts errors between false positives and false negatives, rather than addressing the underlying bias.

4.3. Fairness-Aware Training

We then explored a model-level approach using fairness-aware training. In this method, we modified the loss function to include a penalty on the difference in FPR between groups. The loss is defined as:

$$\mathcal{L} = (1-\gamma)\,\mathcal{L}_{\text{BCE}} + \gamma\,(\text{FPR}_{AAE} - \text{FPR}_{SAE})^2$$

where $\mathcal{L}_{\text{BCE}}$ is the standard binary cross-entropy loss and γ controls the strength of the fairness penalty. The goal is to reduce both classification error and the FPR gap at the same time. However, we observed that training became unstable. This is because BCE is computed per sample, while the FPR gap is a group-level metric. Combining these creates a difficult optimization problem, leading to non-smooth and sensitive training behavior.

4.4. Experiments with Ternary Classification

We also tested a ternary setup with three labels: hate speech, offensive, and non-toxic. The overall trends are similar to the binary case. The FPR for AAE is lowest for the offensive class, likely due to its larger number of training samples, followed by the non-toxic class, and highest for the hate speech class. The gap between FPR_AAE and FPR_SAE remains similar across classes, confirming that data distribution plays a key role in shaping bias.

5. Contributions

We used the Davidson Hate Speech dataset (Davidson et al., 2017) as our primary dataset and augmented it with dialect labels using the pretrained TwitterAAE Model. The original implementation of this model was in Python 2, which we converted to Python 3 for use in our project. We also used the HateXplain dataset for generalization testing (Mathew et al., 2022). We experimented with multiple fairness-related methods proposed in prior research to perform a comparative analysis. For modeling, we used the XGBClassifier from the xgboost Python library, and pretrained BERT-based models using the Hugging Face Transformers library. The code for our project is available at github

6. AI Usage

We used AI tools to assist with converting legacy Python 2 code to Python 3 and for running quick experiments with different loss functions. Additionally, AI was primarily used for generating and refining plots throughout the project.

References

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of African-American English. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://aclanthology.org/D16-1120/.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*, 2017. URL https://api.semanticscholar.org/CorpusID:1733167.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819. doi: 10.48550/arXiv.1610.02413. URL https://doi.org/10.48550/arXiv.1610.02413.

Kini, G. R., Paraskevas, O., Oymak, S., and Thrampoulidis, C. Label-imbalanced and group-sensitive classification under overparameterization. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. Hatexplain: A benchmark dataset for explainable hate speech detection, 2022. URL https://arxiv.org/abs/2012.10289.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. The risk of racial bias in hate speech detection. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163/.

Spliethöver, M., Menon, S. N., and Wachsmuth, H. Disentangling dialect from social bias via multitask learning to improve fairness. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:270521896.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.