Mitigating Biases in Toxic Language Detection through Invariant Rationalization

Yung-Sung Chuang[1,2] **Mingye Gao**[2] **Hongyin Luo**[2]
James Glass[2] **Hung-yi Lee**[1] **Yun-Nung Chen**[1] **Shang-Wen Li**[3*]
[1]National Taiwan University, [2]MIT CSAIL, [3]Amazon AI
{yungsung, mingye, hyluo, glass}@mit.edu,
hungyilee@ntu.edu.tw, y.v.chen@ieee.org, shangwel@amazon.com

Abstract

Automatic detection of toxic language plays an essential role in protecting social media users, especially minority groups, from verbal abuse. However, biases toward some attributes, including gender, race, and dialect, exist in most training datasets for toxicity detection. The biases make the learned models unfair and can even exacerbate the marginalization of people. Considering that current debiasing methods for general natural language understanding tasks cannot effectively mitigate the biases in the toxicity detectors, we propose to use invariant rationalization (INVRAT), a game-theoretic framework consisting of a rationale generator and predictors, to rule out the spurious correlation of certain syntactic patterns (e.g., identity mentions, dialect) to toxicity labels. We empirically show that our method yields lower false positive rate in both lexical and dialectal attributes than previous debiasing methods.[1]

1 Introduction

As social media becomes more and more popular in recent years, many users, especially the minority groups, suffer from verbal abuse and assault. To protect these users from online harassment, it is necessary to develop a tool that can automatically detect the toxic language in social media. In fact, many toxic language detection (TLD) systems have been proposed in these years based on different models, such as support vector machines (SVM) (Gaydhani et al., 2018), bi-directional long short-term memory (BiLSTM) (Bojkovskỳ and Pikuliak, 2019), logistic regression (Davidson et al., 2017) and fine-tuning BERT (d'Sa et al., 2020).

However, the existing TLD systems exhibit some problematic and discriminatory behaviors (Zhou et al., 2021). Experiments show that the tweets containing certain surface markers, such as identity terms and expressions in African American English (AAE), are more likely to be classified as hate speech by the current TLD systems (Davidson et al., 2017; Xia et al., 2020), although some of them are not actually hateful. Such an issue is predominantly attributed to the biases in training datasets for the TLD models; when the models are trained on the biased datasets, these biases are inherited by the models and further exacerbated during the learning process (Zhou et al., 2021). The biases in TLD systems can make the opinions from the members of minority groups more likely to be removed by the online platform, which may significantly hinder their experience as well as exacerbate the discrimination against them in real life.

So far, many debiasing methods have been developed to mitigate biases in learned models, such as data re-balancing (Dixon et al., 2018), residual fitting (He et al., 2019; Clark et al., 2019), adversarial training (Xia et al., 2020) and data filtering approach (Bras et al., 2020; Zhou et al., 2021). While most of these works are successful on other natural language processing (NLP) tasks, their performance on debasing the TLD tasks are unsatisfactory (Zhou et al., 2021). A possible reason is that the toxicity of language is more subjective and nuanced than general NLP tasks that often have unequivocally correct labels (Zhou et al., 2021). As current debiasing techniques reduce the biased behaviors of models by correcting the training data or measuring the difficulty of modeling them, which prevents models from capturing spurious and non-linguistic correlation between input texts and labels, the nuance of toxicity annotation can make such techniques insufficient for the TLD task.

In this paper, we address the challenge by combining the TLD classifier with the selective rationalization method, which is widely used to inter-

* Work is not related to employment at Amazon.

[1]The source code is available at https://github.com/voidism/invrat_debias.

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 114–120

pret the predictions of complex neural networks. Specifically, we use the framework of Invariant Rationalization (INVRAT) (Chang et al., 2020) to rule out the syntactic and semantic patterns in input texts that are highly but spuriously correlated with the toxicity label, and mask such parts during inference. Experimental results show that INVRAT successfully reduce the lexical and dialectal biases in the TLD model with little compromise on overall performance. Our method avoids superficial correlation at the level of syntax and semantics, and makes the toxicity detector learn to use generalizable features for prediction, thus effectively reducing the impact of dataset biases and yielding a fair TLD model.

2 Previous works

Debiasing the TLD Task Researchers have proposed a range of debiasing methods for the TLD task. Some of them try to mitigate the biases by processing the training dataset. For example, Dixon et al. (2018) add additional non-toxic examples containing the identity terms highly correlated to toxicity to balance their distribution in the training dataset. Park et al. (2018) use the combination of debiased *word2vec* and gender swap data augmentation to reduce the gender bias in TLD task. Badjatiya et al. (2019) apply the strategy of replacing the bias sensitive words (BSW) in training data based on multiple knowledge generalization.

Some researchers pay more attention to modifying the models and learning less biased features. Xia et al. (2020) use adversarial training to reduce the tendency of the TLD system to misclassify the AAE texts as toxic speech. Mozafari et al. (2020) propose a novel re-weighting mechanism to alleviate the racial bias in English tweets. Vaidya et al. (2020) implement a multi-task learning framework with an attention layer to prevent the model from picking up the spurious correlation between the certain trigger-words and toxicity labels.

Debiasing Other NLP Task There are many methods proposed to mitigate the biases in NLP tasks other than TLD. Clark et al. (2019) train a robust classifier in an ensemble with a bias-only model to learn the more generalizable patterns in training dataset, which are difficult to be learned by the naive bias-only model. Bras et al. (2020) develop AFLITE, an iterative greedy algorithm that can adversarially filter the biases from the training dataset, as well as the framework to support it. Utama et al. (2020) introduce a novel approach of regularizing the confidence of models on the biased examples, which successfully makes the models perform well on both in-distribution and out-of-distribution data.

3 Invariant Rationalization

3.1 Basic Formulation for Rationalization

We propose TLD debiasing based on INVRAT in this paper. The goal of rationalization is to find a subset of inputs that 1) suffices to yield the same outcome 2) is human interpretable. Normally, we would prefer to find rationale in unsupervised ways because the lack of such annotations in the data. A typical formulation to find rationale is as following: Given the input-output pairs $(\boldsymbol{X}, Y)$ from a text classification dataset, we use a classifier f to predict the labels $f(\boldsymbol{X})$. To extract the rationale here, an intermediate rationale generator g is introduced to find a rationale $\boldsymbol{Z} = g(\boldsymbol{X})$, a masked version of X that can be used to predict the output Y, i.e. maximize mutual information between $\boldsymbol{Z}$ and Y.[2]

$$\max_{\boldsymbol{m} \in \mathcal{S}} I(Y; \boldsymbol{Z}) \quad \text{s.t. } \boldsymbol{Z} = \boldsymbol{m} \odot \boldsymbol{X} \tag{1}$$

Regularization loss $\mathcal{L}_{\text{reg}}$ is often applied to keep the rationale sparse and contiguous:

$$\mathcal{L}_{\text{reg}} = \lambda_1 \left| \frac{1}{N} \mathbb{E}\left[\|\boldsymbol{m}\|_1\right] - \alpha \right| + \lambda_2 \mathbb{E}\left[\sum_{n=2}^{N} |\boldsymbol{m}_n - \boldsymbol{m}_{n-1}| \right] \tag{2}$$

3.2 The INVRAT Framework

INVRAT (Chang et al., 2020) introduces the idea of *environment* to rationalization. We assume that the data are collected from different environments with different prior distributions. Among these environments, the predictive power of spurious correlated features will be variant, while the genuine causal explanations always have invariant predictive power to Y. Thus, the desired rationale should satisfy the following invariant constraint:

$$H(Y|\boldsymbol{Z}, E) = H(Y|\boldsymbol{Z}), \tag{3}$$

where E is the given environment and H is the cross-entropy between the prediction and the ground truth Y. We can use a three-player framework to find the solution for the above equation: an environment-agnostic predictor $f_i(\boldsymbol{Z})$, an environment-aware predictor $f_e(\boldsymbol{Z}, E)$, and a rationale generator $g(\boldsymbol{X})$. The learning objective of the two predictors are:

[2]Real examples of $\boldsymbol{X}$, $\boldsymbol{Z}$ can be found in Table 2.

$$\mathcal{L}_i^* = \min_{f_i(\cdot)} \mathbb{E}\left[\mathcal{L}\left(Y; f_i(\boldsymbol{Z})\right)\right] \quad (4)$$

$$\mathcal{L}_e^* = \min_{f_e(\cdot,\cdot)} \mathbb{E}\left[\mathcal{L}\left(Y; f_e(\boldsymbol{Z}, E)\right)\right] \quad (5)$$

In addition to minimizing the invariant prediction loss $\mathcal{L}_i^*$ and the regularization loss $\mathcal{L}_{\text{reg}}$, the other objective of the rationale generator is to minimize the gap between $\mathcal{L}_i^*$ and $\mathcal{L}_e^*$, that is:

$$\min_{g(\cdot)} \mathcal{L}_i^* + \mathcal{L}_{\text{reg}} + \lambda_{\text{diff}} \cdot \text{ReLU}\left(\mathcal{L}_i^* - \mathcal{L}_e^*\right), \quad (6)$$

where ReLU is applied to prevent the penalty when $\mathcal{L}_i^*$ has been lower than $\mathcal{L}_e^*$.

4 INVRAT for TLD Debiasing

4.1 TLD Dataset and its Biases

We apply INVRAT to debiasing TLD task. For clarity, we seed our following description with a specific TLD dataset where we conducted experiment on, hate speech in Twitter created by Founta et al. (2018) and modified by Zhou et al. (2021), and we will show how to generalize our approach. The dataset contains 32K toxic and 54K non-toxic tweets. Following works done by Zhou et al. (2021), we focus on two types of biases in the dataset: lexical biases and dialectal biases. Lexical biases contain the spurious correlation of toxic language with attributes including Non-offensive minority identity (NOI), Offensive minority identity (OI), and Offensive non-identity (ONI); dialectal biases are relating African-American English (AAE) attribute directly to toxicity. All these attributes are tagged at the document level. We provide more details for the four attributes (NOI, OI, ONI, and AAE) in Appendix A.

4.2 Use INVRAT for Debiasing

We directly use the lexical and dialectal attributes as the environments in INVRAT for debiasing TLD[3]. Under these different environments, the predictive power of spurious correlation between original input texts $\boldsymbol{X}$ and output labels $\boldsymbol{Y}$ will change. Thus, in INVRAT, the rationale generator will learn to exclude the biased phrases that are spurious correlated to toxicity labels from the rationale $\boldsymbol{Z}$. On the other hand, the predictive power for the genuine linguistic clues will be generalizable across environments, so the rationale generator attempts to keep them in the rationale $\boldsymbol{Z}$.

[3]To generalize our method for any other attributes or datasets, one can simply map environments to the attributes in consideration for debiasing.

Since there is no human labeling for the attributes in the original dataset, we infer the labels following Zhou et al. (2021). We match $\boldsymbol{X}$ with TOXTRIG, a handcrafted word bank collected for NOI, OI, and ONI; for dialectal biases, we use the topic model from Blodgett et al. (2016) to classify $\boldsymbol{X}$ into four dialects: AAE, white-aligned English (WAE), Hispanic, and other.

We build two debiasing variants with the obtained attribute labels, INVRAT (lexical) and INVRAT (dialect). The former is learned with the compound loss function in Equation (6) and four lexical-related environment subsets (NOI, OI, ONI, and none of the above); we train the latter using the same loss function but along with four dialectal environments (AAE, WAE, Hispanic, and other). In both variants, the learned $f_i(\boldsymbol{Z})$ is our environment-agnostic TLD predictor that classifies toxic languages based on generalizable clues. Also, in the INVRAT framework, the environment-aware predictor $f_e(\boldsymbol{Z}, E)$ needs to access the environment information. We use an additional embedding layer Emb_{env} to embed the environment id e into a n-dimensional vector $\text{Emb}_{\text{env}}(e)$, where n is the input dimension of the pretrained language model. Word embeddings and $\text{Emb}_{\text{env}}(e)$ are summed to construct the input representation for f_e.

5 Experiment

5.1 Experiment Settings

We leverage RoBERTa-base (Liu et al., 2019) as the backbone of our TLD models in experiments. F_1 scores and false positive rate (FPR) when specific attributes exist in texts are used to quantify TLD and debiasing performance, respectively. The positive label is "toxic" and the negative label is "non-toxic" for computing F_1 scores. When evaluating models debiased by INVRAT, we use the following strategy to balance F_1 and FPR, and have a stable performance measurement. We first select all checkpoints with F_1 scores no less than the best TLD performance in dev set by 3%. Then, we pick the checkpoint with the lowest dev set FPR among these selected ones to evaluate on the test set. We describe more training details and used hyperparameters in Appendix B.

5.2 Quantitative Debiasing Results

In the left four columns of Table 1, we show the F_1 scores and FPR in the entire dataset and in the NOI, OI, and ONI attributes for measuring lexical

		Test	nOI		OI		OnI		AAE	
		$F_1 \uparrow$	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$
	Vanilla	$92.3_{0.0}$	$89.8_{0.3}$	$10.2_{1.3}$	$98.8_{0.1}$	$85.7_{0.0}$	$97.3_{0.1}$	$64.7_{0.8}$	$92.3_{0.0}$	$16.8_{0.3}$
	LMixin-OnI	$85.6_{2.5}$	$87.0_{1.1}$	$14.0_{1.5}$	$98.9_{0.0}$	$85.7_{0.0}$	$87.9_{4.5}$	$\mathbf{43.7}_{3.1}$	-	-
	LMixin-ToxTrig	$86.9_{1.1}$	$85.5_{0.3}$	$11.2_{1.7}$	$97.6_{0.3}$	$\mathbf{71.4}_{0.0}$	$90.4_{1.8}$	$44.5_{1.5}$	-	-
	LMixin-AAE	-	-	-	-	-	-	-	$92.3_{0.1}$	$16.1_{0.4}$
33% train	Random	$92.2_{0.1}$	$89.5_{0.4}$	$9.3_{0.7}$	$\mathbf{98.9}_{0.0}$	$\mathbf{83.3}_{3.4}$	$97.4_{0.1}$	$67.2_{0.6}$	$92.2_{0.1}$	$16.7_{0.6}$
33% train	AFLite	$91.9_{0.1}$	$\mathbf{90.2}_{0.4}$	$11.3_{1.1}$	$98.9_{0.0}$	$85.7_{0.0}$	$97.3_{0.1}$	$68.0_{3.4}$	$91.9_{0.1}$	$16.8_{0.8}$
33% train	DataMaps-Ambig.	$92.5_{0.1}$	$89.2_{0.7}$	$7.4_{1.0}$	$98.9_{0.0}$	$85.7_{0.0}$	$\mathbf{97.5}_{0.0}$	$64.4_{1.4}$	$92.5_{0.1}$	$16.0_{0.4}$
33% train	DataMaps-Hard	$\mathbf{92.6}_{0.1}$	$89.5_{0.4}$	$6.3_{0.9}$	$98.8_{0.0}$	$85.7_{0.0}$	$97.4_{0.0}$	$62.0_{1.1}$	$\mathbf{92.6}_{0.1}$	$\mathbf{13.7}_{0.2}$
33% train	DataMaps-Easy	$91.9_{0.2}$	$86.8_{0.6}$	$\mathbf{5.9}_{0.7}$	$98.9_{0.0}$	$\mathbf{83.3}_{3.4}$	$97.2_{0.1}$	$\mathbf{60.3}_{3.8}$	$91.9_{0.2}$	$19.5_{2.8}$
Ours (RoBERTa-base)										
	Vanilla	$\mathbf{91.7}_{0.1}$	$\mathbf{90.1}_{0.3}$	$8.4_{0.4}$	$\mathbf{98.6}_{0.0}$	$81.0_{3.4}$	$97.0_{0.0}$	$63.4_{1.4}$	$\mathbf{95.9}_{0.2}$	$16.9_{1.0}$
	lexical removal	$90.9_{0.0}$	$86.0_{0.7}$	$18.3_{1.5}$	$98.1_{0.1}$	$78.6_{0.0}$	$96.4_{0.0}$	$61.7_{0.2}$	$95.1_{0.1}$	$18.7_{0.6}$
	InvRat (lexical)	$91.0_{0.5}$	$85.5_{1.6}$	$\mathbf{3.4}_{0.6}$	$97.5_{1.0}$	$76.2_{3.4}$	$\mathbf{97.2}_{0.2}$	$61.1_{1.5}$	$95.0_{0.5}$	$19.6_{1.0}$
	InvRat (dialect)	$91.0_{0.1}$	$85.9_{0.7}$	$\mathbf{3.4}_{0.5}$	$97.6_{0.5}$	$\mathbf{71.4}_{5.8}$	$97.1_{0.1}$	$\mathbf{57.9}_{2.2}$	$93.1_{1.0}$	$\mathbf{14.0}_{1.2}$

Table 1: Evaluation of all debiasing methods on the Founta et al. (2018) test set. We show the mean and s.d. (subscript) of F_1 and FPR across 3 runs. The top two sections contain the scores reported in Zhou et al. (2021). The bottom section contains scores of our methods. When FPR is lower, the model is less biased by lexical associations for toxicity. We used RoBERTa-base, while RoBERTa-large is used in Zhou et al. (2021). Thus, our Vanilla F_1 score is slightly lower than that of Zhou et al. (2021) by 0.5%.

bias. In addition to Vanilla, we include *lexical removal*, a naive baseline that simply removes all words existing in ToxTrig before training and testing.

For our InvRat (lexical/dialect) model, we can see a significant reduction in the FPR of nOI, OI, and OnI over Vanilla (RoBERTa without debiasing). Our approach also yields consistent and usually more considerable bias reduction in all three attributes, compared to the ensemble and data filtering debiasing baselines discussed in Zhou et al. (2021), where no approach improves in more than two attributes (e.g., LMixin-OnI reduces bias in OnI but not the rest two; DataMaps-Easy improves in nOI and OnI but has similar FPR to Vanilla in OI). The result suggests that InvRat can effectively remove the spurious correlation between mentioning words in three lexical attributes and toxicity. Moreover, our InvRat debiasing sacrifices little TLD performance[4], which can sometimes be a concern for debiasing (e.g., the overall performance of LMixin). It is worth noting that the lexical removal baseline does not get as much bias reduction as our method, even inducing more bias in nOI. We surmise that the weak result arises from the limitation of ToxTrig, since a word bank cannot enumerate all biased words, and there are always other terms that can carry the bias to the model.

We summarize the debiasing results for the dialectal attribute in the rightmost column of Table 1. Compared with the Vanilla model, our method effectively reduces the FPR of AAE, suggesting the consistent benefit of InvRat in debiasing dialect biases. Although the results from data relabeling (Zhou et al., 2021) and some data filtering approaches are better than InvRat, these approaches are complementary to InvRat, and combining them presumably improves debiasing performance.

5.3 Qualitative Study

We demonstrate how InvRat removes biases and keeps detectors focusing on genuine toxic clues by showing examples of generated rationales in Table 2. Part (a) of Table 2 shows two utterances where both the baseline and our InvRat debiasing predict the correct labels. We can see that when toxic terms appear in the sentence, the rationale generator will capture them. In part (b), we show three examples where the baseline model incorrectly predicts the sentences as toxic, presumably due to some biased but not toxic words (depend on the context) like *#sexlife, Shits, bullshit*. However, our rationale generator rules out these words and allows the TLD model to focus on main verbs in the sentences like *keeps, blame, have*. In part (c), we show some examples that our InvRat model

[4]There is some degradation in nOI, which may result from some performance fluctuation in the small dataset and the labeling issues mentioned in Zhou et al. (2021). We see the degradation as an opportunity for future dive deep rather than concerns.

		Gold	Vanilla	Ours
(a)	Oh my **god** there's a **f**king** STINKBUG and it's **in my ASS**	⚠	⚠	⚠
	@user yes I hear that it's **great** for a relationship to try and change your partner..	✋	✋	✋
(b)	Other than #kids, what **keeps** you from the #sexlife you want?	✋	⚠	✋
	Shits crazy but bet they'll **blame** us... wait for it	✋	⚠	✋
	@user @user You don't **have** to pay for their bullshit read your rights read the law I don't pay fo…	✋	⚠	✋
(c)	RT @user: my ex so ugly to me now like...i'll **beat** that hoe ass	⚠	⚠	✋
	@user **Stop** that, it's not your **fault** a scumbag decided to steal otems which were obviously meant for someone i…	⚠	⚠	✋
(d)	A shark washed up in the street after a cyclone in Australia	✋	✋	✋

Table 2: Examples from the test set with the predictions from vanilla and our models. ⚠ denotes toxic labels, and ✋ denotes non-toxic labels. The underlined words are selected as the rationale by our ratinoale generator.

fails to generate the true answer, while the baseline model can do it correctly. In these two examples, we observe that our rationale generator remove the offensive words, probably due to the small degree of toxicity, while the annotator marked them as toxic sentences. Part (d) of Table 2 shows another common case that when the sentence can be easily classified as non-toxic, the rationale generator tends not to output any words, and the TLD model will output non-toxic label. It is probably caused by the non-stable predictive power of these non-toxic words (they are *variant*), so the rationale generator choose to rule them out and keep rationale clean and invariant.

6 Conclusion

In this paper, we propose to use INVRAT to reduce the biases in the TLD models effectively. By separately using lexical and dialectal attributes as the environments in INVRAT framework, the rationale generator can learn to generate genuine linguistic clues and rule out spurious correlations. Experimental results show that our method can better mitigate both lexical and dialectal biases without sacrificing much overall accuracy. Furthermore, our method does not rely on complicated data filtering or relabeling process, so it can be applied to new datasets without much effort, showing the potential of being applied to practical scenarios.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59.

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.

Michal Bojkovskỳ and Matúš Pikuliak. 2019. Stufiit at semeval-2019 task 5: Multilingual hate speech detection on twitter with muse and elmo embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 464–468.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4060–4073.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11(1).

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.

Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr. 2020. Bert and fasttext embeddings for automatic

detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, pages 1–5. IEEE.

Marta Dynel. 2012. Swearing methodologically: the (im) politeness of expletives in anonymous commentaries on youtube. *Journal of English studies*, 10:25–50.

Antigoni 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 *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12(1).

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*.

Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. *EMNLP-IJCNLP 2019*, page 132.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

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.

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.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

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.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155.

A Bias attributes

We follow Zhou et al. (2021) to define four attributes (NOI, OI, ONI, and AAE) that are often falsy related to toxic language. NOI is mention of minoritized identities (e.g., *gay, female, Muslim*); OI mentions offensive words about minorities (e.g., *queer, n*gga*); ONI is mention of swear words (e.g., *f*ck, sh*t*). NOI should not be correlated with toxic language but is often found in hateful speech towards minorities (Dixon et al., 2018). Although OI and ONI can be toxic sometimes, they are used to simply convey closeness or emphasize the emotion in specific contexts (Dynel, 2012). AAE contains dialectal markers that are commonly used among African Americans. Even though AAE simply signals a cultural identity in the US (Green, 2002), AAE markers are often falsy related to toxicity and cause content by Black authors to mean suppressed more often than non-Black authors (Sap et al., 2019).

B Training Details

We use a single NVIDIA TESLA V100 (32G) for each experiment. The average runtime of experiments for *Vanilla* model in Table 1 are 2 hours. The INVRAT model in Table 1 need about 9 hours for a single experiment.

The main hyperparameters are listed in Table 3. More details can be found in our released code. We did not conduct hyperparameter search, but follow all settings in the official implementation of Zhou et al. (2021) [5]. One difference is that because INVRAT framework needs three RoBERTa models to run at the same time, we choose to use RoBERTa-base, while Zhou et al. (2021) uses RoBERTa-large. As a result, our F_1 score for the Vanilla model is about 0.5 less than the score in Zhou et al. (2021).

hyperparameter	value
optimizer	AdamW
adam epsilon	1.0×10^{-8}
learning rate	1.0×10^{-5}
training epochs	10
batch size	8
max gradient norm	1.0
weight decay	0.0
sparsity percentage (α)	0.2
sparsity lambda (λ_1)	1.0
continuity lambda (λ_2)	5.0
diff lambda (λ_{diff})	10.0

Table 3: The main hyperparameters in the experiment. Sparsity percentage is the value of α in $\mathcal{L}_{reg}$ mentioned in equation 2; sparsity lambda and continuity lambda are λ_1 and λ_2 in equation 2; diff lambda is λ_{diff} in equation 6.

[5] https://github.com/XuhuiZhou/Toxic_Debias