Challenges in Automated Debiasing for Toxic Language Detection

Xuhui Zhou[♡] **Maarten Sap**[♣] **Swabha Swayamdipta**[♢] **Noah A. Smith**[♣♢] **Yejin Choi**[♣♢]

[♡]Department of Linguistics, University of Washington
[♣]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♢]Allen Institute for Artificial Intelligence
xuhuizh@uw.edu, msap@cs.washington.edu
{swabhas,noah,yejinc}@allenai.org

Abstract

***Warning**: this paper contains content that may be offensive or upsetting.*

Biased associations have been a challenge in the development of classifiers for detecting toxic language, hindering both fairness and accuracy. As potential solutions, we investigate recently introduced debiasing methods for text classification datasets and models, as applied to toxic language detection. Our focus is on lexical (e.g., swear words, slurs, identity mentions) and dialectal markers (specifically African American English). Our comprehensive experiments establish that existing methods are limited in their ability to prevent biased behavior in current toxicity detectors. We then propose an automatic, dialect-aware data correction method, as a proof-of-concept study. Despite the use of synthetic labels, this method reduces dialectal associations with toxicity. Overall, our findings show that debiasing a model trained on biased toxic language data is not as effective as simply relabeling the data to remove existing biases.

1 Introduction

Current hate speech or toxic language detection[1] systems exhibit problematic and discriminatory behavior that causes them to have disparate negative impact on minority populations (Yasin, 2018; Guynn, 2020; Kim et al., 2020; Dias Oliva et al., 2020). Tweets simply containing a minority identity mention are commonly flagged as toxic by current systems, in contrast to those containing majority identity mentions, as illustrated in Figure 1.

At the core of the issue are *dataset biases*, i.e., spurious correlations between surface patterns and annotated toxicity labels (§2), which stem from the data creation process (Sap et al., 2019). Previous work has outlined two such biases for hate speech datasets (both shown in Figure 1): *lexical bias* which associates toxicity with the presence of certain words (e.g., profanities, identity mentions; Dixon et al., 2018; Dinan et al., 2019) and *dialectal bias*, where toxicity is correlated with surface markers of African American English (AAE; Davidson et al., 2019; Sap et al., 2019). When trained on biased datasets, models acquire *and* exacerbate these biases (e.g., flagging text by Black authors as more toxic than by white authors; Sap et al., 2019; Zhang et al., 2018).

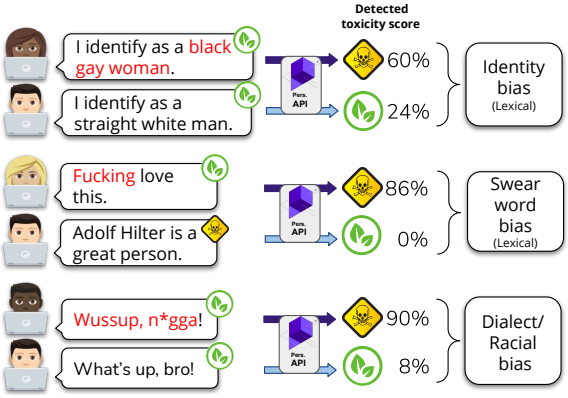

Figure 1: Lexical items and dialect markers cause problematic behavior for toxic language detection systems such as the widely used PerspectiveAPI. In the top two example pairs, statements with minority identity mentions and swear words used inoffensively are flagged as toxic, but majority identity mentions or offensive statements without overt swearing are missed. The bottom pair shows dialect-based racial bias for two inoffensive greetings, where markers of African American English (AAE) trigger the toxicity detector.

Concurrently, there has been elevated interest in developing *debiasing methods* for standard natural language understanding (NLU) tasks, i.e., methods that aim to decrease over-reliance on spurious correlations in NLU models (Clark et al., 2019; He et al., 2019; Karimi Mahabadi et al., 2020; Bras et al., 2020). This raises a natural question: *are*

[1]We use *hate speech* and *toxic language* interchangeably in this work, though their definitions do not perfectly align.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3143–3155
April 19 - 23, 2021.

current debiasing approaches effective for mitigating biases specific to toxic language detection?

In this work, we address the above question by investigating two classes of debiasing approaches to mitigate lexical and dialectal biases—one that employs additional training objectives for bias removal, and another that filters training instances likely exhibiting spurious biases (§3). Through comprehensive experiments, we show that both approaches face major challenges in mitigating biases from a model trained on a biased dataset (in our case, the dataset from Founta et al., 2018) for toxic language detection. While data filtering results in reduced bias associations in the data, *models* trained on filtered datasets still pick up on lexical (§4) and dialectal biases (§5). We find that dialectal biases are particularly challenging to address, as has also been shown by Xia et al. (2020). "Debiased" models still disproportionately flag text in certain dialects as toxic. Notably, mitigating dialectal bias through current debiasing methods does not mitigate a model's propensity to label tweets by Black authors as more toxic than by white authors.

We additionally explore an alternative proof-of-concept study—relabeling supposedly toxic training instances whose automatic translations into a majority dialect are deemed non-toxic by the classifier. To this end, we create a synthetic dataset via few-shot dialect translation system built with GPT-3 (Brown et al., 2020). While only an illustrative solution, it nevertheless takes into account the dialectal context of the tweet, resulting in a model less prone to dialectal and racial biases (§6). Overall, our findings indicate that debiasing a model already trained on biased toxic language data can be challenging, compared to relabeling the data to remove existing biases. Our code and data are publicly available on Github.[2]

2 Biases in Toxic Language Detection

We test the use of debiasing[3] methods for the task of toxic language detection, which aims to flag rude, offensive, hateful, or toxic language on the internet, with the goal of moderating online communities (Roberts, 2019; Vidgen et al., 2019). This task differs in several ways from the natural language understanding (NLU) tasks that debiasing methods have been successful on, such as textual entailment (e.g., SNLI, MNLI; Bowman et al., 2015; Williams et al., 2018) or reading comprehension (e.g., SQuAD; Rajpurkar et al., 2016). First, compared to these NLU tasks where there is one correct label, the toxicity of language is inherently more nuanced, subjective, and contextual, which causes toxic language datasets to have lower agreement in general (Ross et al., 2017). Second, the dataset biases in NLU are predominantly artifacts introduced during data creation (e.g., negations, exaggerations; Schwartz et al., 2017; Gururangan et al., 2018), whereas those in toxic language detection are grounded in the social dynamics of the world (Spears, 1998; Technau, 2018). For example, viewing AAE as a more toxic or less proper variety of English is a form of linguistic discrimination that upholds racial hierarchies in the United States (Rosa and Flores, 2017).

In this work, we consider two broad categories of toxic language dataset biases—lexical (§2.1) and dialectal (§2.2). Our experiments focus on a single, widely used dataset (§2.3) from Founta et al. (2018).

2.1 Lexical Biases (ToxTrig)

Current toxic language detection systems often rely on the presence or absence of certain words (e.g., swear words, identity mentions) to make their predictions (Dixon et al., 2018; Dinan et al., 2019). While most previous analyses of this bias relied on a simple list of "bad" words (Davidson et al., 2019; Dinan et al., 2019),[4] we take a more nuanced view of how lexical items can convey toxicity, inspired by work in pragmatics and sociolinguistics of rudeness (Dynel, 2015; Kasper, 1990, *inter alia*). Specifically, we manually split our full list of words into three distinct categories depending on the extent to which they carry profane or hateful meanings or are simply associated with hateful contexts.[5] We refer to the full set of words as ToxTrig, for Toxicity Triggers, which is included in our released repository.[6]

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

[3] Our definition of "bias" is specific to the social biases in toxic language detection datasets, grounded as lexical and dialectal biases; see Blodgett et al. (2020) for a detailed investigation of the term "bias".

[4] https://tinyurl.com/list-of-bad-words

[5] We note, however, that this categorization is in itself subjective.

[6] https://github.com/XuhuiZhou/Toxic_Debias/blob/master/data/word_based_bias_list.csv

Non-offensive minority identity mentions (NOI) refers to descriptive mentions of minoritized demographic or social identities (e.g., *gay*, *female*, *Muslim*). While these mentions are not usually inherently offensive by themselves, they are often found in offensive statements that are hateful towards minorities (Dixon et al., 2018). We detect these identity mentions in text using a list of 26 regular expressions.

Possibly offensive minority identity mentions (OI) are mentions of minoritized identities that could denote profanity or hate depending on pragmatic and contextual interpretations. This includes slurs and objectifying outdated terms to refer to minority groups, which are usually understood as attacks. Additionally, this includes *reclaimed* slurs (*queer*, *n*gga*), which connote less offensive intent when spoken by in-group members compared to out-group members (Croom, 2013).

Possibly offensive non-identity mentions (ONI) contains swear words and other profanities, which are usually offensive but not associated to any social groups (e.g., *f*ck*, *sh*t*). Note that the pragmatic interpretation of these words is not necessarily always toxic or offensive (Dynel, 2012), as they are often used to convey closeness between the speaker and listener or emphasize the emotionality of a statement (e.g., second example in in Figure 1).

2.2 Dialectal Biases (AAE)

Current toxic language detection systems also associate higher toxicity with dialectal markers of African American English (AAE; Sap et al., 2019; Davidson et al., 2019). Since AAE is a variety of English that is common among African Americans and often signals a cultural identity in the US (Green, 2002), this dialect-based racial bias causes speech by Black authors to be suppressed more often than non-Black authors (Sap et al., 2019), thereby exacerbating racial inequality (Rosa, 2019).

In our experiments, we estimate the dialect of a tweet using a topic model from Blodgett et al. (2016). This model was trained on 60M tweets, where the dialect of the tweet was inferred from the model coordinates, which yielded a probability of a tweet being in one of four dialects (African-American English, white-aligned English, Hispanic, and other). In this study, we only focus on African-American English (AAE) and white-aligned English (WAE) tweets; both definitions are based on US English, as per Blodgett et al. (2016).[7] Our experiments either use the probability of a tweet being in these dialects, or assign tweets their estimated-most-probable dialect.

2.3 Dataset for Toxic Language Detection

We focus our analyses on a widely used hate speech dataset of English tweets (Founta et al., 2018). The tweets were collected using a multi-round bootstrapping procedure, and were labeled out of context[8] for toxic language. We focus on the 86k tweets that are annotated as hateful, abusive, or neither and discard those labelled as spam. We aggregate the abusive and hateful labels into a single *toxic* category, yielding 32k toxic and 54k non-toxic tweets.[9]

3 Debiasing Methods

We consider two types of debiasing methods from current literature. The first type addresses known, pre-defined biases—such as lexical and dialectal biases for hate speech detection, via a model-based approach involving additional training objectives (§3.1). In contrast, the second type is agnostic to prior knowledge about biases, and instead filters out examples that appear "too easy" and might hence contain spurious correlations (§3.2).

3.1 Debiased Training for Pre-Defined Toxicity Biases

We use the LEARNED-MIXIN method of Clark et al. (2019), which achieved high out-of-distribution (OOD) performance on several NLU tasks, for debiased training. This method trains an ensemble containing a *bias-only* model which only uses pre-defined features corresponding to known biases, and a *full* model which uses all features. Intuitively, the ensemble encourages the full

[7] We avoid using disputed terms such as *general American English*, *standard American English*, or *mainstream US English*, which are frequently used for WAE, since we believe that no dialect should be privileged with the designation "general", "standard", or "mainstream" (Rosa, 2019).

[8] Only the tweet text—no profile information or conversational context—was shown to annotators.

[9] We also explored using another widely used hate speech dataset (Davidson et al., 2017), which collected tweets using a seed list of swear words and slurs. However, in line with findings by Xia et al. (2020), debiasing led to degenerate behavior due to the data collection process, as discussed in Appendix B.

model to rely more on features unrelated to the biases. Once trained, the bias-only model is discarded, and only the "bias-free" full model is used for inference, following Clark et al. (2019).

Bias-only model Given its effectiveness on bag-of-words (BoW) features, we use an SVM classifier as the lexical-bias-only model. For example, the TOXTRIG-only model counts the frequency of TOXTRIG words in each tweet. Our dialectal-bias-only model uses the probability of dialects (AAE, WAE, Hispanic, and other) obtained from a dialect detector (Blodgett et al., 2016) as features in a SVM classifier.

Full model We fine-tune a RoBERTa-large classifier (Liu et al., 2019), a state-of-the-art classifier for the toxicity detection task. See Appendix A.1 for more modeling details.

Note that we only consider the LEARNED-MIXIN-ONI and LEARNED-MIXIN-TOXTRIG models for lexical debiasing, due to poor accuracies of the bias-only models for NOI and OI.[10]

3.2 Data Filtering for Spurious Biases

In addition to debiasing methods that handle known biases, we also explore automated approaches which filter out instances exhibiting unspecified, spurious biases. Specifically, we describe below two data selection methods that have shown strong OOD performance.

AFLite (Bras et al., 2020) is an algorithm based on the key intuition that examples predicted correctly by the simplest methods likely exhibit spurious biases. An ensemble of simple linear classifiers is trained and tested on different partitions of the data; test instances which are "predictable", or classified correctly by most classifiers in the ensemble are discarded. The algorithm is iterative, and is repeated until a target data size is achieved. Models trained on this filtered dataset achieve higher performance on OOD and adversarially constructed test sets, compared to the original model, on several text and image classification datasets. This indicates a reduction in spurious biases in the filtered data.

[10]The NOI and OI bias-only models reach 63% and 67% accuracy, respectively, which is empirically hard for the ensemble to use. This is likely due to low coverage in the train set of those categories (4.43% NOI and 4.25% OI).

DataMaps (Swayamdipta et al., 2020) show the presence of distinct regions in a dataset—namely, easy, hard and ambiguous—defined with respect to a given model. These regions are discovered based on the training dynamics of a model, determined by the model's confidence in the true class, for each example, as well as the variability of this confidence, throughout training epochs. Swayamdipta et al. (2020) show that training exclusively on the hard and ambiguous regions of the data results in high OOD performance, indicating lower prevalance of spurious biases. The easy region is the largest in size for RoBERTa; however, experiments showed that training exclusively on these examples hurt OOD generalization on different NLU tasks. Following this work, we create DataMaps-Easy, DataMaps-Ambiguous, and DataMaps-Hard subsets for our dataset.

Following Swayamdipta et al. (2020), we set the target filtered subset size to 33% of the original training set for both filtering methods, but our filtering additionally preserved the original label proportions. We then fine-tune a RoBERTa-large classifer on these filtered subsets; see Appendix A.2 for more details.

4 Experiments: Lexical Biases

We investigate the effect of debiasing approaches (§3) on removing lexical biases in hate speech detection. First, we discuss the evaluation framework for measuring bias reduction (§4.1). We present quantitative (§4.2) and qualitative (§4.3) results on lexical bias removal for all debiasing approaches, and OOD evaluation for debiased training methods (§4.4). See Appendix A.3 for hyperparameters and other experimental settings.

4.1 Evaluation Framework

We report the performance of all models as overall accuracy and F_1 with respect to the toxic class. Given that current hate speech systems tend to rely heavily on the presence of NOI, OI, and ONI mentions (§2.1) for labeling text as toxic, we use false positive rate (FPR) over each of these categories to measure the degree of bias in the model, following Hardt et al. (2016) and Xia et al. (2020). Specifically, we report the FPR of a model on tweets containing NOI (FPR_{NOI}), OI (FPR_{OI}), and ONI (FPR_{ONI}), as well the F_1 corresponding to each of these classes. Intuitively, the lower the FPR_*, the

		$R_{\text{NOI}} \downarrow$	$R_{\text{OI}} \downarrow$	$R_{\text{ONI}} \downarrow$
	Original	0.0445	0.2641	0.6718
33% train	Random	0.0345	0.2603	0.6683
	AFLite	0.0434	0.2458	0.6016
	DataMaps-Ambig.	0.0126	0.1968	**0.5839**
	DataMaps-Hard	**0.0081**	**0.1853**	0.5849
	DataMaps-Easy	0.0772	0.3661	0.7720

Table 1: Lexical associations between toxicity and TOXTRIG mentions in the original dataset (Founta et al., 2018) and various filtered counterparts. Random, AFLite, and DataMaps all contain only 33% of the original data after filtering. Lower Pearson R correlation value indicates less superficial patterns in the dataset, i.e., less bias. ***Takeaway:*** The hard and ambiguous subsets given by DataMaps contain the lowest amount of lexical associations, indicated in boldface.

less the model infers lexical associations for toxicity, and hence is less biased.

Evaluation for Filtered Datasets We additionally consider metrics based on spurious lexical associations for data filtering approaches. This measures prevalence of spurious surface patterns in the filtered datasets, which might propagate to models trained on the data. Specifically, we report the Pearson's correlation between the gold standard toxicity label and whether or not it contains NOI, OI, or ONI mentions. These correlations are denoted as R_{ONI}, R_{NOI}, and R_{OI}, respectively; lower values indicate reduction in lexical biases.

Baselines We consider comparison against two natural baselines: a vanilla RoBERTa-large classifier trained on the original dataset (Original). We also consider a baseline trained on a random selection of the training data (Random), for comparison with data filtering methods for debiasing. Each subset is trained on 33% of the training data.

4.2 Results for Lexical Bias Reduction

First, we measure the reduction in lexical biases in filtered datasets, as given by AFLite and DataMaps. As shown in Table 1, subsets given by AFLite and the ambiguous and hard regions produced by DataMaps reduce the overall associations between TOXTRIG words and toxicity, compared to the original and random baselines; DataMaps-Hard has the largest reduction. On the other hand, as expected, DataMaps-Easy shows an *increased* association between TOXTRIG mentions and toxicity, showing that the these examples display overt lexical biases.

Table 2 shows results for lexical bias reduction using both debiased training approaches, as well as models trained on datasets filtered using AFLite and all three regions from DataMaps. Both debiased training approaches, LMIXIN-ONI and LMIXIN-TOXTRIG, reduce FPR_{ONI} as well as FPR_{OI} by a large amount. However, both approaches also hurt in-distribution test performance, indicating that ONI and other TOXTRIG features are essential for good performance.[11] In contrast, the models trained on hard and ambiguous subsets from DataMaps both preserve in-distribution performance, even though they are trained only a third of the original data. They also reduce the rate of falsely predicting NOI mentions as toxic (FPR_{NOI}), while not showing much improvement for ONI and maintaining FPR_{OI} of the original baseline.

Surprisingly, the model trained on the easy subset from DataMaps shows good bias reduction on the NOI and ONI categories, while matching the random selection baseline for OI. This is despite DataMaps-Easy showing an increased association between TOXTRIG mentions and toxicity (Table 1). Notably, the F_1 for all categories suffers under this model, indicating that it is less competent than the baseline. These results suggest that reduced associations in the data might not necessarily lead to debiased models trained on the same data. Overall, no single approach outperforms all others across different categories for lexical debiasing.

4.3 Qualitative Analysis

A qualitative study of the Founta et al. (2018) test set shows the presence of many annotation errors. We show three representative annotation errors in Table 3. The first example contains an atypical example of toxicity, towards white folks, which the annotators might have been unaware of. It also contains a link which annotators had access to, but not models. The second contains the word *p*ss* which the annotators may have relied for their assessment. The third encourages violence/abuse towards an identity which isn't typically the target of violence. Interestingly, the DataMaps-Easy predictions agree with all the gold standard annotations; perhaps such annotation errors and ambiguity are responsible for the performance discussed

[11] When we combine the bias-only model and the full model, we obtain competitive performance (see Appendix A.4).

		Test (12893)		NOI (602)		OI (553)		ONI (3236)	
		Acc.↑	F_1 ↑	F_1 ↑	FPR_{NOI} ↓	F_1 ↑	FPR_{OI} ↓	F_1 ↑	FPR_{ONI} ↓
	Vanilla	$94.21_{0.0}$	$92.33_{0.0}$	$89.76_{0.3}$	$10.24_{1.3}$	$98.84_{0.1}$	$85.71_{0.0}$	$97.34_{0.1}$	$64.72_{0.8}$
	LMIXIN-ONI	$89.65_{1.5}$	$85.59_{2.5}$	$87.04_{1.1}$	$13.99_{1.5}$	$98.87_{0.0}$	$85.71_{0.0}$	$87.87_{4.5}$	$\mathbf{43.74}_{3.1}$
	LMIXIN-TOXTRIG	$90.44_{0.7}$	$86.94_{1.1}$	$85.47_{0.3}$	$11.15_{1.7}$	$97.64_{0.3}$	$\mathbf{71.43}_{0.0}$	$90.41_{1.8}$	$44.55_{1.5}$
33% train	Random	$94.07_{0.1}$	$92.18_{0.1}$	$89.48_{0.4}$	$9.33_{0.7}$	$\mathbf{98.93}_{0.0}$	$\mathbf{83.33}_{3.4}$	$97.40_{0.1}$	$67.15_{0.6}$
	AFLite	$93.86_{0.1}$	$91.94_{0.1}$	$\mathbf{90.21}_{0.4}$	$11.26_{1.1}$	$98.90_{0.0}$	$85.71_{0.0}$	$97.32_{0.1}$	$67.97_{3.4}$
	DataMaps-Ambig.	$94.33_{0.1}$	$92.45_{0.1}$	$89.16_{0.7}$	$7.39_{1.0}$	$98.87_{0.0}$	$85.71_{0.0}$	$\mathbf{97.54}_{0.0}$	$64.39_{1.4}$
	DataMaps-Hard	$\mathbf{94.50}_{0.0}$	$\mathbf{92.61}_{0.1}$	$89.54_{0.4}$	$6.26_{0.9}$	$98.84_{0.0}$	$85.71_{0.0}$	$97.43_{0.0}$	$61.95_{1.1}$
	DataMaps-Easy	$94.00_{0.1}$	$91.94_{0.2}$	$86.81_{0.6}$	$\mathbf{5.92}_{0.7}$	$98.87_{0.0}$	$\mathbf{83.33}_{3.4}$	$97.17_{0.1}$	$\mathbf{60.33}_{3.8}$

Table 2: Evaluation of lexical bias removal for all debiasing methods on the Founta et al. (2018) test set. Results show the mean and s.d. (subscript) of accuracy and F_1 across 3 runs, as well as F_1 and false positive rate exclusively on test examples containing specific TOXTRIG mentions—NOI, OI and ONI, along with the number of examples in each category. The lower the FPR_*, the less the model infers lexical associations for toxicity. The first block shows debiased training approaches, along with the vanilla classifier, which are trained on the full dataset. The second block shows data filtering approaches, all trained on only 33% of the training data. Best performance in each block is boldfaced. ***Takeaway:*** While data filtering approaches achieve overall higher performance, debiased training approaches perform better on lexical bias reduction, in aggregate.

	Gold	DM-Hard	DM-Easy
@user THIS? LMAOOOOO...do yourself a solid and stay out of Black people's mentions and mind your caucasia...	🍃	☠	🍃
RT @user I wish I wasn't so annoying like I even piss myself off	☠	🍃	☠
@user If you want to attack people, attack fundamentalists of all faiths. Attack those who condemn 1.5bn people out of hand.	🍃	☠	🍃

Table 3: Examples of test set tweets with their gold-standard annotations and predictions from models trained on DataMaps-Hard (DM-Hard) and DataMaps-Easy (DM-Easy) subsets. ☠ denotes tweets with toxic labels, and 🍃 represents non-toxic labels. We anonymize the usernames to protect user privacy.

in §4.2. These annotation ambiguities might also impair our measurement for models' performance and debiasing effects, and expose a limitation of these datasets.

4.4 Adversarial Evaluation: ONI-Adv

To further study the reliance of debiased models on the ONI words, we use the test set from Dinan et al. (2019), which contains 3000 manually created sentences, 300 toxic. The toxic sentences barely contain any ONI mentions, e.g., *Cheese is made by french people who smell*. We call this test set ONI-Adv (for adversarial) since it challenges models with a reversal in the association between toxicity and offensive non-identity words (e.g., "*f*ck*", "*sh*t*").

We report F_1 for all models in Figure 2, which shows how well a model identifies toxicity in offensive tweets that do not contain overtly lexical cues of toxicity. The debiased training approaches improve over the baselines; data filtering methods do not. One reason for this might be that data filtering methods were trained on much less data than both LMIXIN models. Regardless, none of the models we test are good at predicting subtle, non-overt toxicity.

5 Experiments: Dialectal and Racial Biases

We test the efficacy of the bias reduction methods from §3 for dialectal bias (§2.2) reduction.

5.1 Dialectal Biases

For our dialectal bias experiments, we first infer the dialect of a tweet as described in §2.2. Then, analogous to the lexical bias evaluation, we quantify the dialectal debiasing using the Pearson's correlation between estimated probabilities of AAE and toxicity (R_{AAE}), and the false positive rates of models on AAE tweets (FPR_{AAE}). See Appendix A.3 for hyperparameters and other experimental settings.

Results in Table 4 show that almost all data filtering and debiasing methods reduce dialectal biases, with DataMaps-Easy as the exception (con-

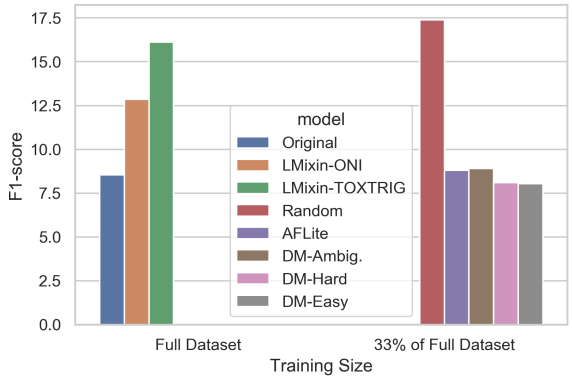

Figure 2: Challenge set evaluation for lexical biases, comparing all debiasing methods with baselines, using the ONI-Adv test set. ***Takeaway:*** $F_1(\uparrow)$ measures show that all models perform poorly at identifying toxic text not containing overtly lexical cues of toxicity. In general, debiased training approaches outperform the original model on this challenge set, while data filtering is not as effective.

		Test		
		$R_{\text{AAE}} \downarrow$	$F_1 \uparrow$	$\text{FPR}_{\text{AAE}} \downarrow$
	Vanilla	0.4079	$92.33_{0.0}$	$16.84_{0.3}$
	LMIXIN-Dialect	-	$92.26_{0.1}$	$16.07_{0.4}$
33% train	Random	0.4027	$92.18_{0.1}$	$16.67_{0.6}$
33% train	AFLite	0.3577	$91.94_{0.1}$	$16.84_{0.8}$
33% train	DataMaps-Ambig.	0.2965	$92.45_{0.1}$	$15.99_{0.4}$
33% train	DataMaps-Hard	**0.2878**	$\mathbf{92.61}_{0.1}$	$\mathbf{13.71}_{0.2}$
33% train	DataMaps-Easy	0.5347	$91.94_{0.2}$	$19.46_{2.8}$
	AAE-relabeled	0.3453	$91.64_{0.3}$	$\mathbf{12.69}_{0.0}$

Table 4: Dialectal bias evaluation for all debiasing methods (§5), as well as the relabeling approach (§6) on the Founta et al. (2018) test set. We report F_1 and the false positive rate with respect to tweets in AAE (FPR_{AAE}), reflecting dialectal bias (lower is less biased), showing mean and s.d. (subscript) across 3 runs. (Top Block) Debiased training approaches, along with the vanilla classifier, are all trained on the full dataset. (Middle Block) Random, AFLite and DataMaps all are trained on only 33% of the training data. Best performance for each training set size is in boldface. ***Takeaway:*** Both debiasing approaches improve performance over baselines, with DataMaps-Hard proving the most effective at debiasing. (Bottom Block) AAE-relabeling results in a model which despite following a noisy process yields even larger improvements for dialectal debiasing.

sistent with Table 1). Notably, DataMaps-Hard performs the best at dialectal debiasing, both in terms of toxicity-AAE correlation (R_{AAE}) and in terms of false flagging of toxicity (FPR_{AAE}). Interestingly, most models' decrease in false flagging is small, suggesting room for improvement.

5.2 Racial Biases

To quantify the real-world impact of dialect-based racial bias, we measure the rates of toxicity predicted by models on a corpus of tweets for which the race of authors is available, but not annotations of toxicity. Specifically, we consider the dataset released by Preoţiuc-Pietro and Ungar (2018), which consists of 5.4M tweets, collected from 4,132 survey participants (3,184 White, 374 African American) with self-reported race/ethnicity and Twitter user handles.[12]

We quantify our models' racial bias by measuring the difference in rates of flagging tweets by African American authors and those by white authors, following Sap et al. (2019).[13]

Listed in Table 5, our results show that automatic debiasing methods do not consistently decrease the racial discrepancy in flagging toxicity. Notably, the toxicity rates on tweets by African American authors—and the diferences compared to white authors—are similar across all debiasing methods and baselines, except for DataMaps-Easy, which shows the most racial bias in toxicity flagging. Surprisingly, DataMaps-Hard, which mitigated dialectal bias the best out of all debiasing methods, also shows high discrepancy between author races. Confirming previous results, this suggests that debiasing these systems requires more than automatic debiasing methods.

6 Towards Data Relabeling

Based on our quantitative and qualitative analyses, we believe there still is room for improvement in debiasing hate speech detection. Therefore, we turn our attention to the role of label noise in datasets. Partly inspired by our qualitative analyses of debiased models' predictions, we design a proof-of-concept study where we automatically correct the label of tweets using a(n automatic) dialectal translation of the tweet, inspired by previous work showing that highlighting AAE tweets' dialect led them to be labeled as less toxic (Sap et al., 2019). We conclude this study by discussing the limitations and ethical implications of the synthetic data, and cautioning against its real-world application.

[12]For efficiency, we randomly select 12k tweets from the dataset as the OOD test set.

[13]Note that we assume that authors from all races have the same likelihood of writing toxic language.

		W-Tox.	AA-Tox.	Δ ↓	AA/W↓
	Original	7.24	12.61	5.37	1.74
	LMIXIN-Dialect	7.50	12.55	5.06	1.67
33% train	Random	8.28	13.24	4.96	1.60
	AFLite	7.32	11.64	4.33	1.59
	DataMaps-Ambig.	6.75	12.17	5.42	1.80
	DataMaps-Hard	6.36	11.67	5.31	1.84
	DataMaps-Easy	8.46	16.30	7.83	1.94
	AAE-relabeled	6.93	10.60	**3.67**	**1.53**

Table 5: Racial disparity in toxicity prediction reported on Preoţiuc-Pietro and Ungar (2018). **W-Tox.** indicates % of white users' tweets being flagged as toxic, **AA-Tox.** indicates % of African American users' tweets being flagged as toxic, Δ refers to the difference between AA-Tox. and W-Tox., and **AA/W** refers to the ratio between AA-Tox. and W-Tox. ***Takeaway:*** Methods generally fail in debiasing on this OOD test set except the relabeling approach shows some benefit.

Focusing on dialectal bias, our key assumption is that an AAE tweet and its corresponding WAE version should have the same toxicity label, therefore toxic AAE tweets whose WAE versions are non-toxic are candidates for label correction.[14]

However, gold-standard translations of AAE to WAE would require qualified translators, and automatic AAE-to-WAE translation systems do not exist, to the best of our knowledge. Therefore, we create a proof-of-concept study—we set up a AAE to WAE "translation" system using the few-shot capabilities of the GPT-3 language model (Brown et al., 2020). Under this mechanism, we prompt GPT-3 with four translation pairs (taken from Spears, 1998) and an AAE tweet from our training data, and generate its WAE "translation". The list of prompts, as well as further details, are provided in Appendix C. Note that we do *not* recommend this approach to build large scale parallel data for dialects, as discussed under ethical implications and limitations.

Next, as per our heuristic, we only relabel toxic AAE tweets whose WAE translation is predicted as non-toxic by either our vanilla classifier trained on the original Founta et al. (2018) dataset, or an identical classifier trained on the WAE translated tweets. The resulting dataset (AAE-relabeled) is the same size as the original dataset, but with 954 (12%) out of 8260 toxic AAE tweets relabeled as non-toxic (examples in Table 6). To assess the validity of the relabeling, the first three authors manually annotated toxicity of 50 randomly selected relabeled tweets. On average, authors agreed with 84% of the relabeling decisions.

Then, we evaluate the dialectal bias of AAE-relabeled and quantify the dialect and racial prediction biases from a RoBERTa-large classifier trained on AAE-relabeled, following §5. As shown in the last row of Table 4, this relabeling scheme decreases dialectal bias more than any other debiasing method, specifically as measured by the FPR on AAE tweets, with one point drop in F_1 score. The F_1 score on the "gold" test data (Table 4) are not fully reliable, as test data contain label biases and better performance could come from exploiting these biases. As shown in Table 5, the model trained on AAE-relabeled has the lowest racial disparity in toxicity flagging rates compared to all other methods.

These results highlight that debiasing methods are much less effective at mitigating dialectal dataset biases compared to data relabeling. For future investigations, we recommend obtaining human-written AAE-WAE pairs (e.g., as done by Groenwold et al., 2020). Additionally, to ensure less biased toxicity labeling, we recommend recruiting AAE speakers or experts for avoiding over-identification of AAE-markers as toxic (Spears, 1998; Croom, 2013). Alternatively, we recommend exploring more holistic representations of social biases or toxicity (e.g., Social Bias Frames; Sap et al., 2020).

Ethical Implications & Limitations

The above synthetic setting is meant to illustrate the role of labeling quality on biases in annotations. We strongly caution against using this approach in real-world applications, such as building parallel datasets for dialects. First, due to how its training data was selected, GPT-3 has likely not been exposed to many African American English varieties during training (Jo and Gebru, 2020). Second, pretrained language models are known to generate toxic language at non-trivial rates (Gehman et al., 2020), which could cause differential toxicity in the translations.

7 Related Work

Debiasing Toxicity Detection As the popularity of hate speech and toxic language detection sys-

[14] Note that this assumption does not hold for lexical items, because substituting lexical items (e.g., swapping a minority mention for a majority mention) would drastically change the denotational meaning of the sentence.

AAE	GPT-3 WAE Translation	Gold	New
RT @user I can't stand a bad texter bruh like don't be mad if I forget about yo ass	RT @user I can't stand a bad texter bro like don't be mad if I forget about you	☠	🍃
RT @user Retweet if you fuck with this!!!!	RT @user Retweet if you like this!	☠	🍃
RT @user That nigga needs anger management	RT @user That guy needs anger management	☠	🍃
RT @user oh fucking hell take a day off man	RT @user oh fuck take a day off man	☠	☠

Table 6: Examples of AAE tweets with their GPT-3 based WAE translation, and original gold standard and new annotations based on AAE-relabeled. For the first three tweets, the (biased) gold labels are changed by models predicting the new labels on their WAE translations. ☠ indicates presence of toxicity, and 🍃 represents non-toxic. We anonymize the usernames to protect user privacy.

tems has grown, several biases have been found in dataset and models, spurring various debiasing efforts to mitigate these individual biases (e.g., gender bias, racial bias; Park et al., 2018; Sap et al., 2019; Davidson et al., 2019). Some work tackles identity-based biases, e.g., using data re-balancing (Dixon et al., 2018), or adversarial feature learning (Vaidya et al., 2019). Less work has tackled racial or dialectal bias. Notably, Xia et al. (2020) use adversarial training to prevent the model from associating toxicity with AAE, showing only small improvements in fairness. Based on those results, we do not explore adversarial methods, opting instead for ensemble-based methods of predefined bias reduction. In contemporary work, Mozafari et al. (2020) use a re-weighting mechanism, which shows some effects in debiasing racial bias. We leave it for future work to evaluate this method in our setting. In contrast to all previous work, our experiments also measure the effectiveness of bias-agnostic methods.

Other General Debiasing Methods Several approaches for debiasing NLU tasks have been proposed lately. Some approaches rely on adversarial training to remove protected attributes (e.g. gender or race), from a model's internal representations (Zhang et al., 2018; Wang et al., 2019; Xia et al., 2020). Other approaches include confidence regularization (Utama et al., 2020), as well as other product of expert approaches (He et al., 2019; Karimi Mahabadi et al., 2020) similar to the debiased training approach from Clark et al. (2019), which is the only debiased training we employ due to its relatively strong performance.

8 Conclusion

We investigate whether toxic language detection systems can be debiased using recently introduced methods for debiasing text classification in NLU tasks. Focusing on two types of biases, lexical and dialectal, our experiments show that these methods face significant challenges in reducing the biased behavior in toxicity detectors. This indicates that biases in toxic language detection might be different in nature compared to spurious associations studied in typical NLU settings. We studied a synthetic scheme for relabeling examples with potential dialectal biases; our results indicate that correcting noisy labels results in better bias reduction. Our findings suggest that instead of solely relying on development of automatic debiasing for existing, imperfect datasets, future work focus primarily on the quality of the underlying data for hate speech detection, such as accounting for speaker identity and dialect. Indeed, such efforts could act as an important step towards making systems less discriminatory, and hence safe and usable.

Acknowledgments

We thank the anonymous reviewers and Laura Vianna for helpful comments on this work. This research was supported in part by NSF grants 1813153 and 1714566.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé, III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proc. of ACL*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proc. of EMNLP*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proc. of ICML*.

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. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proc. of EMNLP*.

Adam M Croom. 2013. How to do things with slurs: Studies in the way of derogatory words. In *Language & communication*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Abusive Language Workshop (at ACL)*.

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*.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2020. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. In *Sexuality & Culture*.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proc. of EMNLP*.

Lucas Dixon, John Li, Jeffrey Scott Sorensen, Nithum Thain, and L. Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proc. of AES*.

Marta Dynel. 2012. Swearing methodologically : the (im)politeness of expletives in anonymous commentaries on youtube. In *Journal of English Studies*.

Marta Dynel. 2015. The landscape of impoliteness research. In *Journal of Politeness Research*.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proc. of WSM*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*.

Lisa Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American vernacular english in Transformer-Based text generation. In *Proc. of EMNLP*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of NAACL*.

Jessica Guynn. 2020. What civil rights groups want from facebook boycott: Stop hate speech and harassment of black users.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Proc. of NeurIPS*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *EMNLP Workshop on Deep Learning Approaches for Low-Resource NLP*.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proc. of FAT*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proc. of ACL*.

Gabriele Kasper. 1990. Linguistic politeness: current research issues. In *Journal of Pragmatics*. Elsevier.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets.

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. In *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. In *PLOS ONE*. Public Library of Science.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proc. of EMNLP*.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *Proc. of COLING*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*, pages 2383–2392.

Sarah T Roberts. 2019. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Jonathan Rosa. 2019. *Looking like a language, sounding like a race*. Oxford University Press.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. In *Language In Society*. Cambridge University Press.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: the case of the european refugee crisis. In *NLP 4 CMC Workshop*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. of ACL*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proc. of ACL*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *Proc. of CoNLL*.

Arthur K Spears. 1998. African-American language use: Ideology and so-called obscenity. In *African-American English: Structure, History and Use*. Routledge New York.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proc. of EMNLP*.

Björn Technau. 2018. Going beyond hate speech: The pragmatics of ethnic slur terms. *Lodz Papers in Pragmatics*, 14(1):25–43.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proc. of ACL*.

Ameya Vaidya, Feng Mai, and Yue Ning. 2019. Empirical analysis of multi-task learning for reducing model bias in toxic comment detection. In *Proc. of ICWSM*.

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019. How much online abuse is there? In *Alan Turing Institute*.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and V. Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proc. of ICCV*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proc. of Social NLP*.

Danyaal Yasin. 2018. Black and banned: Who is free speech for?

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proc. of AES*. Association for Computing Machinery.

Appendix

A Further Details for Models

A.1 Model Debiasing

The LEARNED-MIXIN ensemble allows the model to explicitly determine how much to trust the bias given the input:

$$\hat{p_i} = \text{softmax}\{\log(p_i) + g(\mathbf{x}_i) \log b_i\}$$

where $\mathbf{x}_i$ is the ith input text, p_i and b_i is the toxicity prediction produced by RoBERTa, and bias-only model respectively, and g is a parametric function, which is defined as $\text{softplus}(\mathbf{w} \cdot \mathbf{h}_i)$, where $\mathbf{w}$ is a learned vector, $\mathbf{h}_i$ is the last hidden layer of the model for example $\mathbf{x}_i$, and the $\text{softplus}(x) = \log(1 + \exp x)$. To prevent the LEARNED-MIXIN ensemble from ignoring b_i, Clark et al. (2019) add an entropy penalty (H) to the loss:

$$R = \alpha H(\text{softmax}\{g(\mathbf{x}_i) \log b_i\})$$

Where $H(z) = -\sum_j z_j \log z_j$ is the entropy and α is a hyperparameter.

A.2 Data Filtering

For the data filtering methods, we first filter data to 50% of the original data as in Swayamdipta et al. (2020). Then we further downsample the dataset to 33% of the original data to control that each training set has the same toxic ratio as the original training set. This step is to avoid confounding our results with different toxic ratio among different training sets.

A.3 Training Settings

For all the experiments, we fine-tune RoBERTa-large (Liu et al., 2019) over the corresponding corpus with one GTX2080 Ti. We use the default hyperparameters as provided in the `HuggingFace Transformers` library (Wolf et al., 2019), with two major changes: we use a learning rate of 10^{-5} and 8 batch size in all experiments.

A.4 Prediction Combining with Bias-only Model

To prevent the possibility that our LMIXIN-TOXTRIG/ONI is not well trained, thus resulting in the decrease of models' in-distribution performance, we use the joint-prediction from the main and bias-only model to infer the in-distribution test set and they obtain 94.15% and 94.17% accuracy, respectively. This is competitive performance as shown in Table 2.

B Alternative Dataset of Toxic Language

Davidson et al. (2017) collected data from Twitter, starting with 1,000 terms from HateBase (an online database of hate speech terms) as seeds, which the process relies on lexical biases. We find that performing data filtering methods over this dataset leads to degenerate behaviour. Specifically, as shown in Table 7, the easy region demonstrates least spurious correlation due to its heavily skewed class distribution, which further prevent us from downsampling to control the toxic ratio. We also train LMIXIN-TOXTRIG and LMIXIN-dialect over the dataset. Table 8 shows that FPR of the debiased model increase instead except for the OI category and Table 9's results behave in-line with Table 4.

C Few-shot AAE-to-WAE Translation

Note that we do *not* recommend the following approach to build large scale parallel data for dialects, as discussed under ethical implications and limitations (§6).

We use GPT-3 (Brown et al., 2020) to create a few-shot AAE-to-WAE translation system, using the following set of example translation pairs drawn from Spears (1998):

AAE: Get your triflin' ass out of here.
WAE: Get your trifling self out of here.

AAE: I saw his ass yesterday.
WAE: I saw him yesterday.

AAE: His ass is gonna get fried.
WAE: He is gonna get fried

AAE: Wassup, nigga?
WAE: What's up bro?

AAE: ⟨tweet⟩
WAE:

Note that Spears (1998) refers to WAE as White language varieties, and deals with English prevalent in the United States.

We prepend the formatted example pairs to each AAE tweet in our training data, and generate the translation from GPT-3 using top-0.95 nucleus sampling with a temperature of 0.5. Prompts, formatting, and generation parameters were chosen based on manual inspection of the output.

	Toxic Ratio	$R_{\text{NOI}} \downarrow$	$R_{\text{OI}} \downarrow$	$R_{\text{ONI}} \downarrow$	$R_{\text{AAE}} \downarrow$
Original†	0.8308	0.0287	0.4320	0.2610	0.4061
Random	0.8312	0.0288	0.4312	0.2621	0.4011
AFLite	0.7669	0.0342	0.4708	0.2835	0.4236
DataMaps-Ambig.	0.6736	0.0493	0.4683	0.3230	0.4445
DataMaps-Hard	0.6645	0.0521	0.4533	0.3190	0.4426
DataMaps-Easy	0.9972	0.0135	0.0771	0.0396	0.0928

Table 7: Lexical and dialectal associations between toxicity in the original dataset (Davidson et al., 2017) and various filtered counterparts. Random, AFLite, and DataMaps all contain only 50% of the original data after filtering. (We could not perform downsampling on these datasets due to their heavily skewed label distribution.) Lower Pearson R correlation value indicates less superficial patterns in the dataset, thus are less biased. The easy subset gives the best results here are due to its severe inbalanced label distribution.

	Test		NOI		OI		ONI	
	Acc.↑	$F_1 \uparrow$	$F_1 \uparrow$	$\text{FPR}_{\text{NOI}} \downarrow$	$F_1 \uparrow$	$\text{FPR}_{\text{OI}} \downarrow$	$F_1 \uparrow$	$\text{FPR}_{\text{ONI}} \downarrow$
Original	96.37	97.81	96.42	25.00	99.86	57.14	99.57	63.64
LMIXIN-TOXTRIG	96.15	97.69	96.19	28.57	99.78	42.86	99.28	72.73

Table 8: Lexical bias removal evaluation for debiasing methods. Original refers to the model trained over the full training set. The test set is further categorized into tweets that contained relevant TOXTRIG words. F_1 indicates models' performance while the false positive rate (FPR_*) reflects models' bias. The lower the FPR_* is, the less biased the model tend to be.

Debiasing Method		Test		
	R_{AAE}	Acc. ↑	$F_1 \uparrow$	$\text{FPR}_{\text{AAE}} \downarrow$
Original	0.4079	96.37	97.81	24.76
LMIXIN-Dialect	-	96.48	97.88	22.86

Table 9: Dialectal bias evaluation for all debiasing methods, on both in-distribution test set as well as out-of-distribution dialect and race priming test sets. In addition to accuracy and F_1, we report the false positive rate with respect to tweets in AAE (FPR_{AAE}), reflecting dialectal bias (lower is less debiased). Each method is based on a RoBERTa-large classifier.