Does Robustness Improve Fairness? Approaching Fairness with Word Substitution Robustness Methods for Text Classification

Yada Pruksachatkun *
Amazon Alexa
ypruksac@amazon.com

Satyapriya Krishna *
Amazon Alexa
satyapk@amazon.com

Jwala Dhamala
Amazon Alexa
jddhamal@amazon.com

Rahul Gupta
Amazon Alexa
gupra@amazon.com

Kai Wei Chang
UCLA, Amazon Alexa
kwchang@cs.ucla.edu

Abstract

Existing bias mitigation methods to reduce disparities in model outcomes across cohorts have focused on data augmentation, debiasing model embeddings, or adding fairness-based optimization objectives during training. Separately, certified word substitution robustness methods have been developed to decrease the impact of spurious features and synonym substitutions on model predictions. While their end goals are different, they both aim to encourage models to make the same prediction for certain changes in the input. In this paper, we investigate the utility of certified word substitution robustness methods to improve equality of odds and equality of opportunity on multiple text classification tasks. We observe that certified robustness methods improve fairness, and using both robustness and bias mitigation methods in training results in an improvement in both fronts.

1 Introduction

As natural language processing (NLP) technologies are increasingly used in essential real-world applications, such as social media, healthcare, personal assistants and law (He et al., 2020; Ahmad et al., 2020), it is important to ensure these systems do not create unintended outcomes for end-users or offer disparate experiences to customers from diverse backgrounds. This includes ensuring that model performance does not significantly differ across people belonging to different cohorts, such as different gender or race groups.

A major subset of industry NLP applications lies in text classification, such as domain and intent classification in voice assistants (Su et al., 2018) or code tagging in healthcare (Kemp et al., 2019). In this study, we focus on toxicity classification (Dixon et al., 2018) and occupation classification of Wikipedia biographies (De-Arteaga et al., 2019). For toxicity classification, ensuring fairness means ensuring that a model can identify toxicity to a similar accuracy across all examples regardless of the protected groups present in the example. Past studies (e.g, (Dixon et al., 2018; Zhang et al., 2020; Zhao and Chang, 2020)) have shown that toxicity classification models will falsely classify text containing certain protected attributes as toxic. Leading social media platforms and internet companies use toxicity classification models for content moderation (Gorwa et al., 2020), thus having bias in such models can lead to increased silencing of under-served groups. Similarly, for occupation classification, a fair model should correctly identify occupations given a biography, regardless of the protected group that a person belongs to (De-Arteaga et al., 2019).

Recently, several studies have demonstrated societal bias in NLP systems (Hutchinson et al., 2020; Tan and Celis, 2019; Liang et al., 2020) and various approaches have been proposed to mitigate the bias. These approaches include creating balanced datasets (Park et al., 2018; Zhao et al., 2018a), developing methods optimized for particular fairness notions (Zhang et al., 2017, 2020), model calibration (Zhao et al., 2017; Jia et al., 2020), and reducing representational bias (Bolukbasi et al., 2016b; Zhao et al., 2018b; Liang et al., 2020).

Separately, certified robustness approaches (Jia et al., 2019; Ye et al., 2020) have been developed to ensure robustness against word substitution attacks. Specifically, these strategies ensure small perturbations in the input embedding space do not alter model predictions. Despite never having been discussed in prior literature, this corresponds to notions of fairness, since protected attribute information (e.g. gender) is often irrelevant to the task

* Equal contribution

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3320–3331
August 1–6, 2021.

at hand (i.e. "She is a good singer" and "He is a good singer" should have the same sentiment label). Thus, we posit that word substitution robustness methods can be used to make models invariant to protected attribute tokens and identifiers.

We explore the effect of robustness methods on fairness with GloVe-based CNN models (Kim, 2014) trained with Interval Bound Propagation (IBP) (Jia et al., 2019), and BERT (Devlin et al., 2019a) trained with SAFER (Ye et al., 2020). We compare the effect of these robustness methods to popular bias mitigation methods. We find that robustness methods achieve promising performance on fairness metrics exceeding that of bias mitigation methods in several text classification tasks on gender and sexual orientation dimensions. Furthermore, training on both fairness and robustness exceeds performance over robustness and bias mitigation methods alone. Comprehensive analysis and visualization demonstrate that the robust methods decrease feature importance on gender tokens.

Our contributions are two-fold. First, we show that certified robustness methods can be used and integrated with bias mitigation methods to effectively improve models' performance on several notions of fairness, notably equality of opportunity and equality of odds. Secondly, by integrating robustness methods with fairness, we can improve a model's robustness while reducing bias, which is important in creating trustworthy NLP systems. Our study's practical implications include applications to models used in the industry that can handle customer inputs that may differ from the training data (robust) and that minimize any unintended consequences on the customers (fair). With this study, we aim to motivate future work geared towards developing methods that jointly optimize for multiple trustworthy aspects of models; specifically, those addressing model robustness and fairness.

2 Mitigating Bias through Certified Robustness Methods

In the following, we first define the notions of fairness considered in this paper. Then, we discuss certified robustness methods, and how they can be applied to reduce bias in models.

2.1 Fairness Notions

We focus on measuring two notions of fairness in this paper – equalized odds and equality of opportunity, as they are commonly used in quantifying bias in NLP applications. We describe the metrics associated with these notions in Section 3. We give application examples of these notions on toxicity and occupation classification for English texts.

Equalized Odds A model achieves *Equalized Odds* (Hardt et al., 2016) with respect to a protected attribute A and outcome Y if $P(\bar{Y} = 1|A = 0, Y = y) = P(\bar{Y} = 1|A = 1, Y = y)$, for $y \in \{0, 1\}$. Protected attributes are traits or characteristics that cannot be discriminated against by law[1]. Intuitively, this means that the model should have equal true positive and false positive rates across groups. For toxicity classification, equalized odds implies that a model should be able to effectively detect toxicity on comments that include identifiers across all protected attribute cohorts, while not silencing any one cohort. Prior studies demonstrate that models disproportionately predict sentences associated with LGBTQ individuals as toxic, which may further silence discussion around LGBTQ issues and the voices of LGBTQ people (Oliva et al., 2020).

Equality of Opportunity A model achieves *Equality of Opportunity* (Hardt et al., 2016) with respect to a protected attribute A and outcome Y if $P(\bar{Y} = 1|A = 0, Y = 1) = P(\bar{Y} = 1|A = 1, Y = 1)$. This is a relaxation of Equalized Odds to the positive outcome, in which the model must have equal true positive rates across groups. For occupation classification, equality of opportunity implies that a model is able to correctly classify biographies of people from all groups, thus enabling equity in positive outcomes such as appropriate and useful matches in job recommendation sites.

Due to bias in the training data, off-the-shelf models often contain biases and disparities in model performance against underrepresented groups. Various bias mitigation approaches have been proposed to ensure the fairness in model predictions. We include a diverse array of bias mitigation methods, spanning embedding debiasing, in-training, and post-processing, as baselines. These consist of instance weighting (Zhang et al., 2020), *HardDebias* word embeddings (Bolukbasi et al., 2016a), and adversarial debiasing (Zhang et al., 2018). See more discussion in Sec. 3.

[1] https://www.eeoc.gov/employers/small-business/3-who-protected-employment-discrimination

2.2 Certified Robustness for Bias Mitigation

Designed for a different purpose, certified robustness methods present ways to train models that satisfy guarantees of word substitution robustness. By adapting certified robustness methods to fairness applications, we aim to make models invariant to spurious protected attribute information present in inputs, and thus improve in equality of opportunity and equalized odds.

Formally, a model f is *certifiably robust* if, for any example sentence x, and sentences x' that consist of x modified with word substitutions, $f(x) = f(x') = y$. In the robustness context, word substitution consists of swapping a word with its synonyms (usually defined using retrofitted word embeddings). For example, if x = "The waiter talked to the customer about their problems," x' may consist of the sentences "The waitress talked to the customer about their qualms." In the context of fairness, we consider 'waiter' and 'waitress' or gender pronouns to carry the same meaning in the context of toxicity and occupation classification, and to have the same label.

In this paper, we use two recently developed certified robustness methods, Interval Bound Propagation (IBP) (Jia et al., 2019) and SAFER (Ye et al., 2020). Given a set of perturbations for each word, these two models ensures that word substitution do not affect model predictions. In particular, for each word, and a polytope spanned by the potential substitutions for that word in the embedding space, these methods ensure that swapping the word with any point in the polytope will not change the model predictions. To accomplish this, IBP minimizes the upper bound of the set of losses over perturbation sets, and SAFER uses a model-agnostic randomized smoothing technique.

Both IBP and SAFER encourage models to be robust to spurious word substitutions, which include tokens that contain protected attribute information. The perturbations included in the original paper from Alzantot et al. (2018) are based on a GloVe embedding that has been modified such that synonyms are close together. While the perturbation set does not include explicit gender and sexual orientation swaps ('boy' is not included in the perturbation set for 'girl', while 'girls', and 'women' are), we posit that certified robustness methods can still be applied to bias mitigation by improving robustness in examples that contain identifiers of underrepresented groups. Doing so will decrease model performance disparity in underrepresented group cohorts, and thus fulfill fairness notions.

IBP (Jia et al., 2019) IBP computes bounds on the model loss based on bounds on the input. The robustness goal of the IBP method is to minimize $\max F(x, \theta)$. Here, $F(x, \theta)$ denotes the set of losses of a model over $B_{perturb}$, where $B_{perturb}$ is the set of perturbations for an example sentence x. Formally, $F(x, \theta) = (f(\bar{x}, \theta) | \bar{x} \in B_{perturb})$. The full loss for IBP is $(1 - \lambda) f(x, \theta) + \lambda \mu^{final}(x, \theta)$, where μ^{final} is the upper bound on the loss $f(x, \theta)$ and $\lambda \in [0, 1]$.

SAFER (Ye et al., 2020) Unlike IBP, SAFER does not require any changes to the model training. Instead, it employs a randomized smoothing mechanism in which an input is perturbed before being fed to the model during the training time. Specifically, SAFER creates random word substitutions using a perturbation set derived from a synonym network. Ye et al. (2020) determine certified robustness of a model on an example by certifying that, given an example z, model score $s(z)$, and $y_B = \text{argmax}_{c \in Y, c \neq y}\, s(z)$, the model score of the gold label y is higher than the model score of the highest scoring non-gold label y_B by a constant.

3 Empirical Study on the Connection between Fairness and Robustness

To better understand the connection between fairness and certified robustness in the context of text classification, we empirically analyze models augmented with various combinations of robustness and fairness methods, as enumerated below.

1. **Classifier (Baseline)**: The base text classification models. We consider two types of classification models that widely used in the literature, CNN (Kim, 2014) and BERT (Devlin et al., 2019b).
2. **Classifier + Fairness**: Text classifiers trained with bias mitigation techniques (see Sec. 2).
3. **Classifier + Robustness**: Text classifiers trained with robustness methods (see Sec. 2).
4. **Classifier + Robustness + Gender Word Perturbations**: To ensure that the model becomes robust against gender substitutions, we add definitional gender pairs (e.g., swapping he with she) (Bolukbasi et al., 2016b) in the permutation set of IBP and SAFER.
5. **Classifier + Robustness + Fairness**: Text classifier trained with both fairness and ro-

bustness objectives.

We aim to answer the following research questions based on the aforementioned configurations. (1) *What is the effect of robustness methods on mitigating bias* (compare configuration 3 with 1 and 2)? (2) *What is the effect of adding gender word substitutions to the robustness perturbation sets* (compare configurations 3 and 4)? (3) *What is the effect of integrating bias mitigation and robustness methods* (compare configuration 5 with 1 and 3)?.

In particular, to answer the last question, we consider combining popular bias mitigation approaches with IBP as follows.

- **Debiased Word Embeddings + IBP**: We replace the GloVe embeddings in the baseline CNN model with the *HardDebias* embeddings obtained from (Bolukbasi et al., 2016b), while keeping the rest of the IBP training methodology the same.
- **Instance weighting + IBP**: We add the instance weights to each sample in the loss computation during IBP training.
- **Adversarial Training + IBP**: We perform multitask training, alternating between optimizing for robustness loss and adversarial debiasing loss. We initialized our adversarial training with the IBP-trained model.

Datasets We use the following two text classification datasets to validate our hypothesis on different data distributions.

- *Jigsaw Toxicity*[2] is a dataset for toxicity classification that consists of 1,804,874 training examples, which we split into train and validation sets of size 1,443,900 and 360,974 respectively. We take 97,320 examples from the public leaderboard as the test set.
- *Bias in Bios* (De-Arteaga et al., 2019) [3] is a dataset for occupation classification derived from Common Crawl corpus. It consists of 178,619 train and 91,917 test examples.

Evaluation Metrics We evaluate models on three dimensions: (1) raw task performance, (2) model fairness, and (3) model robustness. For the raw task performance, we follow prior work in using accuracy and area under the ROC curve (AUC) to evaluate the performance of a model on the Bias in Bios dataset and the Jigsaw Toxicity dataset respectively. To measure the robustness of a model, we follow Jia et al. (2019) and Ye et al. (2020) to use the certified robustness accuracy (CRA). For fairness, we follow the discussion in Section 2.1 to evaluate a model based on *equalized odds* and *equal opportunity*.

For fairness, we measure two metrics - i.e, True Positive Equality Difference (TPED) and False Positive Equality Difference (FPED). The FPED and TPED is calculated as:

$$\sum_{z \in Z} |f_z - f_{overall}|,$$

where f is FPR or TPR depending on whether we are computing FPED or TPED, and Z refers to the set of all classes in a protected group. Note that TPED and FPED metrics do not take into account how well the model does - for example, a model that achieves a true positive rate of 0.0 for all groups will still have a TPED of 0.

For Bias in Bios dataset, we chose equality of opportunity to measure fairness, since it is important to ensure job candidates are matched with job recommendations that are relevant to them. Since equality of opportunity necessitates equality in true positive rates across cohorts, we use TPED as the fairness evaluation metric for Bias in Bios. For Jigsaw Toxicity, we define fairness by equalized odds, since it is important for toxicity classifiers to be able to detect toxicity in content containing identifiers across all groups, while not silencing any one. The combination of FPED with TPED aligns with the *Equalized Odds* definition of fairness (Borkan et al., 2019), thus we define a score *EOdds* as FPED + TPED for ease of analysis. Equalized odds is satisfied when FPED = 0 and TPED = 0, and thus when *EOdds* = 0.

For the scope of this paper and the limitations of the dataset, we study binary gender for Bias in Bios, and both gender (male, female, transgender, and non-binary) and sexual orientation (homosexual/straight, heterosexual, gay, lesbian, bisexual) for Jigsaw Toxicity classification. While we acknowledge that there are a multitude of important attributes, we constrain the scope of this study to the attributes present in text classification datasets.

Experiment Details All the experiments were conducted on p3dn.24xlarge and p3.2xlarge AWS compute nodes.[4] The IBP runs took 48 hours for

[2]The data is available at https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[3]The data is available at https://github.com/microsoft/biosbias

[4]https://aws.amazon.com/ec2/instance-types/

Model	Raw task (↑)	Fairness (↓)			Robustness (↑)
	AUC	EOdds	FPED	TPED	CRA
Baseline	**0.957**	0.508	0.197	0.311	0.270
IBP	0.913	0.184	0.005	0.179	**0.934**
$\text{IBP}_{\text{gender}}$	0.947	0.237	0.062	0.175	0.912
Instance weighting	0.955	0.505	0.196	0.309	0.214
HardDebias	0.951	0.525	0.221	0.304	0.404
Adversarial Training	0.955	0.491	0.198	0.293	0.644
IBP + Instance weighting	0.889	**0.165**	**0.002**	**0.163**	0.942
IBP + *HardDebias*	0.923	0.459	0.169	0.290	0.890
IBP + Adversarial Training	0.920	0.473	0.192	0.281	0.901

Table 1: Certified robustness and bias mitigation methods with CNN on Jigsaw dataset. The best performance for each column is boldfaced. Results show that the certified robustness method (IBP) improves both robustness and fairness with performance drops on the raw task accuracy.

Model	Raw task (↑)	Fairness (↓)	Robustness (↑)
	AUC	TPED	CRA
Baseline	**0.787**	0.131	0.115
IBP	0.743	0.127	0.702
$\text{IBP}_{\text{gender}}$	0.749	0.104	0.711
Instance weighting	0.755	0.118	0.095
HardDebias	0.767	0.106	0.070
Adversarial Training	0.773	0.114	0.180
IBP + Instance weighting	0.732	0.113	**0.719**
IBP + *HardDebias*	0.735	**0.101**	0.715
IBP + Adversarial Training	0.725	0.112	0.693

Table 2: Experiment results on CNN models on the Bias in Bios dataset. We see that our best performing model consists of initiating IBP training with $HardDebias$ embeddings.

Jigsaw Toxicity and 34 hours for Bias in Bios, while SAFER took 53 hours with evaluation for Jigsaw Toxicity and 37 hours for Bias in Bios.

For the experiments with CNN, we follow Jia et al. (2019) to configure the IBP schedule and CNN models. In particular, we used a CNN model with a hidden size of 100 and kernel size of 3 with the GloVe embedding (Pennington et al., 2014) as inputs. For IBP, we linearly increased the weight on the certified robustness objective from 0 to 0.8 for 40 epochs, before training for 20 epochs on the full certified robustness objective.

For the experiments with BERT, we follow Ye et al. (2020) to configure the BERT model and SAFER experiment. We use `bert-base-uncased`, and take the top-100 words that are closest in cosine similarity for each token as the token's perturbation set. We describe the remaining hyper-parameter details (learning rate, epochs, dropout probability) in the the appendix, which we obtained after a hyper-parameter search on the development set.

4 Results

The results for Jigsaw Toxicity and Bias in Bios are in Tables 1, 2, 3 and 4.

Effect of certified robustness methods for mitigating bias We observe that adding IBP during training achieves better performance on fairness over othe bias mitigation approaches across Jigsaw Toxicity and Bias in Bios. In Jigsaw Toxicity, EOdds improves from 0.508 to 0.184, and in Bias in Bios, TPED improves from 0.131 to 0.127. Similarly, training models with SAFER results in an improvement in performance in all fairness metrics, with an improvement in EOdds from 0.553 to 0.286 in Jigsaw Toxicity and an improvement in TPED from 0.148 to 0.134 in Bias in Bios.

Effect of adding gender word substitutions to the robustness perturbation sets While adding gender word substitutions further improves fairness in Bias in Bios, it results in worse fairness scores in Jigsaw Toxicity than plain certified robustness methods. In Bias in Bios, IBP_{gender} results

Model	Raw task (↑)	Fairness (↓)			Robustness (↑)
	AUC	EOdds	FPED	TPED	CRA
Baseline	0.914	0.553	0.290	0.263	0.950
SAFER	0.918	**0.286**	**0.144**	**0.142**	**0.967**
$\text{SAFER}_{\text{gender}}$	**0.968**	0.347	0.176	0.171	0.917

Table 3: Model performance on BERT (Baseline) and SAFER on the Jigsaw dataset. Similar to the observation with IBP, SAFER improves both the fairness and robustness metrics.

Model	Raw task (↑)	Fairness (↓)	Robustness (↑)
	AUC	TPED	CRA
Baseline	**0.796**	0.148	0.164
SAFER	0.744	0.134	0.726
SAFER_{gender}	0.761	**0.097**	**0.733**

Table 4: Model performance on BERT (baseline) with SAFER on the Bias in Bios dataset.

in a lower TPED than all fairness only baselines. This trend holds in SAFER, where SAFER_{gender} achieves lower TPED than SAFER. In Jigsaw Toxicity, adding gender words to the perturbation set degrades performance in equalized odds for both IBP and SAFER. This may be because the list of gender word substitutions do not include words relating to sexual orientation and non-binary gender, and thus may only improve fairness amongst examples containing male and female identifiers.

Effect of integrating bias mitigation methods with certified robustness methods Training model with both IBP and bias mitigation methods improves fairness metrics over fairness-only baselines in both datasets. In Bias in Bios, the model that comes closest to fulfilling equality of opportunity is the one trained with both IBP and $HardDebias$, which achieves a TPED of 0.101. In Jigsaw Toxicity, we see a similar trend, with improvements in EOdds after adding IBP training to instance weighting, $HardDebias$, and adversarial training. The model trained with both IBP and instance weighting achieves a EOdds score of 0.165, which is the lowest among all approaches

We also note that for Jigsaw Toxicity, instance weighting mitigates bias more effectively than $HardDebias$ and adversarial training (both in isolation and in combination with robustness methods). This is not the case for Bias in Bios, where $HardDebias$ and adversarial training is more effective than instance weighting in mitigating bias. This may be due to the fact that instance weighting mitigates bias explicitly for a wider array of sexual orientations and gender demographics than the other two methods. The original $HardDebias$ method only projects away the gender direction from embeddings. For adversarial debiasing, we train the adversary with the subset of the training set that is annotated for the presence of protected attribute groups, which is highly skewed towards male and female. Thus, $HardDebias$ and adversarial training may mitigate bias for binary gender, but fall short in mitigating bias for non-binary gender and sexual orientations. Conversely, instance weighting, which mitigates bias for a wider array of demographics, does not mitigate bias on gender in Bias in Bios as well as the other methods.

Additional Observations Outside of the effects of robustness on fairness, we observe differing effects of the methods on certified robustness and raw accuracy. As expected, IBP and SAFER improves performance on certified robustness on both datasets. However, we also observe degradataions in raw task accuracy in experiments with robustness methods. Combining robustness with bias mitigation methods results in a degradataion of raw task performance over fairness-only baselines. Additionally, fairness-only training results in differing effects on certified robustness accuracy. Adversarial debiasing improves certified accuracy in both datasets, while *HardDebias* embedding-initiated training results in an increase in certified robustness in Jigsaw Toxicity, but a decrease in Bias in Bios. This difference in findings may be due to the shorter length of examples in Jigsaw Toxicity, which has a median length of 34, compared with the median length of 72 in Bias in Bios, which in turn determines the number of possible perturbations used to calculate certified accuracy and the difficulty in achieving high CRA.

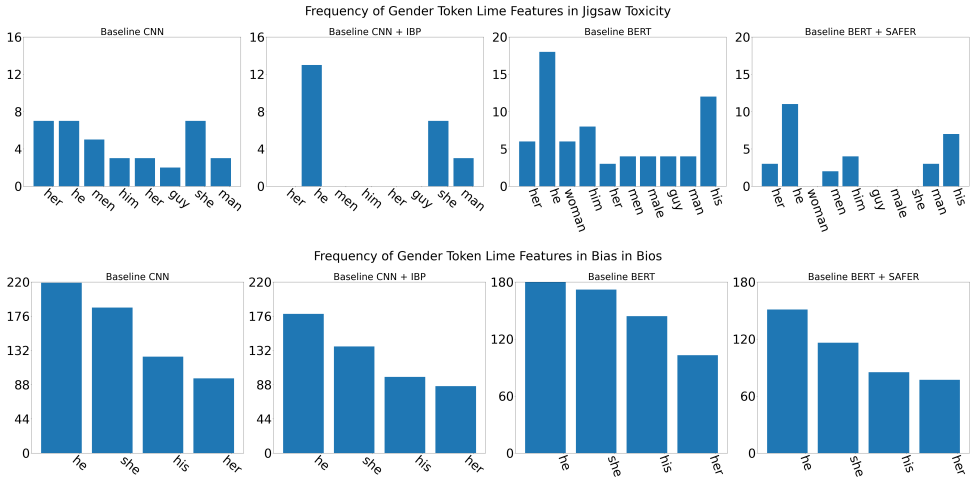

Figure 1: Frequency of gender token features as extracted by LIME for baseline and IBP trained models for Jigsaw Toxicity and Bias in Bios. We see a decrease in number and frequency of gender tokens in the list of top-5 (for Jigsaw Toxicity) and top-50 (for Bias in Bios) most important features.

5 Analysis

In this section, we study how robustness training affects the features our models use for classification. We posit that robustness training encourages models to focus more on predictive attributes than on protected attributes. To gain insight into this, we use LIME (Ribeiro et al., 2016) on the baseline, IBP, and SAFER trained models and extract token features importance as assigned by the model. We run LIME on the subset E of examples that are misclassified by our baseline model as toxic that are correctly classified by the IBP model. We take the top k features for each of the examples (where $k = 5$ for Jigsaw and $k = 50$ for Bias in Bios) over E, and then count the number of gender tokens that appear in that list. For Jigsaw Toxicity, the number of examples that we run LIME on is 488 for CNN experiments and 182 for BERT experiments. For Bias in Bios, we run LIME over a random subset of 500 examples from E for both CNN and BERT experiments.

For Jigsaw Toxicity, we see from Figure 1 that LIME extracts less gender tokens in the top-5 features of the IBP-trained and SAFER-trained model compared to the baseline model. Notably, there are 37 gender tokens that appear in the CNN model, while only 23 in the IBP-trained model. Similarly, 69 gender tokens appear in the baseline BERT model while only 37 appear in the SAFER-trained one. For Bias in Bios, we see a similar trend from Figure 1. The number of important gender token features decreases from 626 to 500 after IBP training, and from 626 to 429 after SAFER.

In addition, we compute the gradient with the output with respect to the input on several examples from Jigsaw Toxicity, which is shown in Table 5. We observe that the baseline model focuses on tokens related to protected groups, while the IBP model takes into account all parts of the sentence.

6 Related Work

Much work has been done in studying fairness in various NLP models (Mehrabi et al., 2019; Sun et al., 2019; Blodgett et al., 2020). In toxicity classification, Adragna et al. (2020) and (Zhang et al., 2020) study the fairness in predicting toxic internet contents in which the contents contain demographic identity-terms (e.g., “gay”, “black”). In occupation classification, De-Arteaga et al. (2019) and Romanov et al. (2019) study the impact of including explicit gender indicators such as a person’s names or a pronoun in online biographies.

Some notable bias mitigation methods, which we also use in this paper, include instance weighting (Zhang et al., 2020), embedding debiasing (Bolukbasi et al., 2016a; Wang et al., 2020), and adversarial debiasing (Zhang et al., 2018). In particular, Bolukbasi et al. (2016a) proposed to reduce representational harm existent in word embeddings. Zhang et al. (2020) proposed instance weighting, a method to debias text classification models for bias against examples containing demographic identity-terms by weighting the instances in the loss function, and that is optimized for demographic pairty. Zhang et al. (2018) presents an adversarial training approach to achieve various notions of fairness

Model	Saliency Map
	Example 1
Baseline	i think that we need to find a real solution to the real problem of abortion abortion represents a loss of hope by a woman that the child she is carrying will have any kind of future i think that the continuing effort to prohibit or limit access to abortion providers does solve the problem any parent can tell you that if you make something forbidden your kids are more likely to seek that thing out i can vouch for this fact a solution that appears to work better as measured by the abortion rate is to give that woman hope by providing free kindergarten the earned income tax credit and showing that woman that by our country investing in innovation there will be a job out there for her kid not to make this a political argument but it is under democratic presidents obama that the abortion rate fell to its lowest level ever our current president and our government is failing at creating hope and that is sad
IBP	i think that we need to find a real solution to the real problem of abortion abortion represents a loss of hope by a woman that the child she is carrying will have any kind of future i think that the continuing effort to prohibit or limit access to abortion providers does solve the problem any parent can tell you that if you make something forbidden your kids are more likely to seek that thing out i can vouch for this fact a solution that appears to work better as measured by the abortion rate is to give that woman hope by providing free kindergarten the earned income tax credit and showing that woman that by our country investing in innovation there will be a job out there for her kid not to make this a political argument but it is under democratic presidents obama that the abortion rate fell to its lowest level ever our current president and our government is failing at creating hope and that is sad
	Example 2
Baseline	i like millions of other americans am not grieving i was overjoyed with the electoral results glad that a woman who advocates for and supports abortion euthanasia gay marriage and the restriction of religious freedom has been defeated hopefully for all time i give thanks that this nation this society was preserved from a second clinton presidency and that we now have the opportunity to reverse years of extreme leftist policies on november it seemed that a new dawn shown on america christmas came a little early this year
IBP	i like millions of other americans am not grieving i was overjoyed with the electoral results glad that a woman who advocates for and supports abortion euthanasia gay marriage and the restriction of religious freedom has been defeated hopefully for all time i give thanks that this nation this society was preserved from a second clinton presidency and that we now have the opportunity to reverse years of extreme leftist policies on november it seemed that a new dawn shown on america christmas came a little early this year

Table 5: Gradient saliency examples on Jigsaw Toxicity. Highlights show larger value of the output gradient with respect to the token embedding. The baseline CNN model focuses on some tokens related to protected groups (e.g., woman), while IBP encourages the model to take into account other parts of the sentence, resulting in less bias.

that is achieved by training an adversary to identify information on protected groups and training the model to minimize the adversary loss. These approaches are designed for reducing specific types of bias exhibited in data.

On the robustness front, it has been shown that models are susceptible to adversarial word substitution attacks (Ebrahimi et al., 2018; Jia and Liang, 2017). Parallel to the development of methods developed to reduce word substitution robustness in the NLP domain (e.g., (Miyato et al., 2017; Huang et al., 2019; Zhou et al., 2021)), many studies has been done in the computer vision domain to ensure that models are robust to image noising (Kannan et al., 2018; Szegedy et al., 2014).

In the intersection of area between fairness and robustness of model training, there is limited prior work in the NLP area. Nanda et al. (2020) investigate and define *robustness bias*, a notion of fairness in which a model must be impervious to perturbations to the same degree for all subgroups, and investigate robustness bias in the computer vision domain. Adragna et al. (2020) examine the use of invariant risk minimization in improving the fairness on out-of-distribution data for toxicity classification. Their robustness approach is inspired from domain generalization and it allows to learn models that have invariant performance across different label distributions. This differs from the word substitution notions of robustness that our methods are optimized for. Chang et al. (2020) shows that achieving equalized odds is incongruent with adversarial robustness on the COMPAS (J. Larson and Angwin, 2017) and the Adult dataset (Dua and Graf, 2017), which is outside the NLP domain. The closest work to ours is in counterfactual logit pairing (Garg et al., 2019), which encourages a model to be robust to protected attributes for counterfactual fairness. However, logit pairing does have the certified characteristic of the robustness methods we use in this study.

7 Conclusion

We present a study that investigates the effect of optimizing for word substitution robustness on fairness. We find that, in both CNN and BERT models, adding robustness methods such as IBP and SAFER to the training process improves fairness metrics over adding bias mitigation methods alone. Given these promising results, we encourage future explorations in using robustness methods to not only improve fairness metrics, but to also optimize for both fairness and robustness, two important aspects of creating trustworthy NLP.

Future work may include studying the effects of robustness and fairness in attributes other than gender and sexual orientation, extending our study to other word substitution based robustness methods, and exploring more sophisticated methods to combine robustness and bias mitigation methods during training. We also intend on extending the study to investigating the impact of privacy pre-

serving training methods on both, robustness and fairness.

Broader Impact

We limit the scope of this paper to gender and sexual orientation in this initial effort, and future work must be done on mitigating bias in other protected attribute dimensions such as race, ethnicity, neurodiversity, etc. Additionally, this work draws importance to the need to extend fairness methods to groups beyond binary gender. In our IBP_{gender} experiments, we only consider swapping binary gender pairs from prior literature to provide an anchor for our analysis. We see from our results that methods that mitigate for binary gender such as *HardDebias* and IBP_{gender} do not reduce harm for all gender or sexual orientation, especially for non-binary gender and non-heterosexual sexual orientation groups. We will extend the study in the future by developing fairness methods that directly mitigate for non-heterosexual sexual orientations and non-binary genders pairs using sociology literature.

The language used in this paper is English. We recognize that the presented methods rely on the availability and quality of the set of words associated to a fairness task. Scaling to languages beyond English–such as gendered languages like Spanish–need more careful analysis. Another limitation of this method is that word substitution may lead to non-sensible sentences and inappropriate grammar especially in complex fairness domains where it is difficult to find word-to-word mapping (e.g., mapping names of religious artifacts like Christmas tree or Diwali lights, etc are not trivial).

Our experiments and results show that pursuing fairness can help in improving robustness and vice versa. With these findings, we hope to inspire researchers to investigate novel approaches that focus on jointly achieving robust and fair models. We also hope that this work will lead to more investigations around achieving multiple objectives such as privacy, robustness and fairness together in the NLP research community.

References

Robert Adragna, Elliot Creager, David Madras, and R. Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *ArXiv*, abs/2011.06485.

Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. 2020. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3529–3530.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, M. Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *EMNLP*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and A. Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *ArXiv*, abs/1607.06520.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Hongyan Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and Reza Shokri. 2020. On adversarial bias and the robustness of fair machine learning. *ArXiv*, abs/2006.08669.

Maria De-Arteaga, Alexey Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, Sahin Cem Geyik, K. Kenthapadi, and A. Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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.

Dheeru Dua and Casey Graf. 2017. Ouci machine learning repository. http://archive.ics.uci.edu/ml.

J. Ebrahimi, Anyi Rao, Daniel Lowd, and D. Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *ACL*.

Sahaj Garg, V. Perot, Nicole Limtiaco, Ankur Taly, Ed Huai hsin Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Society*, 7.

M. Hardt, E. Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *NIPS*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. 2019. Achieving verified robustness to symbol substitutions via interval bound propagation. 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 4083–4093.

B. Hutchinson, Vinodkumar Prabhakaran, Emily L. Denton, K. Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. *ArXiv*, abs/2005.00813.

L. Kirchner J. Larson, S. Mattu and J. Angwin. 2017. Compas dataset. https://github.com/propublica/compas-analysis.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *ArXiv*, abs/1707.07328.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. *ArXiv*, abs/1909.00986.

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. Mitigating gender bias amplification in distribution by posterior regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942.

Harini Kannan, A. Kurakin, and Ian J. Goodfellow. 2018. Adversarial logit pairing. *ArXiv*, abs/1803.06373.

J. Kemp, Alvin Rajkomar, and Andrew M. Dai. 2019. Improved hierarchical patient classification with language model pretraining over clinical notes. *arXiv: Learning*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

P. P. Liang, Irene Z Li, E. Zheng, Yao Chong Lim, R. Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *ACL*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and A. Galstyan. 2019. A survey on bias and fairness in machine learning. *ArXiv*, abs/1908.09635.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Proceedings of the International Conference on Learning Representations*.

Vedant Nanda, S. Dooley, Sahil Singla, S. Feizi, and John P. Dickerson. 2020. Fairness through robustness: Investigating robustness disparity in deep learning. *ArXiv*, abs/2006.12621.

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. *Sexuality & Culture*, pages 1–33.

J. Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Alexey Romanov, Maria De-Arteaga, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, Sahin Cem Geyik, K. Kenthapadi, Anna Rumshisky, and A. Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. *ArXiv*, abs/1904.05233.

Chengwei Su, Rahul Gupta, Shankar Ananthakrishnan, and Spyridon Matsoukas. 2018. A re-ranker scheme for integrating large scale nlu models. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 670–676.

T. Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, M. ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding-Royer, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *ArXiv*, abs/1906.08976.

Christian Szegedy, W. Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and R. Fergus. 2014. Intriguing properties of neural networks. *CoRR*, abs/1312.6199.

Y. Tan and L. Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *NeurIPS*.

Tianlu Wang, Xi Victoria Lin, Nazneen Rajani, Vicente Ordonez, and Caimng Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. *ArXiv*, abs/2005.00965.

M. Ye, Chengyue Gong, and Qiang Liu. 2020. Safer: A structure-free approach for certified robustness to adversarial word substitutions. *ArXiv*, abs/2005.14424.

B. H. Zhang, B. Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *ACL*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Jieyu Zhao and Kai-Wei Chang. 2020. LOGAN: Local group bias detection by clustering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021. Defense against synonym substitution-based adversarial attacks via dirichlet neighborhood ensemble. In *ACL*.

A Appendices

Appendix A. Hyperparameter Settings We perform hyperparameter search on the dev set using random search with 12 trials, with initial learning rate range between $1 * 10^{-2}$ to $1 * 10^{-7}$, a dropout probability range of 0.1 to 0.5, and number of epochs between 10 and 60.The final hyperparameter settings are shown in Table 6. We choose our hyperparameters based on the one that minimizes FPED + TPED + (1 - CRA) + (1 - *tp*), where *tp* refers to task performance.

Additionally, for adversarial debiasing, we tune the adversary loss weight from $\alpha = 0.1$ to $\alpha = 3$, and choose $\alpha = 1$ for the weight. We pretrain our classifier and adversary for 2 epochs each.

Experiment	**Learning Rate**	**Dropout Prob**	**Number of epochs**
GloVe + CNN (Jigsaw)	1e-2	0.5	20
GloVe + CNN (Bias in Bios)	1e-3	0.1	15
BERT + SAFER (Jigsaw)	5e-6	0.1	20
BERT + SAFER (Bias in Bios)	1e-5	0.1	15

Table 6: Hyperparameter settings for our experiments. We use the same hyperparameters across our fairness and robustness experiments.