Joint Mitigation of Interactional Bias

Anonymous ACL submission

Abstract

Machine learning algorithms have been found discriminative against groups of different social identities, e.g., gender and race. With the detrimental effects of these algorithmic biases, researchers proposed promising approaches for bias mitigation, typically designed for *individual* bias type. Due to the complex nature of social bias, we argue it is important to study *how different biases interact with each other*, i.e., *how mitigating one bias type (e.g., gender) influences the bias results regarding other social identities (e.g., race and religion).* We further question whether jointly debiasing multiple types of bias is desired in different contexts, e.g., when correlations between biases are different. To address these research questions, we examine bias mitigation in two NLP tasks – toxicity detection and word embeddings – on three social identities, i.e., race, gender, and religion. Empirical findings based on benchmark datasets suggest that different biases can be correlated and therefore, warranting attention for future research on joint bias mitigation.

1 Introduction

The increasing reliance on automated systems, e.g., systems helping decide who is hired, has led many people, especially the minorities, to be unfairly treated. Take the two well-studied tasks in NLP as examples: Toxicity classifiers are found to use demographic terms such as "black" as the key features (Zhang et al., 2020; Zhou et al., 2021), and word embeddings trained on human-generated corpus such as Google News also present occupational stereotypes (Bolukbasi et al., 2016).

These findings are well-received, as witnessed by tremendous attentions from industry, academia, and many other quarters to various social biases in machine learning (Cheng et al., 2021b). Mitigating social bias is challenging due to its variety and complexity. As shown in the example of toxicity detection in Fig. 1, multiple terms are labeled as being related to two social identities: gender and race. A biased toxicity classifier may incorrectly identify it as "Toxic" simply because of these identity terms, indicating gender and/or racial bias. With various social identities, a model debiased for gender may further reduce or amplify the racial bias. When treating different biases independently – as studied by existing works in bias and fairness, e.g., (Dixon et al., 2018; Bolukbasi et al., 2016), we may overlook the implicit interactions between biases, rendering amplified total bias.

Figure 1: A sample with labeled social identities from the benchmark dataset for toxicity detection, Jigsaw[1].

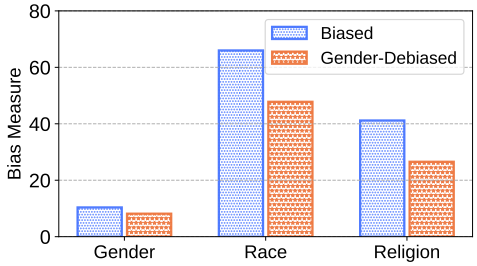


Figure 2: An illustration of interactional bias in debiasing toxicity detection using Jigsaw dataset. We observe that a gender-debiased toxicity classifier also reduces racial and religious biases.

Aware of the potential correlations between different biases, we study a novel and practical problem of *interactional bias* and ask *how to debias different biases that correlate with each other*. In a preliminary experiment, we found that a gender-debiased toxicity classifier (Gencoglu, 2020) also reduces racial and religious biases, as shown in Fig.

[1]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

2. Similar problems can surface when debiasing at the data level. For example, will approaches to debiasing for gender amplify or mitigate racial and religious biases in word embeddings? With multiple **correlated** biases, we further question whether joint debiasing can outperform conventional approaches that mitigate individual biases.

To address the interactional bias problem, we examine biases against multiple social identities from both *data* and *algorithm* aspects, respectively, on two representative NLP tasks – word embeddings and toxicity detection. Particularly, we seek to answer the following research questions:

- **RQ. 1.** Are biases of different social identities (e.g., gender, race) correlated? Are the correlations different across various tasks?
- **RQ. 2.** Will a joint bias mitigation strategy outperform approaches for individual biases regarding reducing the total bias of all social identities?
- **RQ. 3.** When jointly mitigating multiple biases, do we have to trade off accuracy for debiasing?

2 Related Work

2.1 Debiasing Toxicity Detection

The majority of existing works on debiasing toxicity detection addresses this problem by data augmentation (Park et al., 2018; Sap et al., 2019) and modified model training (Zhou et al., 2021; Xia et al., 2020). The pioneering work by Dixon et al. (2018) first proposed a data augmentation strategy to eliminate demographic biases. Specifically, it balanced the training dataset by adding external labeled data. Other data augmentation methods include gender swapping (Park et al., 2018), instance weighting (Mozafari et al., 2020), and using debiased word embedding (Prost et al., 2019). By attributing discrimination to selection bias, Zhang et al. (2020) proposed to mitigate biases in the training data by assuming a non-discrimination data distribution, and reconstructing the distribution by instance weighting.

Another line of research debiases during model training. Vaidya et al. (2020) trained a debiased toxicity classifier using adversarial learning. They designed a multi-task learning model to jointly predict the toxicity of a comment and the involved identities. Gencoglu (2020) modified the model training by imposing debiasing constraints on the classification objective. A recent survey summarizes biases of commonly studied social identities in toxicity detection (Zhou et al., 2021). By considering a sequence of comments in a social media post, Cheng et al. (2021a) proposed a reinforcement learning framework to mitigate biases in a sequential manner.

Though showing promising results, previously mentioned research works in a single-bias context. They have not discussed bias interaction and joint debiasing with multiple biases.

2.2 Debiasing Word Embeddings

Debiasing word embeddings has been focused on gender, e.g., (Zhao et al., 2019; Basta et al., 2019; Prost et al., 2019). For example, the occupational stereotypes were found in word2vec (Mikolov et al., 2013) trained on the Google News dataset (Bolukbasi et al., 2016). In an extension to the pioneering work (Bolukbasi et al., 2016), Manzini et al. (2019) further identified racial and religious biases in word2vec trained on a Reddit corpus. Contextualized word embeddings such as ELMo (Peters et al., 2018) also inherit gender bias (Zhao et al., 2019; Kurita et al., 2019; Basta et al., 2019). Informed by the Implicit Association Test (IAT) (Greenwald et al., 1998) in social psychology, Caliskan et al. (2017) proposed the Word Embedding Association Test (WEAT) to examine the associations in word embeddings between concepts captured in IAT. WEAT aims to assess implicit stereotypes such as the association between female/male names and groups of words stereotypically assigned to females/males, e.g., arts vs. science.

To mitigate gender bias, the post-processing method proposed in (Bolukbasi et al., 2016) simply removed the component along the gender direction. For contextualized word embeddings, Zhao et al. (2018) proposed to explicitly restrict gender information in certain dimensions when training a coreference system. Due to the demanding computational sources of this approach, an encoder-decoder model was proposed by Kaneko and Bollegala (2019) to re-embed existing pre-trained word embeddings. While most of these works were found to remove biases superficially (Gonen and Goldberg, 2019), they have successfully raised the awareness of bias issues in the NLP community. As our work studies bias interactions, we use debiasing word embeddings as an illustration given its popularity and significance.

In summary, bias interactions and joint bias mitigation with different types of bias are rarely understood in literature of both debiasing toxicity detection and word embeddings. Therefore, we complement prior research by providing the first systematic evidence on the interactional bias and proposing joint mitigation strategies. By examining the correlations between biases within different tasks, our research emphasizes the necessity to develop joint bias mitigation strategies in the presence of multiple social biases.

3 Preliminaries

3.1 Debiasing Toxicity Detection

A common bias mitigation strategy in toxicity detection adopts the metrics widely used to assess discrimination in classification tasks, i.e, False Negative Equality Difference (FNED) and False Positive Equality Difference (FPED) (Dixon et al., 2018). FNED/FPED is defined as the sum of deviations of group-specific False Negative Rates (FNRs)/False Positive Rates (FPRs) from the overall FNR/FPR. Given N demographic groups (e.g., female and male in gender), denote each group as $G_{i\in\{1,...,N\}}$, FNED and FPED are calculated as:

$$\begin{aligned} FNED &= \sum_{i\in\{1,...,N\}} |FNR - FNR_{G_i}|, \\ FPED &= \sum_{i\in\{1,...,N\}} |FPR - FPR_{G_i}|. \end{aligned} \tag{1}$$

A debiased model is expected to have similar FNR and FPR for different groups belonging to the same identity, therefore, small FNED and FPED. Ideally, the sum of FNED and FPED is close to zero.

To reach this goal, a standard debiasing practice is the constrained model training (Chen et al., 2020; Zafar et al., 2017), which imposes debiasing constraints during the training process. Specifically, it aims to reach equitable performances for different demographic groups of interest. In this work, we employ the method in (Gencoglu, 2020), which simultaneously minimizes the deviation of each group-specific FNR/FPR from the overall FNR/FPR and the toxicity classification loss, i.e.,

$$\begin{aligned} &\min_{\theta} f_L(\theta) \\ s.t. \quad &\forall i, \ |FNR - FNR_{G_i}| < \tau_{FNR} \\ &|FPR - FPR_{G_i}| < \tau_{FPR}, \end{aligned} \tag{2}$$

where τ_{FNR} and τ_{FPR} are the tolerances of group deviation (corresponding to biases) from overall FNRs and FPRs, respectively. The model training process is then considered as a robust optimization problem. We use this approach as the base model since (1) it examines bias mitigation from the algorithmic aspect, therefore, complementing the task of debiasing word embeddings, which considers biases from the data aspect; and (2) It is a generalizable and data-independent approach, and can be easily extended to the joint bias mitigation scenario, which we will detail in Sec. 4.1.

3.2 Debiasing Word Embeddings

A pioneering work in debiasing word embedding (Bolukbasi et al., 2016) seeks to remove gender bias in word2vec (Mikolov et al., 2013) by identifying the gender bias subspace. Manzini et al. (2019) further extended the debiasing strategy from a binary setting to the multi-class setting to account for other biases such as racial bias. Note that it is possible to use other approaches for debiasing word embeddings, e.g., (Zhao et al., 2017), as the baseline model to study joint bias mitigation. We leave this for future research.

Identifying the bias subspace. The bias subspace is identified by the *defining sets* of words (Bolukbasi et al., 2016) and words in each set represent different ends of the bias. The defining sets for gender can be *{she, he}* and *{woman, man}*. There are two steps to define a bias subspace: (1) computing the vector differences between the word embeddings of words in each set and the mean word embedding over the set; and (2) identifying the most k significant components $\{\boldsymbol{b_1}, \boldsymbol{b_2}, ..., \boldsymbol{b_k}\}$ of the resulting vectors using dimensionality reduction techniques such as Principal Component Analysis (PCA) (Abdi and Williams, 2010).

Removing bias components. The next step completely or partially removes the subspace components from the embeddings, e.g., *hard-debiasing* (Bolukbasi et al., 2016). For non-gendered words such as *doctor* and *nurse*, hard-debiasing method removes their bias components; for gendered words such as *man* and *woman*, it first centers their word embeddings and then equalizes the bias components. Formally, given a bias subspace $\mathcal{B}$ defined by a set of vectors $\{\boldsymbol{b_1}, \boldsymbol{b_2}, ..., \boldsymbol{b_k}\}$, we get the bias component of an embedding in this subspace by

$$\mathbf{w}_{\mathcal{B}} = \sum_{i=1}^{k} \langle \mathbf{w}, \boldsymbol{b_i} \rangle \boldsymbol{b_i}. \tag{3}$$

We then *neutralize* word embeddings by removing the resulting component from non-gendered words: $\mathbf{w}' = \frac{\mathbf{w} - \mathbf{w}_{\mathcal{B}}}{\|\mathbf{w} - \mathbf{w}_{\mathcal{B}}\|}$, where $\mathbf{w}'$ are the debiased word

embeddings. We further *equalize* the gendered words in the equality set E. Specifically, we debias $\mathbf{w} \in E$ by

$$\mathbf{w}' = (\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathcal{B}}) + \sqrt{1 - \|\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathcal{B}}\|^2} \frac{\mathbf{w}_{\mathcal{B}} - \boldsymbol{\mu}_{\mathcal{B}}}{\|\mathbf{w}_{\mathcal{B}} - \boldsymbol{\mu}_{\mathcal{B}}\|}, \quad (4)$$

where $\boldsymbol{\mu} = \frac{1}{|E|} \sum_{\mathbf{w} \in E} \mathbf{w}$ is the average embedding of the words in the set. $\boldsymbol{\mu}_{\mathcal{B}}$ denotes the bias component in the identified gender subspace and it can be obtained via Eq. 3. In real-world applications, the equality set is often the same set of words as the defining set (Manzini et al., 2019).

4 Joint Bias Mitigation

When a model aims to mitigate one type of bias, how it influences other biases is unknown in part due to the *implicit correlations* between different biases. With interactional biases that are correlated with each other, a joint bias mitigation strategy might further help alleviate the total bias regarding all social identities. Here, we introduce our strategies for jointly debiasing toxicity detection and word embeddings with multiple biases.

4.1 Joint Debiasing for Toxicity Detection

The joint bias mitigation strategies for toxicity detection is an extension of the method described in Sec. 3.1. Let T=$\{gender, racial, religion\}$ be the social identity set and G_t=$\{G_{t1}, G_{t2}, ..., G_{tj}, ...\}$ be the demographic group set of $t \in T$ (e.g., G_{gender}=$\{female, male\}$). An intuitive solution is to extend the debiasing method in Sec. 3.1 by imposing uniform constraints on each demographic group G_{tj}. Take identities *gender* and *race* as an example, we enforce the model to have similar FPR and FNR for *male* and *black*. However, simply enforcing uniform performances across all identities may lead to sub-optimal solutions, considering the unique bias distribution and language characteristics within each social identity. For instance, it is more common to observe gender-related insulting words in text biased against gender than that biased against religion. Therefore, we propose to pair each demographic group only with groups from the same identity in the joint debiasing setting. That is, we enforce the model to have similar FPR and FNR for *male* and *female*, but not for *male* and *black*. The *joint debiasing constraint* is defined as:

$$\begin{aligned} &G_{tj} \in G_t \quad \forall t \in T \\ &|FNR_{G_t} - FNR_{G_{tj}}| < \tau_{FNR}, \\ &|FPR_{G_t} - FPR_{G_{tj}}| < \tau_{FPR}. \end{aligned} \quad (5)$$

Accordingly, the *joint bias metric* is defined as the sum of $FNED_J$ and $FPED_J$ across all identities, where $FNED_J$ and $FPED_J$ are defined as:

$$\begin{aligned} FNED_J &= \sum_{t \in T} \sum_{G_{tj} \in G_t} |FNR_{G_t} - FNR_{G_{tj}}|, \\ FPED_J &= \sum_{t \in T} \sum_{G_{tj} \in G_t} |FPR_{G_t} - FPR_{G_{tj}}|. \end{aligned} \quad (6)$$

4.2 Joint Debiasing for Word Embeddings

Given an identity $t \in T$, n defining sets of word embeddings $\{D_{t1}, D_{t2}, ..., D_{tn}\}$, and word embedding $\mathbf{w} \in \mathbb{R}^d$ of word w, the bias subspace $\mathcal{B}_t$ is defined by the first k components of the following PCA evaluation (Manzini et al., 2019):

$$\mathcal{B}_t = \mathbf{PCA}\Big(\bigcup_{i=1}^{n} \bigcup_{\mathbf{w} \in D_{ti}} \mathbf{w} - \boldsymbol{\mu}_{ti}\Big), \quad (7)$$

where $\boldsymbol{\mu}_{ti} = \frac{1}{|D_{ti}|} \sum_{\mathbf{w} \in D_{ti}} \mathbf{w}$ is a vector averaged over all word embeddings in set i. $\bigcup$ denotes concatenation by rows.

Joint bias mitigation requires to simultaneously remove the bias components from a word embedding in all three identity subspaces. One simple solution is to take the mean of all bias subspaces as the joint bias subspace. However, the unique information of each bias type may be averaged out. Alternatively, we can concatenate the bias subspace $\mathcal{B}_t$ w.r.t. individual identities such that the resulting subspace $\mathcal{B}$ contains the significant components of all identity biases:

$$\mathcal{B} = \bigcup_{t \in T} \mathcal{B}_t. \quad (8)$$

$\mathcal{B} \in \mathbb{R}^{3k \times d}$ allows us to reserve the unique bias information of each identity in joint bias mitigation.

Quantifying Bias Removal. We use the mean average cosine similarity (MAC) (Manzini et al., 2019) to evaluate the individual bias in collections of words. Suppose we have a set of target word embeddings $\mathcal{S}$ that inherently contains certain form of social bias (e.g., *Jew, Muslim*) and a set of attribute sets $\mathcal{A} = \{A_1, A_2, ..., A_N\}$. A_j consists of embeddings of words $\boldsymbol{a}$ that should not be associated with any word in $\mathcal{S}$ (e.g., *violent, terrorist*). We define function $f(\cdot)$ to compute the mean cosine distance between $S_i \in \mathcal{S}$ and $\boldsymbol{a} \in A_j$:

$$f(S_i, A_j) = \frac{1}{|A_j|} \sum_{\boldsymbol{a} \in A_j} \cos(S_i, \boldsymbol{a}), \quad (9)$$

where $\cos(S_i, \boldsymbol{a}) = 1 - \frac{S_i \cdot \boldsymbol{a}}{\|S_i\|_2 \cdot \|\boldsymbol{a}\|_2}$. MAC is then defined as

$$\text{MAC}(\mathcal{S}, \mathcal{A}) = \frac{1}{|\mathcal{S}||\mathcal{A}|} \sum_{S_i \in \mathcal{S}} \sum_{A_j \in \mathcal{A}} f(S_i, A_j). \quad (10)$$

Table 1: Statistics of the Jigsaw dataset. We report the percentage of each demographic group and proportion of toxic sentences within each group.

Group	**Gender**		**Race**	
Type	Male	Female	Black	White
% data	11.0%	13.2%	3.7%	6.2%
% toxicity	15.0%	13.7%	31.4%	28.1%
Group	**Religion**			**Overall**
Type	Christian	Jewish	Muslim	
% data	10.0%	1.9%	5.2%	100.0%
% toxicity	9.1%	16.2%	22.8%	11.4%

Table 2: Toxicity detection and bias mitigation performances of biased, individually debiased, and jointly debiased methods on the Jigsaw dataset.

Model	**AUC**↑	**F1**↑	**Acc.**↑	**Individual Bias Metric**↓			
				Ge	**Ra**	**Re**	**Total**
Baseline	**87.78**	51.21	85.07	10.36	65.98	41.16	117.50
Gender	87.73	54.93	89.83	8.15	**47.79**	26.54	**82.48**
Race	87.15	54.93	89.50	**7.78**	57.34	34.95	100.06
Religion	87.71	55.42	89.36	11.26	66.53	**26.41**	104.20
Model	**AUC**↑	**F1**↑	**Acc.**↑	**Joint Bias Metric**↓			
				Ge	**Ra**	**Re**	**Total**
Ge+Ra	87.38	55.30	89.06	6.98	8.28	22.13	37.40
Ge+Re	87.61	**55.43**	89.74	6.13	5.97	22.88	34.98
Ra+Re	86.86	54.52	**90.10**	**4.44**	5.98	22.37	32.79
Joint	87.62	54.74	90.01	5.65	**4.72**	**20.29**	**30.65**

A larger MAC score denotes greater bias removal.

5 Experiments

In this section, we present the major results of this work. In particular, we answer **RQ. 1 - RQ. 3** that seek to examine (1) the interactions between different types of bias; (2) the effectiveness of the proposed joint bias mitigation strategies within different contexts; and (3) the debiasing-accuracy tradeoff of the joint approaches in different tasks. Data and code can be found in supplementary materials.

5.1 Task 1: Toxicity Detection

We first describe the data source and basic experimental settings. We then discuss the main results.

5.1.1 Data

We use the Perspective API's Jigsaw dataset of 403,957 sentences with toxicity and identity annotations. We use the same data split strategy as (Gencoglu, 2020), that is 70% for training, 15% for validation, and 15% for testing. The detailed statistics for each group are shown in Table 1. All data in this study are publicly available and used under ethical considerations.

5.1.2 Compared Approaches

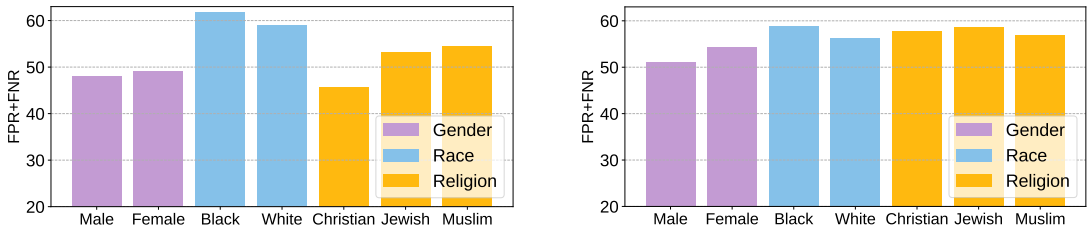
We consider models debiased at different levels, i.e., biased **Baseline** model, models debiased for individual social identities (i.e., **Ge**nder, **Ra**ce and **Re**ligion), models debiased simultaneously for two identities (i.e., **Ge+Ra**, **Ge+Re**, and **Ra+Re**) as well as models debiased simultaneously for all three identities (i.e., **Joint**). Parameter settings can be found in Appendix A.

5.1.3 Evaluation Metrics

We use the standard AUC, micro-F1, and accuracy (Acc.) as the evaluation metrics for classification. For bias mitigation, following (Dixon et al., 2018; Gencoglu, 2020), we use the standard *individual bias metric* introduced in Section 3.1 for individually debiased models (i.e., **Ge**nder, **Ra**ce, and **Re**ligion). As the individual bias metric is not suitable when multiple types of bias are present, we measure bias using the *joint bias metric* described in Section 4.1 for methods debiasing for multiple biases: Sequential debiasing (i.e., **Ge+Ra**, **Ge+Re**, and **Ra+Re**) and **Joint**.

5.1.4 Results

We have the following observations from Table 2:
RQ. 1. Different types of bias are inherently ***correlated*** and the correlations tend to be ***positive***: debiasing for individual social identities will alleviate total bias w.r.t. all identities, e.g., comparing results of *Baseline* with **Ge**nder, **Ra**ce, and **Re**ligion. While aimed for gender bias, **Ge**nder even achieves better performance on mitigating racial bias compared to **Ra**ce. Similar findings can be observed from models debiased simultaneously for two identities, e.g., results for **Ra+Re**. This indicates that with positive bias interactions in toxicity detection, an individually debiased model might be helpful to mitigate multiple types of bias.
RQ. 2. Our proposed joint debiasing strategy outperforms individual debiasing methods in terms of total bias mitigation. As observed from the last two rows in Table 2, the **Joint** model effectively reduces multiple types of bias simultaneously, achieving the least bias on race, religion, as well as total bias. Meanwhile, it presents competitive AUC, F1, and Accuracy scores. Instead of forcing equal model performances among all demographic groups, **Joint** only seeks for the closeness of FNR and FPR between groups within the same social identity, therefore better capturing the unique debiasing characteristics for individual iden-



(a) Baseline (b) Joint

Figure 3: FPR+FNR w.r.t. different groups within each identity. We compare the biased **baseline** model with the **Joint** debiased model. The difference of FPR+FNR between groups within the same identity reflects model bias, i.e., larger difference indicates higher level of bias. We observe that **Joint** presents smaller group differences and results in lower bias. Best viewed in color.

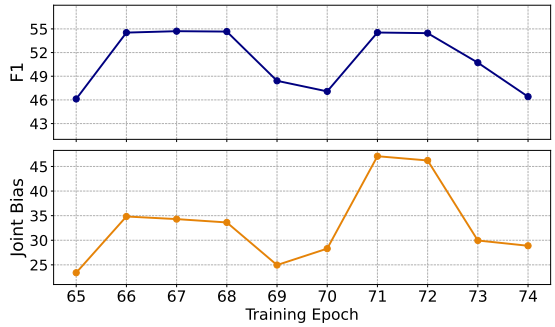


Figure 4: The debiasing-accuracy trade-off of **Joint** after its loss converges.

ties. We zoom in and further examine the sum of FPR and FNR (FPR+FNR) of each demographic group. The results for biased model and **Joint** are presented in Fig. 3. Our observation is that the FPR+FNR of **Joint** on different demographic groups within the same social identity are closer to each other, e.g., the three groups in Religion in Fig. 3(b). As bias is defined as the *difference* between FPR+FNR of different demographic groups, this result further validates the effectiveness of the proposed **Joint** on joint bias mitigation.

RQ. 3. Biases inherently baked in data will propagate to toxicity detection models, and it is challenging to completely remove such data bias while ensuring good classification performance. The debiasing-accuracy trade-off of **Joint** in Fig. 4 shows that the joint bias increases as F1 score increases, i.e., a better jointly debiased toxicity classifier can present poorer classification performance. This suggests that bias mitigation by debiasing the model alone may not be ideal, especially with multiple types of bias. As shown in Table 1, data distributions of different social identities are imbalanced, which can lead to optimization challenge and further contribute to the debiasing-accuracy trade-off.

5.2 Task 2: Word Embeddings

We first introduce the social bias, linguist data sources, and tasks used for evaluation. We then present and discuss the major results.

5.2.1 Social Bias and Linguistic Data

We use the L2-reddit corpus (Rabinovich et al., 2018), a collection of Reddit posts and comments by both native and non-native English speakers. For gender, we use vocabularies curated by (Bolukbasi et al., 2016) and (Caliskan et al., 2017). For race and religion, we use the list of lexicons curated by (Manzini et al., 2019). The initial biased word embeddings are obtained by training word2vec on approximately 56 million sentences. We debias word embeddings w.r.t. individual bias using the method in (Manzini et al., 2019). All in all, we have four sets of pretrained word embeddings: *Initial-biased,* ***Ge****nder-debiased,* ***Ra****ce-debiased*, and ***Re****ligion-debiased* word embeddings, denoted as **Initial, Ge, Ra**, and **Re**.

5.2.2 Evaluation Tasks

We perform the following experiments to answer **RQ. 1-3**: (1) we first quantify the biases in both the biased and debiased pretrained word embeddings w.r.t. the other two social identities, e.g., MACs (bias removal) of *Initial* and *Ge* w.r.t. race and religion. (2) We evaluate the effectiveness of bias removals of all methods. With multiple types of bias, baselines apply the hard-debiasing method to *Initial* **sequentially** whereas our approach (*Joint*) **jointly** debiases for all identities. For instance, to mitigate both gender and racial biases, the baseline debiases *Ge* for race whereas *Joint* simultaneously debiases *Initial* for race and gender.

We report MACs for individual social identities to better show bias interactions. We also perform a

Table 3: MACs (↑) w.r.t. different social identities **After** debiasing for various word embeddings. **Target Identity** is the bias type we aim to mitigate and of which we present the MACs. Take the target identity "**Ge**nder" as an example. **Initial'** shows gender MACs **After** debiasing for gender on **Initial**. Similarly, **Re'** and **Joint** under **Re**ligion are gender MACs after *sequentially* and *jointly* debiasing for religion and gender, respectively. Under **Re**ligion + **Ra**ce are gender MACs after debiasing base embeddings (i.e., **Re-Ra** and **Initial**) *sequentially* (i.e., the resulting embedding is **Re-Ra'**) and *jointly* (i.e., **Joint**) for all three identities. All the resulting embeddings after debiasing are denoted as **XX'** (same as Table 3). * indicates statistically insignificant results.

Target Identity	**Initial'**	**Debiasing Multiple Identities**			
		Religion		**Re**ligion + **Ra**ce	
Gender	.695	**Re'**	**Joint**	**Re-Ra'**	**Joint**
		.654	.790	.655	**.794**
		Religion		**Re**ligion+**Ge**nder	
Race	.925	**Re'**	**Joint**	**Re-Ge'**	**Joint**
		.889	**.940**	.891	.880*
		Gender		**Ge**nder + **Ra**ce	
Religion	.937	**Ge'**	**Joint**	**Ge-Ra'**	**Joint**
		.865	**.941**	.865	.930

Table 4: MACs (↑) w.r.t. different social identities for various sets of word embeddings. **Before** and **After** denote embeddings before and after we apply the hard-debiasing method, respectively. Under **Before** are the base embeddings we aim to debias. **Target Identity** is the bias type we aim to mitigate and of which we present the MACs. Take the third row "**Ra**ce" as an example: Under **Before** are the race MACs of base embeddings **Initial** and **Ge** before debiasing for race. The resulting embeddings are **Initial'** and **Ge'** under **After**. That is, **Initial'** is debiased only for race and **Ge'** is debiased sequentially for gender and race. Under **Joint** is the race MAC after applying the proposed debiasing strategy to **Initial** to jointly reduce gender and racial biases.

Target Identity	**Before**		**After**		
	Initial	**Ge**	**Initial'**	**Ge'**	**Joint**
Race	.892	.894	**.925**	.892	.924
Religion	.859	.857	.937	.865	**.941**
	Initial	**Ra**	**Initial'**	**Ra'**	**Joint**
Gender	.623	.624	**.695**	.654	**.695**
Religion	.859	.857	.937	.865	**.940**
	Initial	**Re**	**Initial'**	**Re'**	**Joint**
Gender	.623	.624	.695	.654	**.790**
Race	.892	.890	.925	.889	**.940**

paired t-test on the distribution of average cosine distance used to compute MAC (Manzini et al., 2019). Unless otherwise noted, results of MAC below are statistically significant at level 0.05. (3) To examine the influence of joint bias mitigation on the utility of word embeddings, we further perform downstream tasks following (Manzini et al., 2019), including NER, POS tagging, and POS chunking. Data are provided by the CoNLL 2003 shared tasks (Sang and De Meulder, 2003). There are two evaluation paradigms: replacing the biased embeddings with the debiased ones or retraining the model on debiased embeddings.

5.2.3 Results

RQ. 1. We can observe from Table 4-3 that: (i) hard-debiasing method designed for individual identity has little influence on the results of other identities. For example, in Table 4, race MACs of *Initial* (.892) and *Ge* (.894) are similar. (ii) Sequential hard-debiasing method has ***negative*** influence on debiasing for the second and the third identities. For example, *Initial'* (.925) achieves better race MAC than *Ge'* (.892). (iii) Similar findings can be observed in Table 3. Take the target identity **Ge**nder as an example: in Row 4, both gender MACs for sequential debiasing (.654 and .655) are worse than directly debiasing gender on *Initial* (.695).

RQ. 2. Under **After** in Table 4, *Joint* outperforms sequential debiasing and achieves competitive performance compared to method that focuses on debiasing for individual identity, i.e., *Initial'*. Similarly, on row 4 in Table 3, gender MACs of jointly debiasing for religion and gender (.790) or all three identities (.794) are better than the results of corresponding sequential debiasing methods, i.e., *Re'* (.654) and *Re-Ra'* (.655). We further generate the top five analogies for {*man, woman*} using various word embeddings debiased for gender. We observe from Table 5 that *Initial'* and *Joint* generate same analogies for both *man* and *woman*. The sequential debiasing methods (i.e., *Re'* and *Re-Ra'*), however, generate discriminative analogies as highlighted.

RQ. 3. We examine the effects of mitigating multiple biases on three downstream tasks: NER Tagging, POS Tagging, and POS Chunking. We compare the utility of *Initial* with that of word embeddings debiased at different levels: individual identity (**Re**ligion), two (**Re**ligion and **Ra**ce), and all identities. We report results of embedding matrix replacement in Table 6. Results of model retraining can be found in Appendix B. We observe that for all word embeddings, the semantic utility only slightly changes. Student t test further testifies that these differences are insignificant. We may conclude that the hard-debiasing method does not have

Table 5: Top five analogies of {*man, woman*} generated by various word embeddings after being **debiased for Ge**nder. Each entry below can be interpreted as "man is to XX as woman is to XX".

Initial'	**Re**ligion		**Re**ligion + **Ra**ce	
	Re'	**Joint**	**Re-Ra'**	**Joint**
(executive, executive)	**(chairman, secretary)**	(homemaker, homemaker)	(executive, executive)	(stylist, stylist)
(homemaker, homemaker)	(executive, executive)	(stylist, stylist)	**(chairman, secretary)**	(homemaker, homemaker)
(manager, manager)	(homemaker, homemaker)	(manager, manager)	(homemaker, homemaker)	(clerk, clerk)
(clerk, clerk)	(secretary, secretary)	(programmer, programmer)	(secretary, secretary)	(executive, executive)
(secretary, secretary)	(manager, manager)	(supervisor, supervisor)	(manager, manager)	(singer, singer)

Table 6: Utility of word embeddings debiased at various levels in NER Tagging, POS Tagging, and POS Chunking. **Seq** and **Joint** denote sequential and joint debiasing, respectively. Δ denotes the change before and after debiasing.

Target Identity	**Re**ligion			**Re**ligion +**Ge**nder			**Re**ligion + **Ge**nder + **Ra**ce		
	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking
Biased F1	.9930	.9677	.9968	.9930	.9677	.9968	.9930	.9677	.9968
	NA			**Seq / Joint**	**Seq / Joint**	**Seq / Joint**	**Seq / Joint**	**Seq / Joint**	**Seq / Joint**
Δ F1	+.007	-.026	+.003	+.004 / +.004	-.011 / -.009	+.004 / +.005	+.004 / +.004	-.012 / -.013	+.004 / +.005
Δ Precision	.0	-.029	.0	.0 / .0	-.023 / -.019	.0 / .0	.0 / .0	-.026 / -.026	.0 / .0
Δ Recall	+.025	-.073	+.012	+.015 / +.015	-.020 / -.018	+.016 / +.017	+0.014 / +.016	-.021 / -.025	+.014 / +.019

significant influence on the utility of word embeddings, regardless of the number of bias types. This applies to both sequential and joint bias mitigation.

5.3 Discussions

Based on the empirical evaluations, we summarize **key findings** about interactional bias: (1) Correlations between biases can be positive or negative. As a toxic comment often includes terms indicating multiple social identities, the bias interactions in toxicity detection tend to be positive. For word embeddings, bias interaction is weak in part because it is less common to see words associated with different identities in a general text. However, there might be negative bias interactions during debiasing process as evidenced by the poor performance of sequential hard-debiasing. We conjecture this is partly due to the difference between debiasing data and debiasing models. (2) With multiple correlated biases, a joint bias mitigation approach is more effective in reducing total bias, regardless of the correlation being positive or negative. However, the improvement of this joint approach appears to be more significant under negative bias interactions. This suggests the need to simultaneously consider multiple biases, especially under the negative correlation. (3) The debiasing-accuracy trade-off appears to exist in joint bias mitigation. The issue might not be equally serious across tasks, e.g., it is more evident to see the trade-off when jointly debiasing toxicity detection.

The study is not without limitations. First, there might be other sources contributing to data biases, such as the selection bias and annotation bias due to the diverse belief and background of annotators. Second, future research on other NLP/ML tasks and datasets is needed to have an in-depth understanding of our findings. Third, while showing promising results, our joint bias mitigation strategy is straightforward. More advanced approaches might better capture bias correlations, facilitate the performance of joint bias mitigation, and address the issue of debiasing-accuracy trade-off.

6 Conclusions

This work initiates the discussions of *interactional* bias using two representative NLP tasks. It examines the correlations between biases w.r.t. different social identities and explores joint bias mitigation strategies. We present findings of how biases might interact differently dependent on the task and show promising results of simple joint mitigation approaches. The goal of this study is to bring forefront the discussions of interactional biases and joint mitigation strategy that might have been neglected by the community before.

Our work opens up several key future research avenues. Some prospective works include investigating interactional bias in other NLP and machine learning tasks, developing more principled approaches and evaluation metrics for joint bias mitigation, conducting in-depth analyses of the optimization trade-off between different biases, as well as examining the application of debiased word embeddings in downstream debiasing tasks.

Ethics Statement

This work aims to advance collaborative research efforts in joint mitigation of interactional bias in machine learning and NLP, a topic that has yet to be well understood. Here, we provide the first solution, which is simple yet effective. However, much work remains to elucidate how to build a debiased and effective framework in the presence of multiple biases that are correlated with each other. All data in this study are publicly available and used under ethical considerations. Text and figures that contain terms considered profane, vulgar, or offensive are used for illustration only, they do not represent the ethical attitude of the authors.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

You-Lin Chen, Zhaoran Wang, and Mladen Kolar. 2020. Provably training neural network classifiers under fairness constraints. *arXiv preprint arXiv:2012.15274*.

Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. 2021a. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In *Proceedings of ACL*.

Lu Cheng, Kush R Varshney, and Huan Liu. 2021b. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*.

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.

Oguzhan Gencoglu. 2020. Cyberbullying detection with fairness constraints. *IEEE Internet Computing*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, pages 609–614.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *ACL*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *NAACL*, pages 615–621.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing model bias in toxic comment detection. In *ICWSM*.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *SocialNLP*.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR.

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*, pages 4134–4145.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *NAACL*, pages 629–634.

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 *EMNLP*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*, volume 2.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection. In *European Chapter of the Association for Computational Linguistics*.

A Appendices

A.1 Parameter Settings

A.1.1 Debiasing Toxicity Detection

For debiasing toxicity detection with constraints, we use the publicly available implementation in paper (Gencoglu, 2020)[2]. In particular, we employ a multilingual language model, sentence-DistilBERT (Reimers and Gurevych, 2019), for extracting sentence embeddings to represent each post/comment. The toxicity classifier is a simple 3-layer fully-connected neural network. The size of each layer is 512, 32, and 1, respectively. We use sentence embeddings output from sentence-DistilBERT as input features. The model is trained for 25 epochs with Adam as the optimizer. In the repository, they also release their trained models along with the source code. To account for the performance variance caused by the system environment, we ran the source code from scratch and reported our experimental results. Both baselines and the constrained models are trained in a mini-batch manner. Models that maximize the F1 score on the validation set are used for experimentation. We ran each experiments for five times and reported the average performances. We introduce the details of major parameter setting in Table 7. The descriptions of the major parameters are as follows:

- Embedding Dimension: the dimension of BERT sentence embedding
- LR Constrains: the updating rate of proxy-Lagrangian state
- Adam_beta_1: the exponential decay rate for the 1st moment estimates
- Adam_beta_2: the exponential decay rate for the 2nd moment estimates
- FNR Deviation: the maximum allowed deviation for false negative rate
- FPR Deviation: the maximum allowed deviation for false positive rate

A.1.2 Debiasing Word Embeddings

For hard debiasing word embeddings, the major parameter k (the most k significant components in PCA) is set to 2 for all bias subspace identifications. We use the same data split strategy as (Manzini et al., 2019) by randomly splitting the dataset into 80% for training, 10% for validation, and 10% for testing. We also follow their parameter settings for all downstream tasks. We ran each experiments for five times and reported the average performances. The major parameters in debiasing word embeddings are:

- Max Seq Len: the threshold to control the maximum length of sentences.
- Embedding Size: the dimension of embedding layer.
- Debias_eps: the threshold for detecting words that had their biases altered.

We describe major parameter settings in Table 8.

[2]https://github.com/ogencoglu/fair_cyberbullying_detection

Parameter	Setting	Parameter	Setting
Batch Size	128	Embedding Dimension	512
Learning Rate (LR)	5e-4	LR Constrains	5e-3
Adam_beta_1	0.9	Adam_beta_2	0.999
FNR Deviation	0.02	FPR Deviation	0.03

Table 7: Details of the parameters in the debiasing toxicity detection experiment.

Parameter	Setting	Parameter	Setting
Max Seq Len	128	Embedding Size	50
Learning Rate (LR)	1e-3	Epochs	25
RMSprop Decay	1e-3	RMSprop Momentum	0.25
Debias_eps	1e-10	Batch Size	64

Table 8: Details of the parameters in the debiasing word embedding experiment.

A.2 Supplementary Experimental Results

A.2.1 Downstream Tasks of Word Embedding

In the tasks of NER tagging, POS tagging, and POS chunking, we can either replace the biased word embeddings with debiased ones or retrain the model on the debiased embeddings. In addition to the results for replacement shown in Sec. 5.2, here, we examine the semantic utility of word embeddings in the second scenario. Table 9 presents the similar results to those for embedding matrix replacement. The major difference is that after the word embeddings sequentially debiased for religion and gender (the middle part in the table), their utility in POS Tagging is slightly improved whilst utility of embeddings debiased for single bias or jointly debiased for two types of bias decreases. As the changes are statistically insignificant, this further supports the conclusion that the hard-debiasing method does not have significant influence on the

Table 9: Utility of word embeddings debiased at different levels in NER Tagging, POS Tagging, and POS Chunking with model retraining. **Seq** and **Joint** represent sequential and joint bias mitigation, respectively. Δ denotes the change before and after debiasing.

Target Identity	**Re**ligion			**Re**ligion +**Ge**nder			**Re**ligion + **Ge**nder + **Ra**ce		
	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking	NER Tagging	POS Tagging	POS Chunking
Biased F1	.9930	.9677	.9968	.9930	.9677	.9968	.9930	.9677	.9968
	NA			**Seq / Joint**	**Seq / Joint**	**Seq / Joint**	**Seq / Joint**	**Seq / Joint**	**Seq / Joint**
Δ F1	+.007	-.011	+.003	+.004 / +.004	+.003 / -.011	+.004 / +.005	+.004 / +.004	-.010 / -.007	+.004 / +.005
Δ Precision	.0	-.011	.0	.0 / .0	+.006 / -.023	.0 / .0	.0 / .0	-.018 / -.007	.0 / .0
Δ Recall	+.025	-.031	+.012	+.015 / +.015	+.005 / -.019	+.016 / +.017	+.014 / +.016	-.021 / -.023	+.014 / +.019

utility of word embeddings, regardless of the number of bias types. This applies to both sequential and joint bias mitigation.