ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230473

Adversarial Discriminator to Mitigate Gender Bias in Abusive Language Detection

Jaeil Park* **and Sung-Bae Cho****

Department of Computer Science, Yonsei University, Seoul 03722, South Korea
ORCiD ID: Jaeil Park https://orcid.org/0000-0003-3124-1651,
Sung-Bae Cho https://orcid.org/0000-0002-0185-1769

Abstract. Abusive language detection models tend to have a gender bias problem in which the model is biased towards sentences containing identity words of specific gender groups. Previous studies to reduce bias, such as projection methods, tend to lose information in word vectors and sentence context, resulting in low detection accuracy. This paper proposes a novel method that mitigates gender bias while preserving original information by regularizing sentence embedding vectors based on information theory. Latent vectors generated by an autoencoder are debiased through dual regularization using a gender discriminator, an abuse classifier, and a decoder. While the gender discriminator labels are randomized, the discriminator confuses the gender feature, and the classifier retains the abuse information. Latent vectors are regularized through information theoretic adversarial optimization that disentangles and mitigates gender features. We show that the proposed method successfully orthogonalizes the direction of the correlated information and reduces the gender feature through calculation of subspaces and embedding vector visualization. Moreover, the proposed method maintains the highest accuracy among the four state-of-the-art bias mitigation methods and shows superior performance in reducing gender bias in four different Twitter datasets for abusive language detection.

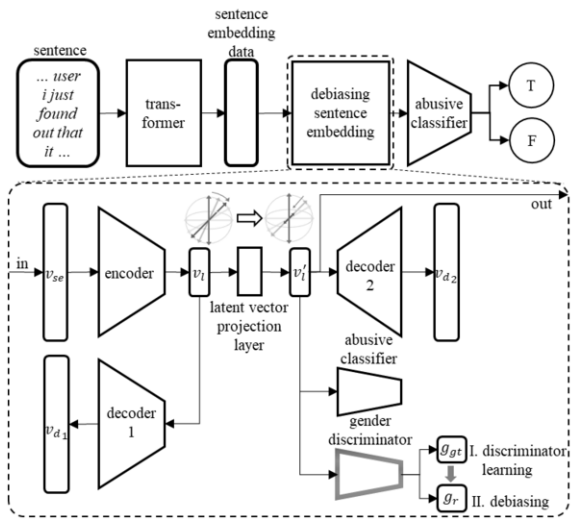

Figure 1. Overall architecture of reducing gender bias in abusive language detection with latent vector regularization.

1 Introduction

As social media become increasingly important in social life, abusive language over there is raising a significant problem like cyberbullying [28]. Many researchers solve the problem with machine learning, such as BERT [44, 24]. However, due to sexism, abusive language with female-related phrases is generated, and gender bias in the abuse classification models occurs, especially for females [39].

For the biases, Dixon et al. [14] defined unintended bias as 'performance difference between comments that contain some identity terms and those that do not', and false-positive bias as 'unreasonably high toxic scores given to clearly non-toxic statements containing some identity terms', and they claimed that those biases raised by various features such as gender and race were important for fairness. Park et al. [31] mentioned that sentences like "you are a good woman" got a high abusive score because of the term 'woman', which represents a false positive bias of gender.

Previous researchers have studied that those biases are associated with a correlation in the direction of gender and abuse subspaces in embedding space [9, 41]. To solve this problem, several methods such as extracting gender-corresponding space from word embedding space and projecting them have been studied, but they have a problem of losing information about the context of word embedding vectors. Bolukbasi et al. [9] suggested a debiasing method using projection in word embedding vector space to gender features subspace representing a difference of gender pair-word embeddings.

Derived from this method, various researchers have investigated several methods to reduce gender bias on word embedding vectors [21, 32] and sentence embedding vectors [23, 8]. Park et al. [31] reduced gender bias of word embeddings in the model using projection, gender swap, and bias fine-tuning. Shin et al. [37] and Kaneko et al. [19] announced that the methods with projection could restrict maintaining original information except for gender and solved it with debiasing word embeddings using latent vector. Thus, a method without projection is necessary for preserving original information.

In this paper, we propose a method to project sentence embedding into the latent space and apply regularization based on information theory to mitigate gender bias for abusive language detection. For tighter bias mitigation, we have applied the information bottleneck theory to project the latent vector into the mutated latent vector, as

* Email: wodlf603@yonsei.ac.kr
** Corresponding Author. Email: sbcho@yonsei.ac.kr

Please check ArXiv or contact the authors for any appendices or supplementary material mentioned in the paper.

Table 1. Previous studies for reducing bias in natural language processing area.

Target embedding	Method	Description
None	Data alteration	Rebalanced dataset and length sensitive upsampling [29]
		Certified mitigation mechanism with metamorphic testing [25]
		Entropy-based attention regularization [6]
		Dynamic fair sampling with selection strategy [36]
	Else	Movement pruning for BERT [18]
Word embedding	Projection	Linear projection [21]
		Hard debias [9]
		Interactive null space projection [32]
		Subspace orthogonal word embedding [1]
	Else	Hard debias, gender swap, and bias fine-tuning [31]
		Iterative adversarial disentanglement [17]
		Dictionary definitions leverage based train-time debiasing [4]
Sentence embedding	Data alteration	Counterfactual data substitution [7]
		Data augmentation and neutralization [46]
	Projection	Hard debias on sentence embedding [23]
		Layered gender subspace projection [8]
	Else	Cross-lingual method based on knowledge distillation with probabilistic rules [12]
		Causal mediation analysis for hidden layer [42]
		Mixup on knowledge distillation [2]
		Adversarial training with CNN and Transformer [45]
Both	Projection	Orthogonal subspace correction and rectification [13]

shown in Figure 1. Moreover, the proposed method limits bias mitigation to the sentences containing gender words and minimizes contextual information loss by reducing the data loss of sentences that do not contain gender words. Comparing to the previous state-of-the-art models on four well-known Twitter datasets of abusive language detection, we verify that the proposed method has an advantage in maintaining accuracy while reducing the gender bias in abusive language detection.

In brief, our contributions are as follows:

- We propose a method of mitigating bias while maintaining original information by regularization with adversarial autoencoder rather than projection, based on information theory.
- We maintain the context while removing the bias by controlling the target of bias mitigation as sentences containing gender-pharse and maintaining the information of sentences not containing the gender-pharse.
- The proposed method compared with the state-of-the-art models on four Twitter datasets of abusive language detection confirms that it mitigates the gender bias without accuracy deterioration, which demonstrates its fairness and superiority.

2 Related Works

Language models in abusive language detection. Pre-trained language models such as BERT and ALBERT or ensembles of them produced good performance compared to other models like LSTM at the competitions for abusive language detection [44]. Liu et al. [24] preprocessed the dataset, fine-tuned it on BERT, and won the 1st prize in sub-task A of the OLID competition. This showed high accuracy by retraining the language dataset on domains that work with a specific model such as BioBERT [22] and FinBERT [5], which are models for other domains of natural language processing. fBERT [35] is a BERT model retrained to SOLID [33], one of the abusive language detection competition datasets, and showed high accuracy in abusive language detection. We use the fBERT as the state-of-the-art abusive language detection model.

Gender bias mitigation methods. Bolukbasi et al. [9] suggested a debiasing method that used projection in vector space of word embeddings to a subspace of gender feature which is generated from the difference of gender pair-words. Followed by this method, several researchers investigated reducing gender bias of NLP tasks, word embedding and sentence embedding, as shown in Table 1. However, there is a problem that the method using projection shows a large decline in performance which is a side effect of the mitigation of bias. On the other hand, Tan and Celis [40] argued that contextual approach for debiasing is needed as some biases like racial biases are strongly encoded in contextual models.

3 Adversarial Discriminator to Mitigate Bias

3.1 Latent vector debiasing via regularizers

The proposed method mitigates bias by focusing how to proceed with regularization of latent vectors in adversarial autoencoders [26]. They compress the original data, generate latent vectors containing the information as much as possible, and adjust the distribution by regularizing the distributions of the original data and the generated data. Adversarial autoencoders exploit the convergence of this distribution and learn the distribution of the original data.

Inspired by the idea of converging and regularizing the data distribution using latent vectors, we attempt to adjust the data distributions of sentences containing female phrases and those containing male phrases for abusive language detection, based on the disentangled representation as Moon et al. [27] did. An adversarial discriminator determines whether the input sentence contains male phrases or female phrases with their gender label, and the data distribution

of them is learned. The distribution of the two data converges with each other through a regularization that confuses gender information using the adversarial discriminator and maintains abuse information using an abuse classifier and original sentence information using a decoder. For producing data that is not tilted in either direction, regularization proceeded by fixing the discriminator and randomizing the gender label.

We aim to learn the debiased representation for the mutated latent vector v_l' that contains all the information of original data x except that of gender feature g on the proposed model, while v_l shows the disentangled representation of gender feature, as shown in Figure 1. It can be shown as equation (1) setting the objective function with information theory, where I represents the mutual information, X is the original sentence embedding data, V_l' is the mutated latent vector, G is the gender attribute, and β is a coefficient to balance the two terms

$$\max \mathcal{L} = I\left(V_l'; X\right) - \beta I\left(V_l'; G\right) \tag{1}$$

Through equation (1), maximizing $I\left(V_l'; X\right)$ which represents the value of association between the original data X and the latent vector V_l, we can maximize the lower bound of equation (1). Inspired by Chen et al. [11], the lower bound of $I\left(V_l'; X\right)$ is calculated as shown in equations (2) to (5) since entropy is non-negative. D_2 is the function of decoder 2, q is encoder function, p is real distribution, m is projection function, and H is the entropy. Let the first term of equation (5) as $\mathcal{L}_r$.

$$I\left(V_l'; X\right) \tag{2}$$

$$= \mathbb{E}_{v_l \sim q(v_l|x), v_l' \sim m(v_l'|v_l)}\left[\mathbb{E}_{v_{d_2} \sim p(x|v_l')}\left[logp(v_{d_2}|v_l')\right]\right] + H(x) \tag{3}$$

$$\geq \mathbb{E}_{v_l \sim q(v_l|x), v_l'\ m(v_l'|v_l)}\left[\mathbb{E}_{v_{d_2} \sim p(x|v_l')}\left[logD_2(v_{d_2}|v_l')\right]\right] + H(x) \tag{4}$$

$$= \mathbb{E}_{v_{d_2} \sim P\left(v_{d_2}\right), v_l' \sim m(v_l'|v_l), v_l \sim q(v_l|x)}\left[logD_2(v_{d_2}|v_l')\right] + H(x) \tag{5}$$

At the same time, minimizing $I\left(V_l'; G\right)$ which represents the value of association between the latent value R and the gender attribute G, we can also maximize the lower bound of equation (1). Inspired by Alemi et al. [3], we can minimize $I\left(V_l'; G\right)$ by reducing the upper bound as shown in equations (6) to (10) since KL divergence is non-negative. $h(z)$ is a variational approximation and D_{KL} is KL divergence. Let the value of equation (10) be C_1.

$$I\left(V_l'; G\right) \leq I\left(V_l'; X, G\right) \tag{6}$$

$$= \mathbb{E}_{p\left(v_l', x, g\right)}\left[\log p\left(v_l' \mid x, g\right) - \log p\left(v_l'\right)\right] \tag{7}$$

$$= \mathbb{E}_{p\left(v_l', x, g\right)}\left[\log p\left(v_l' \mid x, g\right) - \log h\left(v_l'\right) - \log p\left(v_l'\right) + \log h\left(v_l'\right)\right] \tag{8}$$

$$= \mathbb{E}_{p(x,g)}\left[\mathcal{D}_{KL}\left(p\left(v_l' \mid x, g\right) \parallel h\left(v_l'\right)\right) - \mathcal{D}_{KL}\left(p\left(v_l'\right) \parallel h\left(v_l'\right)\right)\right] \tag{9}$$

$$\leq \mathbb{E}_{p(x,g)}\left[\mathcal{D}_{KL}\left(p\left(v_l' \mid x, g\right) \parallel h\left(v_l'\right)\right)\right] \tag{10}$$

However, it is not ideal as considering information of X might cause over-elimination. For the tighter upper bound when assuming optimal dual regularization, inspired by Song et al. [38], $I\left(V_l'; G\right)$ can be approximated as shown in equations (11), where l is positive such that $D_{KL}(p(g|v_l')||h(g|v_l')) \leq l$. Let the right term of equation (11) be C_2.

Algorithm 1: Learning to debias latent vector

Data: Data X and corresponding gender feature G
Result: Fully trained encoder q
Initialize q, D_1, D_2;
for *epochs* **do**
  for *batches* **do**
    Sample x,g, $\widetilde{g}$ from X,G, $U(0,1)$ respectively;
    $\theta_q \leftarrow \theta_q - \eta \frac{\partial \mathcal{L}_r}{\partial \theta_q}(x, \widetilde{g})$;
    $\theta_{D_1} \leftarrow \theta_{D_1} - \eta \frac{\partial \mathcal{L}_r}{\partial \theta_{D_1}}(x, \widetilde{g})$;
    $\theta_q \leftarrow \theta_q - \eta \frac{\partial (C_1 + C_2)}{\partial \theta_q}(x, \widetilde{g})$;
    $\theta_t \leftarrow \theta_t - \eta \frac{\partial (C_1 + C_2)}{\partial \theta_t}(x, \widetilde{g})$;
    $\theta_{D_2} \leftarrow \theta_{D_2} - \eta \frac{\partial C_2}{\partial \theta_{D_2}}(x, g)$;
  end
end
return q ;

$$I\left(V_l'; G\right) \leq \mathbb{E}_{p\left(v_l', g\right)}\left[\log p\left(g \mid v_l'\right) - \log p\left(g\right)\right] + l \tag{11}$$

Final objective function can be formulated as equation (12), which can be applicable to V_l [20].

$$\max \mathcal{L}_r - \beta_1 C_1 - \beta_2 C_2 \tag{12}$$

The latent vector v_l and the mutated latent vector v_l' are learned by L_r and C_i as different directions: L_r learns v_l, v_l' to store all the features of x, but C_i forces v_l, v_l' to be representation without the information of protected feature. To overcome this problem, another formulation for v_l is used for the upper bound based on the information bottleneck theory. A modified upper bound of $I(V_l; G)$ can be obtained as equations (13) to (14) where $s(z)$ is a variational approximation. Algorithm 1 shows its training process.

$$I\left(V_l; G\right) = I\left(V_l; D_2\left(m\left(V_l\right)\right)\right) \leq I\left(V_l; m\left(V_l\right)\right) \tag{13}$$

$$\leq \int m\left(v_l' \middle| v_l\right) p\left(v_l\right) log \frac{m(v_l'|v_l)}{s(v_l')} \tag{14}$$

3.2 Sentence embedding by debiased model

We generate a mutated latent vector that passes through the projection layer such that it can inherit information from the vector. To ensure that the mutated latent vectors maintain the information about the original input data as much as possible, the vectors are also decoded and compared with the original sentence embedding data. Mutated latent vectors of a sentence containing a female phrase and those containing a male phrase are classified through a gender discriminator, so that the discriminator learns the distribution of the two latent vectors. For the discriminator to be learned, gender labels of input sentences are generated automatically by checking that gender words are contained in sentences, which does not take much time. We use generic gender words as dictionary-defined male/female words [13]. Their examples are described in Table 2. An abuse classifier is also trained with the ground-truth abuse label.

After training, the discriminator and the classifier are fixed, and the gender label of the mutated latent vector is rumpled. As the discriminator and the classifier are fixed, the mutated latent vector is regularized as softening gender information and preserving abuse information. The projection layer is trained through randomized labels

Table 2. Examples of male terms and female terms.

Male terms	Female terms
he him masculine actor	she her feminine actress
boy brother count daddy	girl sister countess
duke man emperor father	mummy duchess woman
grandfather heir host	empress mother
husband king master	grandmother heiress
nephew prince sir son	hostess wife queen
uncle wizard waiter	mistress niece princess
boyfriend dad gentleman	madam daughter aunt
monk priest baron abbot	witch waitress girlfriend
	mom lady nun priestess
	baroness abbess

to learn the distribution of data in the direction of convergence between mutated latent vectors of the sentence with the gender phrase. Abusive information is maintained by the abuse classifier, and the original information without them is maintained where mutated latent vectors compare the restored vector by the decoder with the original sentence embedding vector. Thus, through adversarial learning between the discriminator and the classifier within the decoder, mutated latent vectors are transformed where the gender attribute is removed and the abuse data are maintained.

Although the main problem is to mitigate bias between the sentences containing female terms and those containing male terms, there are sentences that do not contain both and sentences that contain both, and it is also necessary to preserve the content of these sentences [19]. This problem is solved by adjusting the classification target of the discriminator. Neutral sentences are excluded as the target of the discriminator, minimizing the effect due to bias mitigation, and comparing them with the original data through a decoder of a latent vector to determine whether the data information is well preserved. Although those sentences do not train the discriminator, data preservation on those data is guaranteed, passing through the bias mitigation algorithm and its decoder and comparing the results with the original sentence embedding data.

4 Experiments

4.1 Experimental setting

Experiments are conducted in an environment using the Transformer-based fBERT learned with the Twitter datasets for abusive language classification. fBERT is a model that has trained BERT on a Twitter dataset, and shown high accuracy in abusive language classification [35]. In this paper, the results between the models of applying the bias mitigation method to fBERT are compared and analyzed. The alternative methods for comparison include OSCaR [13], SentDebias [23], INLP [32] and the baseline without debiasing method. Experiments are conducted on the four Twitter datasets. Founta [16] is a large-scale dataset with abusive, offensive, and hate-speech labels by considering various data. OLID [44] is a relatively small dataset for abusive language detection, considering the performance on small data where regularization is more challenging than projection. Waseem [43] is a dataset that considers sexism part of abusive language with gender, and CMSB [34] considers sexism in psychological scales, which makes it hard to detect. For those datasets, we classify the sentences as male, female, or else, according to whether the sentence contains gender terms or not, to generate gender labels. If the sentence only contains gender terms about males, we classify it as a male sentence and, otherwise, similarly, as a female sentence.

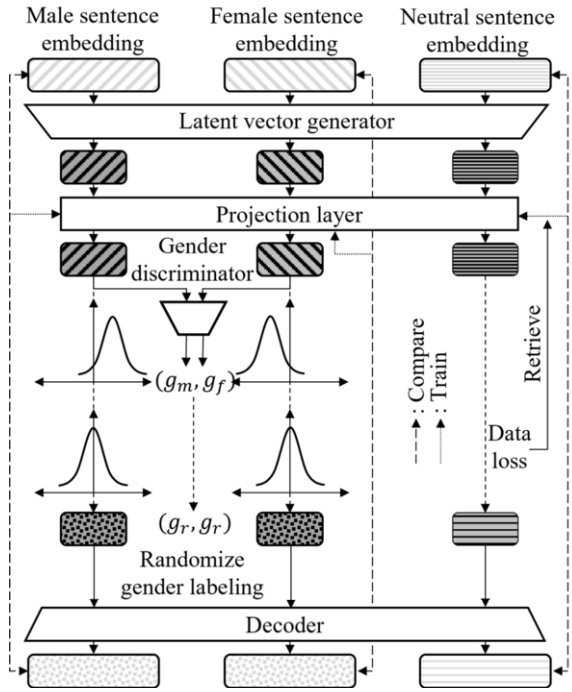

Figure 2. Learning with debiasing in context-based sentence embeddings.

Detailed information for pre-processing of datasets is described in Appendix A.

For experiments, we implement the methods on Windows operating system with the Intel Core i7 9700KF CPU and the NVIDIA GeForce RTX 3090. For PLMs, we set the same hyperparameters in the paper of fBERT. The autoencoder's input and output dimensions are set to 768 which is the same as fBERT's [CLS] token dimension, and the latent vector dimension is set to 400. The autoencoder's learning rate is 1e-4, and 7 patience early stopping is applied to every 5000 epochs, and fine-tuning is conducted.

4.2 Latent vector debiasing via regularizers

The experimental results are evaluated on the two types of datasets, original and generated datasets. Original datasets can show how bias appears against the original data distribution that is close to the real situation. Generated datasets extract the sentences containing female and male phrases from original datasets, and then even convert them into male and female phrases respectively, and the generated dataset that makes the distributions of sentences containing female and male phrases have the same distribution of bias. As Park et al. [31] used, FPED and FNED are used for the bias evaluation, and AUC is used for the accuracy evaluation. FPED and FNED are calculated as follows,

$$FPED = \Sigma_{t \in T} |FPR - FPR_t| \tag{15}$$

$$FNED = \Sigma_{t \in T} |FNR - FNR_t| \tag{16}$$

where FPR is a false positive rate, FNR is a false negative rate, t is each group, and T is a set of all groups. The results with 10-fold cross-validation of the four datasets are shown in Table 3. For each item, the best figures are in bold, and the second best figures are underlined.

Table 3. Experimental results with 10-fold cross-validation on the original and generated datasets (%). First place written in **bold** and second place written in underlined.

Type	Dataset	Metric	Original	OSCaR [13]	SentDebais [23]	INLP [32]	Ours
Original dataset	Founta [16]	AUC	**93.8**	93.5	93.6	93.7	93.7
		FPED	2.32	**1.20**	2.53	1.92	1.87
		FNED	3.71	6.21	3.46	6.34	**3.44**
	OLID [44]	AUC	**84.1**	79.2	83.4	75.5	82.3
		FPED	0.630	1.98	0.649	0.385	**0.329**
		FNED	3.47	1.33	3.45	0.418	**0.201**
	CMSB [34]	AUC	**96.5**	94.3	96.3	88.2	95.1
		FPED	0.121	0.443	**0.0117**	0.502	0.060
		FNED	9.54	3.61	4.43	12.4	**3.21**
	Waseam [43]	AUC	**90.8**	88.4	90.6	86.2	90.5
		FPED	1.57	1.32	1.36	24.6	**0.452**
		FNED	9.35	4.43	5.27	6.13	**3.85**
Generated dataset	Founta [16]	AUC	92.3	91.9	92.5	91.5	**92.8**
		FPED	0.262	0.654	0.131	0.314	**0.0654**
		FNED	0.251	**0.036**	0.0835	0.332	0.167
	OLID [44]	AUC	**83.9**	81.2	83.7	77.1	82.7
		FPED	0.0927	1.72	**0.0432**	0.331	0.210
		FNED	0.314	0.627	0.537	1.33	**0.0615**
	CMSB [34]	AUC	94.7	89.2	**94.9**	84.8	94.3
		FPED	0.0584	0.0562	**0.0137**	0.0192	0.0188
		FNED	3.01	1.01	0.0442	0.0257	**0.0218**
	Waseam [43]	AUC	**86.7**	84.2	84.5	77.9	86.5
		FPED	0.391	0.146	0.142	16.8	**0.0751**
		FNED	0.941	**0.0132**	0.651	9.87	1.42

OSCaR shows a good performance on bias mitigation, and it has advantages in reducing the number of FPEDs and FNEDs. However, due to the low contextual influence, the accuracy deteriorate due to the poor maintenance of the information. SentDebias maintains the contextual information, but it can be confirmed that the simple projection method results in data loss and relatively low bias mitigation. INLP mitigates bias well on generated datasets, but it couldn't maintain the accuracy, mainly on difficult datasets like CMVB and Waseam.

For most of the metrics of the datasets conducted in the experiment, the proposed method occupies the first or second place. Unlike OSCaR, which exhibits a significant loss, it shows a low numerical drop in AUC and particularly good performance for the generated dataset. Unlike other algorithms, it also shows a particular strength in the FNED figures, signifying that it better solves "unintended bias" including false-negative error. Our method mitigates bias superiorly among other comparable methods while preserving accuracy.

It should be noted that the proposed model exhibits higher utility in the generated dataset. The sentences containing female phrases and those containing male phrases that have the same distribution are more effective than other algorithms, i.e., theoretical algorithmic bias in the ideal data state. This is because the method involves equating the data distribution, resulting in higher results in ideal situations. The proposed method takes shorter time than OSCaR. It controls only the result of fBERT, while OSCaR and SentDebias are applied to inner layers of fBERT. In the experiment, it has taken around 188 minutes to learn for Founta dataset, whereas OSCaR requires over six hours. However, as SentDebias takes less than an hour, there is a room to improve the method to reduce the training time while retaining the performance of bias mitigation.

Figure 3 shows the results of case analysis to determine whether the proposed method has strength for a particular case or not. The sentences with (a) sexist comments, (b) occupation words related to gender, (c) gender-related spam sentences, and (d) LGBT-related words are classified, and the results of the proposed method for each case are graphically presented. Detailed information of sentence classification is explained in Appendix B. The sentences with sexist comments are analyzed to confirm whether the removal of the bias affects the content of gender discrimination that is the basis of the problem. The sentences containing occupation words related to gender are analyzed to confirm whether the problem presented by Bolukbasi et al. [9] is affected by the bias between gender words and occupation words. The sentences with gender-related spam term or LGBT-related words are analyzed to confirm whether the sentences with bias that are not explicitly affected by the model are well mitigated or not.

In the case (a), the proposed model shows lower bias mitigation metric value, which confirms that the proposed method has achieved bias mitigation suitable for the purpose. In the case (b), the proposed model also shows lower bias mitigation metric value, and it can be determined that this also affects the connectivity between gender-occupations, which has been raised as a problem. (c) and (d) do not produce good results by the proposed method.

4.3 Disentanglement on sentence embedding vector

In order to establish the relationship between the transformation of the gender feature and bias mitigation, the latent vectors of the model, v_l and v_l', have been measured through a metric to determine how well the information theory-based adversarial discrimi-

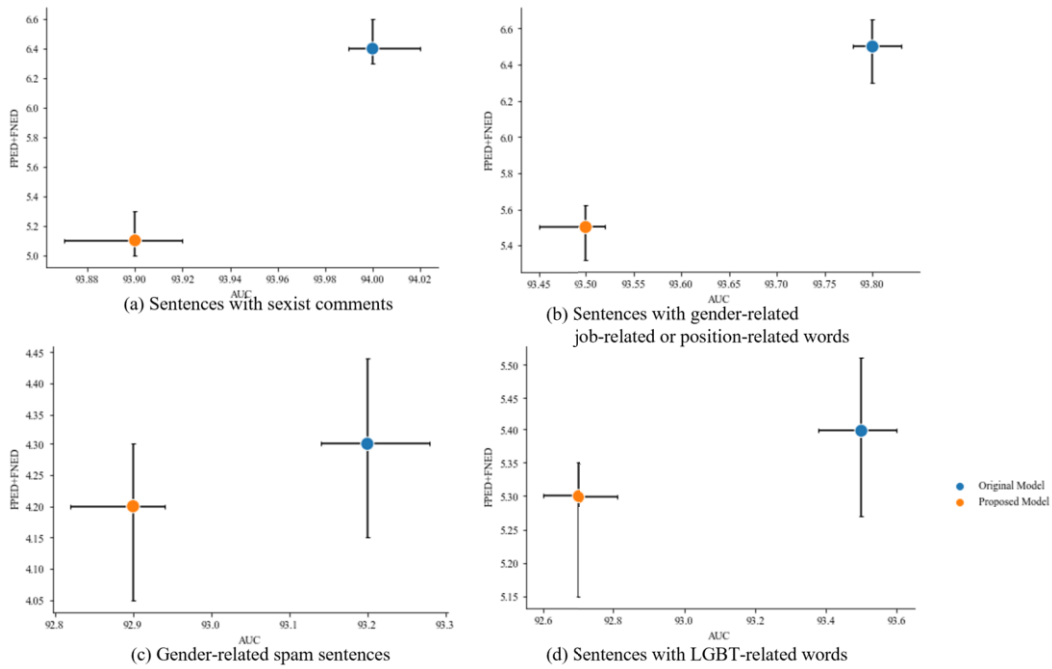

Figure 3. Results of the case study on Founta dataset. This represents the average and standard deviation for 10 runs of experiments.

nator performs. The experiment was executed based on the original version of the large-scale dataset, Founta dataset, and the distribution of v_l before the randomization of gender feature, v_l, and v_l' after the randomization have been measured through visualization using t-SNE and metric. Metric has been used DCI-Random Forest [10] by adopting the analysis results of Carbonneau et al. [15]. The results of this experiment can be found in Figure 4. Blue dots represent male sentences, red dots represent female sentences, dark colors represent abusive sentences and light colors represent non-abusive sentences. (a) is the result of v_l before the gender label is randomized, (b) is the result of v_l after randomization, and (c) is the result of v_l'.

Visualization can confirm the distribution of each gender sentence. (a) suggests that each gender sentence is not completely separated based on its relationship with abuse, indicating that bias exists. The lowest DCI-Random Forest metric among the three environments supports this. (b) can be seen that each gender is separated, and abuse is distributed within it. The highest metric level proves that discrimination has performed well in the environment. Finally, (c) is a vector from which the gender attribute has escaped, and it can be seen that each gender appears relatively close. The metric level of the median value supports this. However, in this figure, it can be seen that the distribution of abuse within a gender is separated. At the same time, the boundaries of the joints are more clearly visible than (a), which shows many light red points positioned inside of abusive cluster and many dark blue points positioned inside of non-abusive cluster.

4.4 *Training with different learning ratios*

The decoder, the classifier, and the discriminator are learned with the same ratio. We have experimented with adjusting the degree of regularization by controlling the learning ratio of the discriminator to the decoder and the classifier, which relates to the beta value in equation (1). The corresponding experiment adjusts the degree of regularization, confirming how adversarial learning among the decoder, the classifier, and the discriminator influences the bias mitigation and reduced accuracy. This experiment is conducted with the models with different learning ratios of decoders and discriminators for the Founta dataset. Table 4 illustrates its result.

The accuracy gets lower through adversarial learning as shown in the AUC of the original dataset, and the range of bias mitigation is significantly lowered when the ratio exceeds. This means that the bias mitigation that can be expected when learning at the same rate is high, and the higher rate is an appropriate learning ratio for bias mitigation because the bias is not properly mitigated compared to the accuracy decrease.

However, as shown in Table 4, the adversarial learning ratio does not have the same value as the direct accuracy and trade-off between bias mitigations. Also, beta does not have a logical causal relationship with the trade-off [30]. Accordingly, to obtain the desired trade-off value, we need to conduct the same experiment as this section.

4.5 *Ablation study*

In the proposed model, the results of the absence of the discriminator and decoder are checked to confirm the effect of each component on bias mitigation. For the proposed method, experiments are conducted on the original model, the model without discriminator, and the model without decoder applied to the original latent vector without the projection layer. An experiment is conducted by checking the metric values for the Founta dataset. Table 5 shows its result.

If the decoder is removed, the original data cannot be maintained during projection, resulting in a significant decrease in accuracy and an increase in bias metric due to data loss. As a result, it can be confirmed that the decoder plays an important role in maintaining original data. If the discriminator is removed, the original data is retained, but the bias is not mitigated properly, resulting in high FPED and FNED for the dataset.

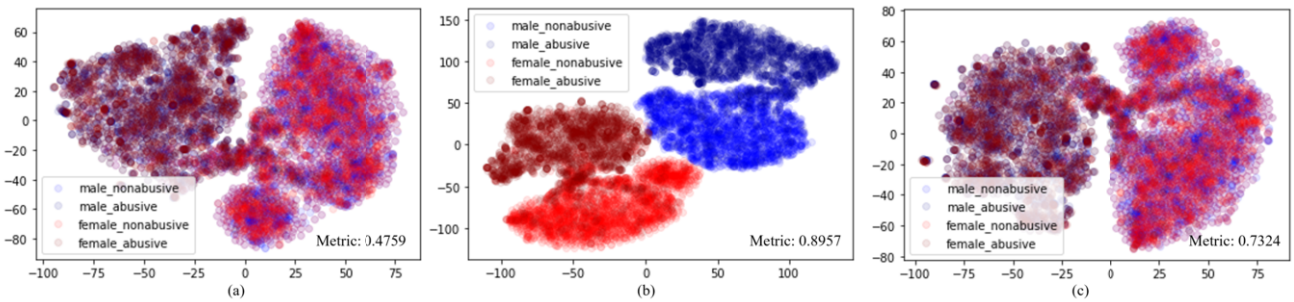

Figure 4. Results of visualization and DCI-Random Forest metric analysis on Founta dataset.

Table 4. Training with different adversarial learning ratios.

Dataset	Metric	Adversarial learning ratio				
		0	0.5	1	1.5	2
Original	AUC	93.9	94.1	93.9	93.7	93.8
	FPED	2.50	1.41	1.84	1.92	2.21
	FNED	3.46	3.97	3.46	3.46	3.26
Generated	AUC	92.4	92.6	92.8	92.8	92.9
	FPED	0.392	0.0654	0.0654	0.0654	0
	FNED	0.251	0.251	0.167	0.167	0.167

Table 5. Ablation study. (a) is the result of decoder removal and (b) is that of discriminator

Dataset	Metric	Original	(a)	(b)
Original	AUC	93.9	50.3	93.9
	FPED	1.84	20.2	2.50
	FNED	3.46	18.3	3.46
Generated	AUC	92.8	48.9	92.4
	FPED	0.0654	22.1	0.392
	FNED	0.167	17.5	0.250

5 Concluding Remarks

In this paper, we argue that bias exists for abusive language classification depending on whether gender-related words are included, and present the limitations of the existing methods in natural language processing. A method of regularizing the distribution of data between genders is presented by removing the bias using latent vectors based on sentence embeddings, and experiments show that the method yields good results for "unintended bias", which minimizes the reduced accuracy and mitigates the bias. However, to find the optimal beta value and control the learning rate, optimizing the learning rate between the discriminator and the classifier through experiments also needs to be improved. Nevertheless, the high bias mitigation performance and the high maintenance of the original information demonstrate the superiority of the proposed method.

For the future work, the proposed method will be applied to other domains like translation in natural language processing, where the bias occurs more frequently. Furthermore, we will investigate how to increase the expandability of the method by reasoning with the latent vector and by adjusting parameters. Moreover, we need to devise the customized method of bias mitigation according to the users.

Acknowledgements

This work was supported by the Yonsei Fellow Program funded by Lee Youn Jae, and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No. 2022-0-00113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework).

References

[1] Prince Osei Aboagye, Yan Zheng, Jack Shunn, Chin-Chia Michael Yeh, Junpeng Wang, Zhongfang Zhuang, Huiyuan Chen, Liang Wang, Wei Zhang, and Jeff Phillips, 'Interpretable debiasing of vectorized language representations with iterative orthogonalization', in *The Eleventh International Conference on Learning Representations*, (2023).

[2] Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh, 'Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of distilbert', in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 266–272, (2022).

[3] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, 'Deep variational information bottleneck', *arXiv preprint arXiv:1612.00410*, (2016).

[4] Haozhe An, Xiaojiang Liu, and Donald Zhang, 'Learning bias-reduced word embeddings using dictionary definitions', in *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1139–1152, (2022).

[5] Dogu Araci, 'Finbert: Financial sentiment analysis with pre-trained language models', *arXiv preprint arXiv:1908.10063*, (2019).

[6] Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis, 'Entropy-based attention regularization frees unintended bias mitigation from lists', in *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1105–1119, (2022).

[7] Marion Bartl, Malvina Nissim, and Albert Gatt, 'Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias', in *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pp. 1–16, (2020).

[8] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria, 'Investigating gender bias in bert', *Cognitive Computation*, **13**(4), 1008–1018, (2021).

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai, 'Man is to computer programmer as woman is to homemaker? debiasing word embeddings', *Advances in neural information processing systems*, **29**, (2016).

[10] Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon, 'Measuring disentanglement: A review of metrics', *IEEE Transactions on Neural Networks and Learning Systems*, (2022).

[11] Irene Chen, Fredrik D Johansson, and David Sontag, 'Why is my classifier discriminatory?', *Advances in neural information processing systems*, **31**, (2018).

[12] Pieter Delobelle and Bettina Berendt, 'Fairdistillation: mitigating stereotyping in language models', in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part II*, pp. 638–654. Springer, (2023).

[13] Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar, 'Oscar: Orthogonal subspace correction and rectification of biases in word embeddings', *arXiv preprint arXiv:2007.00049*, (2020).

[14] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman, 'Measuring and mitigating unintended bias in text classification', in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, (2018).

[15] Cian Eastwood and Christopher KI Williams, 'A framework for the quantitative evaluation of disentangled representations', in *International Conference on Learning Representations*, (2018).

[16] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis, 'Large scale crowdsourcing and characterization of twitter abusive behavior', in *Twelfth International AAAI Conference on Web and Social Media*, (2018).

[17] Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem, 'Iterative adversarial removal of gender bias in pretrained word embeddings', in *Proceedings of the 37th ACM/SIGAPP Symposium On Applied Computing*, pp. 829–836, (2022).

[18] Przemyslaw Joniak and Akiko Aizawa, 'Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning', *arXiv preprint arXiv:2207.02463*, (2022).

[19] Masahiro Kaneko and Danushka Bollegala, 'Gender-preserving debiasing for pre-trained word embeddings', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1641–1650, (2019).

[20] Jin-Young Kim and Sung-Bae Cho, 'An information theoretic approach to reducing algorithmic bias for machine learning', *Neurocomputing*, **500**, 26–38, (2022).

[21] Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić, 'A general framework for implicit and explicit debiasing of distributional word vector spaces', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8131–8138, (2020).

[22] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, 'Biobert: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, **36**(4), 1234–1240, (2020).

[23] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency, 'Towards debiasing sentence representations', *arXiv preprint arXiv:2007.08100*, (2020).

[24] Ping Liu, Wen Li, and Liang Zou, 'Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers.', in *SemEval@ NAACL-HLT*, pp. 87–91, (2019).

[25] Pingchuan Ma, Shuai Wang, and Jin Liu, 'Metamorphic testing and certified mitigation of fairness violations in nlp models', in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 458–465, (2021).

[26] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, 'Adversarial autoencoders', *arXiv preprint arXiv:1511.05644*, (2015).

[27] Hyung-Jun Moon, Seok-Jun Bu, and Sung-Bae Cho, 'Learning disentangled representation of residential power demand peak via convolutional-recurrent triplet network', in *2020 International Conference on Data Mining Workshops (ICDMW)*, pp. 757–761. IEEE, (2020).

[28] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang, 'Abusive language detection in online user content', in *Proceedings of the 25th international conference on world wide web*, pp. 145–153, (2016).

[29] Debora Nozza, Claudia Volpetti, and Elisabetta Fersini, 'Unintended bias in misogyny detection', in *Ieee/wic/acm international conference on web intelligence*, pp. 149–155, (2019).

[30] Ziqi Pan, Li Niu, Jianfu Zhang, and Liqing Zhang, 'Disentangled information bottleneck', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9285–9293, (2021).

[31] Ji Ho Park, Jamin Shin, and Pascale Fung, 'Reducing gender bias in abusive language detection', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, (2018).

[32] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg, 'Null it out: Guarding protected attributes by iterative nullspace projection', *arXiv preprint arXiv:2004.07667*, (2020).

[33] Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov, 'Solid: A large-scale semi-supervised dataset for offensive language identification', *arXiv preprint arXiv:2004.14454*, (2020).

[34] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner, '"call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples', in *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pp. 573–584, (2021).

[35] Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia, 'fbert: A neural transformer for identifying offensive content', in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1792–1798, (2021).

[36] Lele Sha, Yuheng Li, Dragan Gasevic, and Guanliang Chen, 'Bigger data or fairer data? augmenting bert via active sampling for educational text classification', in *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1275–1285, (2022).

[37] Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon, 'Neutralizing gender bias in word embeddings with latent disentanglement and counterfactual generation', in *Empirical Methods in Natural Language Processing conference (EMNLP) 2020*, pp. 3126–3140. Association for Computational Linguistics, (2020).

[38] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon, 'Learning controllable fair representations', in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2164–2173. PMLR, (2019).

[39] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang, 'Mitigating gender bias in natural language processing: Literature review', *arXiv preprint arXiv:1906.08976*, (2019).

[40] Yi Chern Tan and L Elisa Celis, 'Assessing social and intersectional biases in contextualized word representations', *Advances in neural information processing systems*, **32**, (2019).

[41] Francisco Vargas and Ryan Cotterell, 'Exploring the linear subspace hypothesis in gender bias mitigation', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2902–2913, (2020).

[42] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber, 'Investigating gender bias in language models using causal mediation analysis', *Advances in neural information processing systems*, **33**, 12388–12401, (2020).

[43] Zeerak Waseem and Dirk Hovy, 'Hateful symbols or hateful people? predictive features for hate speech detection on twitter', in *Proceedings of the NAACL student research workshop*, pp. 88–93, (2016).

[44] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar, 'Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)', in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 75–86, (2019).

[45] Guanhong Zhang, Sophia Ananiadou, et al., 'Examining and mitigating gender bias in text emotion detection task', *Neurocomputing*, **493**, 422–434, (2022).

[46] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang, 'Gender bias in coreference resolution: Evaluation and debiasing methods', *arXiv preprint arXiv:1804.06876*, (2018).