Words Can Be Confusing: Stereotype Bias Removal in Text Classification at the Word Level

Shaofei Shen[1], Mingzhe Zhang[1], Weitong Chen[2(✉)], Alina Bialkowski[1], and Miao Xu[1,3]

[1] The University of Queensland, Brisbane, Australia
{shaofei.shen,mingzhe.zhang,alina.bialkowski,miao.xu}@uq.edu.au
[2] University of Adelaide, Adelaide, Australia
t.chen@adelaide.edu.au
[3] RIKEN, Tokyo, Japan

Abstract. Text classification is a widely used task in natural language processing. However, the presence of stereotype bias in text classification can lead to unfair and inaccurate predictions. Stereotype bias is particularly prevalent in words that are unevenly distributed across classes and are associated with specific categories. This bias can be further strengthened in pre-trained models on large natural language datasets. Prior works to remove stereotype bias have mainly focused on specific demographic groups or relied on specific thesauri without measuring the influence of stereotype words on predictions. In this work, we present a causal analysis of how stereotype bias occurs and affects text classification, and propose a framework to mitigate stereotype bias. Our framework detects potential stereotype bias words using SHAP values and alleviates bias in the prediction stage through a counterfactual approach. Unlike existing debiasing methods, our framework does not rely on existing stereotype word sets and can dynamically evaluate the influence of words on stereotype bias. Extensive experiments and ablation studies show that our approach effectively improves classification performance while mitigating stereotype bias.

Keywords: Text Mining · Text Classification · Stereotype Bias · Causal Inference

1 Introduction

Text classification tasks in natural language processing (NLP) can be influenced by **stereotype words**, which are words associated with specific categories or groups based on emotions, politics, or demographic features [1]. Studies have shown that stereotype words such as *pink* and *blue* are often associated with girls and boys, respectively [2]. Additionally, words such as *Varicella* or *Alzheimer* are associated with specific age groups. The distribution of stereotype words across different document categories can result in **stereotype bias** in classification models trained on datasets containing these words [14]. This bias can have significant

H. Kashima et al. (Eds.): PAKDD 2023, LNAI 13938, pp. 99–111, 2023.
https://doi.org/10.1007/978-3-031-33383-5_8

implications, particularly in sentiment analysis, where the results can influence decision-making processes.

Stereotype bias in text classification is caused by oversimplified correlations between stereotype words and text categories. However, in text classification, the semantic relationships of the words should be the basis for classification, not the existence of specific words [15]. In this paper, we focus on the problem of detecting and alleviating stereotype bias in text classification. To solve this problem, three challenges need to be addressed. Firstly, identifying the set of stereotype words that can potentially introduce stereotype bias is critical [18]. Such words are usually domain-specific and difficult to identify universally across different domains, given that document classification tasks are often domain-specific, such as in movie reviews or medical research. Secondly, accurately estimating the degree to which a word contributes to stereotype bias in predictions for a document is challenging. This is especially true for complex deep models, and the same words may contribute differently in different documents due to the interdependence between words. Simply removing a particular stereotype word may not eliminate stereotype bias. Thirdly, aside from stereotype bias in classifiers, there may be stereotype bias in widely-used pre-trained word embedding models [24]. Accessing the training data of these models to detect stereotype words is difficult, and it is challenging to measure and reduce stereotype bias without the original training data. Addressing these challenges is essential for developing more inclusive NLP models that are free from stereotype bias.

Previous works have attempted to alleviate various forms of stereotype bias, with a particular focus on removing gender bias in language models [4,24,25]. In [25], the authors reduced gender bias through data augmentation using an occupation word set associated with gender bias. Similarly, [24] proposed reducing gender bias in the word embedding stage. [4] demonstrated bias amplification in language models and used posterior regularization to address it. Other forms of bias, including label bias [17], context-word bias [17], race bias [6], demographic bias [10], and implicit bias [9], have also received attention. However, these works typically rely on fixed word sets considered as stereotype words derived from a different domain, which may not be effective in the current domain. While some works [17] have proposed selecting stereotype words, their selection strategy is based solely on the TextRank score [13], without considering other important metrics, such as word imbalance. Moreover, these works assume that all words contribute equally as stereotype bias creators, without taking into account the fact that different words may contribute differently in different documents.

In this work, a causal graph is constructed to analyze how the stereotype bias from texts and pre-trained models affects classifications, building on previous works. Based on this causal graph, a novel framework is proposed for detecting and alleviating stereotype bias using a counterfactual method to address the three challenges mentioned earlier. Initially, word distribution statistics and word importance (i.e. SHAP value) in prediction are used to determine a dynamic stereotype word set. Subsequently, a fusion model is adopted to learn the relationship between semantics, stereotype words, and text categories. During the prediction stage, the counterfactual approach was used to alleviate the bias from

stereotype words. In contrast to previous works, this study utilizes real-time word importance in document-level predictions and domain-level word distributions to identify stereotype words in different document domains, and focuses more on semantically relevant words instead of context words [17]. In summary, the contributions of this work are three-fold.

- We investigate the stereotype bias in text classification from a causal perspective, analyzing how stereotype words from both texts and pre-trained models influence classification results.
- We propose a novel framework to detect and remove stereotype bias, which involves detecting stereotype words based on word importance and word distribution statistics, training a fusion model to learn the relationship between semantics, stereotype words, and text categories, and utilizing a counterfactual approach for unbiased prediction. To the best of our knowledge, this is the first work that systematically addresses the stereotype bias caused by semantic words without relying on a prior thesaurus.
- We conduct extensive experiments to demonstrate the effectiveness of our framework in achieving unbiased classification, and we compare our results with state-of-the-art approaches for unbiased text classification [17].

Related Work. The word-level bias in language models has attracted the growing interest of researchers. Apart from the aforementioned works on gender bias [24, 25], the most recent work handled gender bias via an adaptation perspective and treated gender groups as different domains [3]. Apart from the gender bias, other works also focus on intended bias [22] and the stereotype bias generated from words [3,17,25], word embeddings [24], and pre-trained models [14]. These works inspired us to model the causal relationship among texts, sources of stereotypes bias, and predictions.

As for the debiasing methods, the counterfactual approach is attracting increasing attention. The counterfactual approach utilizes a dummy value as the counterfactual and aims to remove the indirect effect of confounders on the treatment variables [16]. [17] proposed to use a counterfactual method to remove the bias from imbalanced labels and semantically-irrelevant words. [21] removed the bias in fake news classification and [20] mitigated the bias in text understanding and hypothesis inference via counterfactual debiasing. As discussed in the aforementioned challenges, training data is generally unavailable or inaccessible in pre-trained models. In this case, the counterfactual approach is applicable to mitigate the potential stereotype bias in pre-trained models and text classifiers.

2 Methodology

2.1 Problem Formulation

Let $\mathbf{D}$ and $\mathbf{Y}$ denote the text documents and text categories, respectively. Considering a pre-trained word embedding model $\mathbf{h}$ and classification model $\mathbf{g}$, the goal of the text classification is to train the classification model $\mathbf{g}$ to maximize

the classification accuracy of $(\mathbf{g} \circ \mathbf{h})(\mathbf{D})$. In the ideal view of the training process, the semantics of documents can be learned through a two-stage model, which first classifies semantics and then performs text classification. Then the semantics will be the main basis of the text classification [14]. However, when training from a pre-trained model, the text classification model will inherit any existing stereotype relationships of the pre-trained model.

To construct the causal relationship among these variables, we analyse the word-level stereotypes first. The words in one document can be divided into three groups: the semantic-irrelevant words which have no contribution to the semantics and the further text category predictions, the normal words which are related to the semantics but will not involve stereotypes in the predictions, and the stereotype words that affect the semantics and introduce the stereotype bias in the predictions meantime. In addition, the pre-trained word embedding model may also involve stereotype bias due to the pre-training dataset. Figure 1(a) demonstrates the causal relationship among these groups of words, semantics, and text category. Ideally, the pre-trained word embedding M should also contribute to the semantics X and be the confounder of causal path $X \rightarrow Y$. However, the causal effect from M on X is hard to estimate and we remove the path $M \rightarrow X$ for easier implementation in the experiments. Considering all the sources of bias, the prediction results can be denoted as:

$$Y_{x(d,s),m,s} = Y(X = x(d,s), M = m, S = s) \tag{1}$$

where Y is a prediction function based on word embedding m, normal words d, stereotype words s, and the specific semantics $x(d,s)$.

The unbiased prediction requires using the semantic as the only direct causal variable to the predictions. Then we need to remove the causal effect from the other two causal variables: stereotype bias from pre-trained model $\mathbf{M}$ and stereotype words $\mathbf{S}$ as shown in Fig. 1(b). Then the debiasing goal can be denoted as:

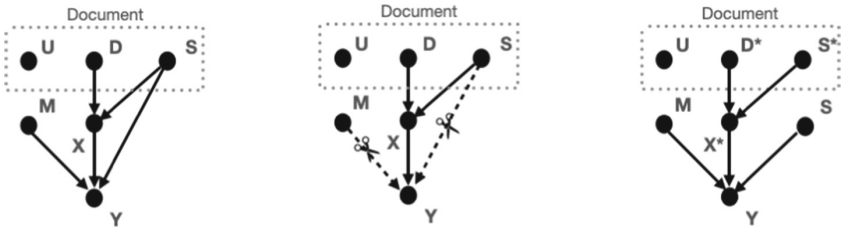
$$Y_{x(d,s)} = Y(X = x(d,s)) \tag{2}$$

where Y is only decided on the semantic $x(d,s)$ of the texts.

Based on the analysis and causal relationship in Fig. 1, our framework contains three stages: stereotype word set construction, fusion model training to learn the causal effect from the sources of bias, and unbiased prediction.

2.2 Stereotype Words Detection

To mitigate the stereotype bias from the training documents, the first stage of our framework is to select the potential words that may lead to stereotypes. As Fig. 1 shows, we assume the stereotype words have a direct causal effect on the semantics and predictions at the same time. Therefore, we can focus on the words that contribute to predictions and utilize the **word importance on predictions** to detect stereotype words. [17] proposes to use the TextRank-based method to calculate the word importance in the document. However, TextRank can only select the keywords and does not consider whether these words affect



(a) Causal graph after involving stereotypes

(b) Cut off stereotypes to semantics in training

(c) Cut off stereotypes to predictions in test

Fig. 1. Conventional text classification and debiased text classification. The words in the document are composed of three parts: **D** donates the normal words, **S** donates the potential stereotype words from the dataset, and **U** donates the semantic-irrelevant words. **M** denotes the stereotypes in existing pre-trained models, and **X** is the semantic embedding of the document and **Y** is the prediction results of the documents. (a) shows a causal graph after introducing stereotype bias from texts and pre-trained models, (b) illustrates the goal of mitigating stereotype bias: removing the direct causal effect from **S** and **M** to **Y**, and (c) is our proposed method of unbiased predictions via a counterfactual approach. In language models, the causal effects from words **D** and **S** to predictions **Y** needs the mediator variable **X**. Therefore, we focus on the causal effect in path $S \rightarrow Y \leftarrow X$ instead of the path $S \rightarrow Y \leftarrow D$. Moreover, in this causal graph, we remove the path $M \rightarrow X$ for easier implementation.

downstream tasks while we are aiming to get the word importance on the downstream predictions in this work. Therefore, we adopt the post-training SHAP value, which can provide the feature importance (i.e. word importance in this work) for the predictions [7,12] and provide the contribution of each word to the predictions based on the same prediction model. Moreover, another characteristic to select the stereotype words is the word distribution in different classes.

Specifically, as shown in stage 1 of Fig. 2, after the initial training stage, we calculate the SHAP values as the word importance for the words in training data and select the set of words $\mathbf{D} + \mathbf{S}$ which contributes to the predictions. Then to select the potential stereotype words, we calculate the word distributions in each document class and rank them via information entropy:

$$H(w) = -\sum_{c \in C} p(c|w) \log p(c|w) \tag{3}$$

where w is all semantic-relevant words and C is the text category set. Lower $H(w)$ means a more imbalanced distribution in different text classes and a higher potential to involve stereotypes in the predictions. Then we set the proportion of stereotype words as a parameter and select the percentage of data from the ranking of $H(w)$.

2.3 Fusion Model Training

After selecting the potential stereotype words, we build a fusion model to estimate the direct causal effect from the pre-trained model and stereotype words,

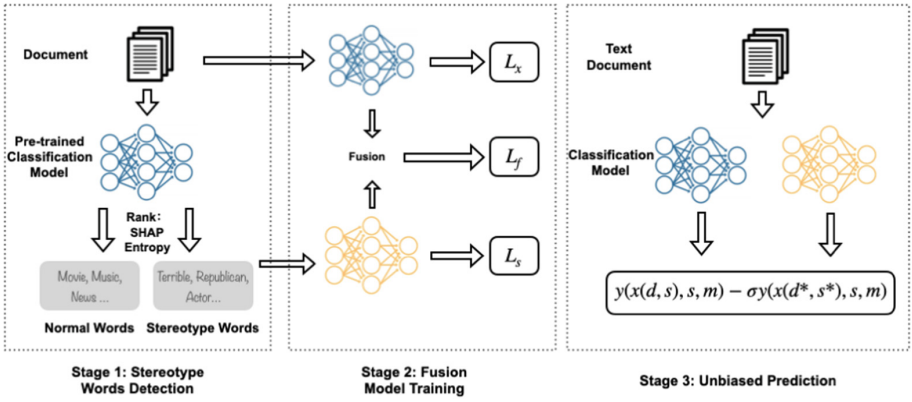


Fig. 2. The proposed framework to mitigate stereotype bias in text classification. This framework contains three stages: stereotype word detection after the first training; fusion model training to learn the causal relationships of stereotype bias and prediction results; unbiased prediction to alleviate stereotype bias. Moreover, the blue and yellow models in the framework indicate the pre-trained and de-biased classification model respectively (Color figure online)

which is shown as $S \rightarrow Y$ and $M \rightarrow Y$ in Fig. 1. Inspired by [20,21], we build two models $\hat{y}_{d,s}$ and $\hat{y}_s$. We use the original texts as input of the first model and predict the corresponding text categories to capture the causal relationships in the paths $S \rightarrow X \leftarrow D$ and $S \rightarrow Y$. This model is trained to learn the causal effect from semantic-relevant words to the semantic and then to the prediction results. To train this model, we use the cross-entropy loss function as shown in

$$L_x = \sum_c y_c \log(\hat{y}_{d,s,c}) \tag{4}$$

where y_c and $\hat{y}_{d,s,c}$ means the ground-truth and prediction probability on the text category c.

Another model $\hat{y}_s$ is used to estimate the causal effect in $S \rightarrow Y$. In this model, we preserve the stereotype words **S** and semantic-irrelevant words **U** and mask the normal words **D** as the model input. The output of the model is still the corresponding text category predictions. Then we optimize the second model by the following loss function:

$$L_s = \sum_c y_c \log(\hat{y}_{s,c}) \tag{5}$$

where $\hat{y}_{s,c}$ means the prediction probability of the word-based classification model on the text category c. Moreover considering that the stereotype influence from the pre-trained model is intrinsic and does not affect the training process, we assume that the overall prediction is a linear combination of stereotype influ-

ence $\hat{y}_m$ and the fusion of $\hat{y}_{x(d,s)}$ and $\hat{y}_s$:

$$\begin{aligned} P(Y = y|x(d,s),m,s) &= y(x(d,s),m,s) \\ &= f(\hat{y}_{x(d,s)},\hat{y}_s,\hat{y}_m) \\ &= f(\hat{y}_{x(d,s)},\hat{y}_s) + \hat{y}_m \end{aligned} \tag{6}$$

As for the fusion model of $f(\hat{y}_{x(d,s)},\hat{y}_s)$, we adopt the following fusion strategy to combine the model predictions: $\hat{y}_{x(d,s)} + \alpha\tanh(\hat{y}_s)$, where α is a hyperparameter, σ is sigmoid function, and tanh is the tanh activation function. And the corresponding loss function of this fusion model is set to be:

$$L_f = \sum_c y_c \log(f(\hat{y}_{x(d,s)},\hat{y}_s)) \tag{7}$$

2.4 Unbiased Prediction

The third stage of our framework is to mitigate the stereotype bias from the text category predictions. As we have mentioned in Sect. 3.1, we aim to remove the direct causal effect from the set of stereotype words **S** and pre-trained word embedding model **M**. The total effect (**TE**) stands for all the direct and indirect causal effects of the causal variable on the outcome, which can be denoted as:

$$\text{TE} = P(Y = y|x(d,s),m,s) - P(Y = y|x(d^*,s^*),m^*,s^*) \tag{8}$$

Then the direct causal effect of m and s can be represented by the natural direct effect (**NDE**):

$$\text{NDE} = P(Y = y|x(d^*,s^*),m,s) - P(Y = y|x_{c^*,s^*},m^*,s^*) \tag{9}$$

where d^*,s^*, and m^* represent the counterfactual value of **D**, **S**, and m respectively. Specifically, the counterfactual values d^* and s^* can be obtained by the masked values based on the training dataset and the value of m^* can be set as any value and will not influence the indirect effects shown in (10).

Finally, as shown in Fig. 1 (b), we aim to cut all the direct causal effects from **M** and **S** to the text categories **Y**. Therefore, we use the total indirect effects (**TIE**) to remove the stereotype bias from the pre-trained model and stereotype words:

$$\begin{aligned} \text{TIE} &= \text{TE} - \text{NDE} \\ &= P(Y = y|x(d,s),m,s) - P(Y = y|x(d^*,s^*),m,s) \\ &= y(x(d,s),m,s) - \sigma y(x(d^*,s^*),m,s) \\ &\approx f(\hat{y}_{x(d,s)},\hat{y}_s) - \sigma f(\hat{y}_{x(d^*,s^*)},\hat{y}_s) \end{aligned} \tag{10}$$

where σ is a hyperparameter to control the influence of the stereotype bias on the prediction results.

3 Experiments

3.1 Settings

To validate the effectiveness of our debiasing framework, we conduct experiments on multiple text classification datasets using different classifiers and mitigate the stereotype bias via our proposed framework. Then we compare our results with two state-of-the-art works on bias mitigation [17,22]. In experiments, we concentrate on the effectiveness of our proposed methods on classification results and word-level fairness. The code of the experiments is available[1].

Baseline. We choose three representative text classifiers as the baselines of our framework. The first one is **TextCNN** [5] which is based on the convolutional neural network (CNN) to extract the textual features and the TextCNN requires word embedding as the input, which can utilize pre-trained word embeddings. Another model is **TextRCNN** [8], which uses the bi-directional recurrent networks to capture the contextual information and utilize CNN for future feature extraction. The last one is **RoBERTa** [11], which uses dynamic masking and a larger pre-training set than BERT. RoBERTa can reach better generalization and robustness in text classification tasks. The three models all require pre-trained word embedding as inputs to involve the stereotype bias from the pre-trained model in downstream text classification tasks. To compare with the SOTA debiasing works, we choose two methods: **IPS-Weight** [22] and **CORSAIR** [17] for comparison. IPS-Weight uses the inverse propensity score as the instance weights to reduce the intended bias while CORSAIR removes the bias from the context words and imbalanced classes by removing counterfactual predictions.

Dataset. We conduct experiments on nine text classification datasets. Among these datasets, six datasets are the same as used in [17]: HyperPartisan, Twitter, ARC, SCIERC, Economy, and Parties. In addition, we also use an Amazon product review datasets [23]. We adopt the same pre-processing procedures on these datasets as [17].

Evaluation. We evaluate the framework from two perspectives: classification performance and word-level fairness. Considering the class imbalance of our dataset, we use Macro-F1 to measure the text classification performance. As for the word fairness, we adopt the evaluation framework shown in [19]. For each word in the dataset, we compare the prediction distribution of the data that contains this word with the even distribution and calculate the Jensen-Shannon divergence (JS). We use the average JS of all the words as the fairness metric.

Parameters. Then we use the grid search method to decide the specific value of parameters in the experiments. We set the batch size as 32 for the training and test data and then we use the Adam optimizer with a learning rate of 5e-4 for all three classification models during both the initial training and fusion training stages which have 20 epochs respectively. Then we set the proportion of

[1] https://github.com/DATA-Transpose/StereotypeWords.

the stereotype words as 5%, and set the α in the fusion training stage as 0.1. As for the parameter σ in the unbiased predictions, we search for the best σ from 0 to 2 with a stride of 0.05 based on the validation results. The experiments are conducted on three servers with NVIDIA RTX A5000 GPUs and the results are the average results of three rounds of experiments using different seeds.

3.2 Classification Performance

Table 1 shows our proposed methods' classification performance (Macro F1 score) and the comparison with two SOTA methods. The higher results mean a better classification performance. The rows of **BASELINE** stand for the results without any debiasing methods. The rows of **KEYWORD** and **IMBWORD** represent the results of ablation studies where we regard all the words that have positive contributions to predictions as stereotype words in the experiments of KEYWORD and we mark all the words that have large entropy as stereotype words in the rows of KEYWORD.

From the results, our proposed methods have average improvements of 4.23%, 4.88%, and 4.82% using TextCNN, TextRCNN, RoBERTa from the baselines across the nine datasets. As for the two comparison methods, the improvements of our proposed methods are much more significant and stable across different datasets. Then compared with the results of two ablation studies, the proposed method considering both the word importance and entropy can reach better classification results in most of the datasets and the results are slightly lower than the ablation methods in SCIERC and Parties datasets. Moreover, the results that

Table 1. Classification performances compared with the State-of-the-art methods (%)

Model	Method	HYP	TWI	ARC	SCI	ECO	PRT	AMA
TextCNN	BASELINE	59.63	80.41	38.80	44.25	56.16	57.70	72.71
	IPS	45.48	63.76	13.87	9.79	44.19	57.75	65.60
	CORSAIR	51.20	69.26	17.57	22.06	58.23	55.16	67.33
	KEYWORD	66.12	74.62	48.72	43.13	57.97	57.88	72.84
	IMBWORD	65.95	75.03	41.65	**46.32**	60.08	**58.90**	72.50
	PROPOSED	**66.88**	**82.24**	**52.96**	45.14	**60.67**	57.94	**73.46**
TextRCNN	BASELINE	60.33	69.84	52.08	62.29	56.42	**60.68**	72.45
	IPS	36.89	60.58	11.23	9.79	44.19	52.60	68.57
	CORSAIR	48.88	74.72	22.21	23.03	56.04	57.84	64.19
	KEYWORD	**74.59**	80.08	55.96	62.01	61.91	57.78	**73.50**
	IMBWORD	73.61	77.31	54.24	**63.59**	61.78	59.15	73.28
	PROPOSED	70.55	**82.11**	**58.31**	62.00	**63.51**	58.97	72.83
RoBERTa	BASELINE	65.60	71.18	24.92	26.54	60.36	**64.33**	89.45
	IPS	50.66	70.77	15.88	9.79	44.19	54.12	74.15
	CORSAIR	59.62	68.17	17.17	20.30	57.57	53.64	71.67
	KEYWORD	72.23	74.80	27.52	31.34	64.91	64.25	89.83
	IMBWORD	73.23	76.64	23.35	30.46	**65.46**	63.91	89.67
	PROPOSED	**75.80**	**79.17**	**28.71**	**33.04**	65.45	63.61	**90.33**

Table 2. Word-level fairness of proposed methods

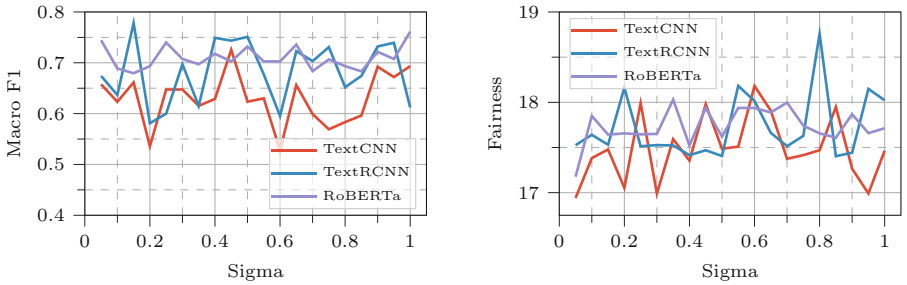
Model	Method	HYP	TWI	ARC	SCI	ECO	PRT	AMA
TextCNN	BASELINE	17.87	19.46	42.35	39.91	18.20	13.80	16.27
	IPS	19.63	20.41	45.18	47.78	21.54	13.14	16.39
	CORSAIR	**16.28**	19.45	40.76	35.67	16.58	**13.00**	**15.82**
	KEYWORD	17.12	18.64	41.73	37.12	19.85	17.41	16.34
	IMBWORD	17.18	**18.27**	41.83	39.37	17.60	16.14	16.16
	PROPOSED	17.21	18.60	**37.00**	**37.11**	**16.31**	13.70	16.44
TextRCNN	BASELINE	19.08	19.11	41.48	37.34	20.00	13.54	16.30
	IPS	**16.05**	19.22	43.32	**35.53**	**15.33**	13.17	16.17
	CORSAIR	16.94	19.14	43.49	37.79	16.97	**13.12**	**16.15**
	KEYWORD	17.50	18.21	41.60	37.23	20.74	20.12	16.22
	IMBWORD	17.48	**16.52**	**41.50**	36.43	20.11	19.88	16.32
	PROPOSED	17.87	18.03	41.89	36.26	20.16	14.45	16.20
RoBERTa	BASELINE	17.60	18.78	43.44	42.75	19.40	14.57	16.41
	IPS	17.43	18.42	42.21	40.36	17.09	14.36	16.48
	CORSAIR	18.58	18.69	46.27	**39.89**	17.36	**13.97**	16.52
	KEYWORD	17.86	18.23	40.81	41.51	**16.93**	14.16	16.34
	IMBWORD	17.73	18.63	**40.70**	41.36	16.93	14.69	16.37
	PROPOSED	**17.17**	**17.90**	41.11	41.47	19.76	14.10	**16.34**

only consider the word importance(KEYWORD) result in a higher Macro F1 score than the results of IMBWORD in all three baseline models, which implies that semantic words are one source of bias in text classification.

3.3 Stereotype Word Fairness

Table 2 shows the word-level fairness of our proposed methods, ablation studies, and two comparison techniques. Considering that the stereotype words are not fixed among different classification models, we calculate the fairness of all the words instead of the stereotype words in the texts in Table 2. Lower results mean better fairness in the prediction results.

Compared with the baselines, our proposed methods can reach average improvements of 1.64, 0.28, and 0.73 respectively across all the datasets. From the results, we can find our proposed methods can reach lower fairness metrics in TextCNN and RoBERTa models in most datasets. The improvements in fairness are not as significant as the improvements in F1 scores because we use the fairness on the whole word set in the documents instead of the stereotype word set. The larger stereotype word set easily leads to a smaller word fairness. Compared with three methods: CORSAIR, KEYWORD, and IMBWORD, our proposed methods select a smaller and more accurate stereotype word set for debiasing and can reach competitive results.



(a) F1 against Sigma

(b) Fairness against Sigma

Fig. 3. F1 and Fairness results under different proportion of stereotype words

3.4 Proportion of Stereotype Words

In this section, we implement further experiments on the influences of stereotype word proportion in the document set. We choose HyperPartisa and try 20 different proportions from 5%, 10%, to 100%. The Macro F1 scores and fairness are recorded in Fig. 3. For the TextRCNN model, the proportion of 15% can reach the highest F1 scores of more than 0.78 and rather low fairness around 17.00 while the best stereotype word proportion for the TextCNN model is around 50%, where the classification F1 is the highest: 0.73 and the word fairness is about 17.50. In addition, for the RoBERTa model, the F1 score can reach 0.75 when we set the proportion as 5%. In the meantime, the fairness is around 17.17 under this proportion. The different proportions show the effects of pre-trained embedding models and classification models on the stereotype bias. Similar to HyperPartisa, in other datasets, we can also select the best stereotype word proportions that can retain higher F1 scores and lower word fairness.

4 Conclusion

In this work, we follow previous works and focus on the potential word-level stereotype bias in text classification. We analyse the generation of bias in a causal view and propose a novel framework for bias mitigation. Our framework includes stereotype detection, fusion model training and unbiased prediction. Different from previous works, our framework can detect words that have a direct contribution to the predictions and does not rely on an external thesaurus. The experiments show better and more stable performances on multiple datasets and the ablation studies prove the effectiveness of the two parts in stereotype word detection. Moreover, we also explore the influences of the proportions of selected stereotype words. In future work, we will refine and weaken our assumptions on the proposed causal graph. We will include the causal path from the pre-trained model to the semantic variables and model the corresponding causal effects.

References

1. Badjatiya, P., Gupta, M., Varma, V.: Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In: WWW (2019)
2. Frassanito, P., Pettorini, B.: Pink and blue: the color of gender. Childs Nerv. Syst. **24**, 881–882 (2008)
3. Huang, X.: Easy adaptation to mitigate gender bias in multilingual text classification. In: NAACL (2022)
4. Jia, S., Meng, T., Zhao, J., Chang, K.: Mitigating gender bias amplification in distribution by posterior regularization. In: ACL (2020)
5. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP (2014)
6. Kiritchenko, S., Mohammad, S.M.: Examining gender and race bias in two hundred sentiment analysis systems. In: NAACL-HLT (2018)
7. Kokalj, E., Skrlj, B., Lavrac, N., Pollak, S., Robnik-Sikonja, M.: BERT meets shapley: extending SHAP explanations to transformer-based classifiers. In: EACL (2021)
8. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: AAAI (2015)
9. Liu, H., Jin, W., Karimi, H., Liu, Z., Tang, J.: The authors matter: Understanding and mitigating implicit bias in deep text classification. arXiv preprint arXiv:2105.02778 (2021)
10. Liu, J., et al.: Fair representation learning: An alternative to mutual information. In: Zhang, A., Rangwala, H. (eds.) KDD (2022)
11. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
12. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NIPS (2017)
13. Mihalcea, R., Tarau, P.: Textrank: bringing order into text. In: EMNLP (2004)
14. Nadeem, M., Bethke, A., Reddy, S.: Stereoset: measuring stereotypical bias in pretrained language models. In: ACL/IJCNLP (2021)
15. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: K-CAP (2003)
16. Pearl, J.: Direct and indirect effects. In: Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 373–392 (2022)
17. Qian, C., Feng, F., Wen, L., Ma, C., Xie, P.: Counterfactual inference for text classification debiasing. In: ACL/IJCNLP (2021)
18. Sun, P., Wu, B., Li, X., Li, W., Duan, L., Gan, C.: Counterfactual debiasing inference for compositional action recognition. In: MM (2021)
19. Sweeney, C., Najafian, M.: A transparent framework for evaluating unintended demographic bias in word embeddings. In: ACL (2019)
20. Tian, B., Cao, Y., Zhang, Y., Xing, C.: Debiasing NLU models via causal intervention and counterfactual reasoning. In: AAAI (2022)
21. Wu, J., Liu, Q., Xu, W., Wu, S.: Bias mitigation for evidence-aware fake news detection by causal intervention. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR (2022)
22. Zhang, G., Bai, B., Zhang, J., Bai, K., Zhu, C., Zhao, T.: Demographics should not be the reason of toxicity: mitigating discrimination in text classifications with instance weighting. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) ACL (2020)

23. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. In: NIPS (2015)
24. Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.: Gender bias in contextualized word embeddings. In: NAACL-HLT (2019)
25. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: NAACL-HLT (2018)