Causal-Debias: Unifying Debiasing in Pretrained Language Models and Fine-tuning via Causal Invariant Learning

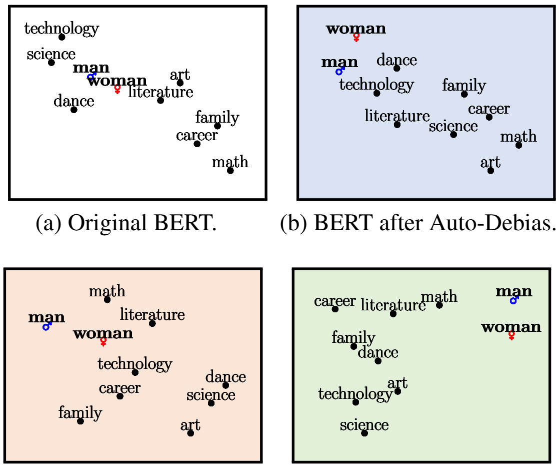
Fan Zhou[1] Yuzhou Mao [1] Liu Yu [1] * Yi Yang [2] Ting Zhong [1,3]
[1]University of Electronic Science and Technology of China
[2]Hong Kong University of Science and Technology
[3]Kashi Institute of Electronics and Information Industry
fan.zhou@uestc.edu.cn, yuzhou.mao@outlook.com, liu.yu@std.uestc.edu.cn,
imyiyang@ust.hk, zhongting@uestc.edu.cn

Abstract

Demographic biases and social stereotypes are common in pretrained language models (PLMs), and a burgeoning body of literature focuses on removing the unwanted stereotypical associations from PLMs. However, when fine-tuning these bias-mitigated PLMs in downstream natural language processing (NLP) applications, such as sentiment classification, the unwanted stereotypical associations resurface or even get amplified. Since pretrain&fine-tune is a major paradigm in NLP applications, separating the debiasing procedure of PLMs from fine-tuning would eventually harm the actual downstream utility. In this paper, we propose a unified debiasing framework Causal-Debias to remove unwanted stereotypical associations in PLMs *during* fine-tuning. Specifically, Causal-Debias mitigates bias from a causal invariant perspective by leveraging the specific downstream task to identify bias-relevant and label-relevant factors. We propose that bias-relevant factors are non-causal as they should have little impact on downstream tasks, while label-relevant factors are causal. We perform interventions on non-causal factors in different demographic groups and design an invariant risk minimization loss to mitigate bias while maintaining task performance. Experimental results on three downstream tasks show that our proposed method can remarkably reduce unwanted stereotypical associations after PLMs are fine-tuned, while simultaneously minimizing the impact on PLMs and downstream applications.

1 Introduction

Pretrained language models (PLMs) have achieved remarkable success in many natural language processing (NLP) tasks. However, PLMs often encoded undesired social stereotypes and biases, and thus mitigating such biases has become an emerging and important task (Meade et al., 2022). Prior bias mitigation methods often focus on removing unwanted stereotypical associations in PLMs. For example, some works (Zmigrod et al., 2019) pretrain a language model using original and counterfactual corpus in order to cancel-out biased associations, some works (Liang et al., 2020) focus on debiasing post-hoc sentence representations, and others (Guo et al., 2022; Cheng et al., 2021) design bias-equalizing objectives to fine-tune PLM's parameters.



(a) Original BERT. (b) BERT after Auto-Debias.

(c) Auto-Debias on SST-2. (d) Ours on SST-2.

Figure 1: Motivation of Causal-Debias. t-SNE plots of average sentence representations of each word across its sentence templates on SST-2 task (Socher et al., 2013).

However, a problem with existing debiasing strategies is that they are separate from downstream NLP tasks. If people take a debiased PLM which is supposed to have certain stereotypical associations removed, and then fine-tune it on downstream task, the unwanted associations will re-enter or even get amplified in the fine-tuned language model (Goldfarb-Tarrant et al., 2021). Consider gender debiasing using Auto-Debias (Guo et al., 2022) as an example. Fig. 1a and 1b shows the BERT-based sentence embeddings using t-SNE (Van der Maaten and Hinton, 2008) before and after Auto-Debias, respectively. The gender bias is clearly less

[*]Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4227–4241
July 9-14, 2023 ©2023 Association for Computational Linguistics

prominent than the original BERT, i.e., non-gender-specific concepts (in black) are more equidistant to both genders after Auto-Debias (Fig. 1b). However, when applying BERT model that is debiased by Auto-Debias to downstream tasks (Fig. 1c), the fine-tuning procedure almost, if not all, “neutralizes" the effect of PLM debiasing. The phenomenon that biases encode in a fine-tuned PLM is as worrisome as in a vanilla PLM, because the fine-tuned PLM is more likely to be deployed in real-world scenarios serving thousands to millions of end users.

Addressing the aforementioned bias resurgence issues is non-trivial. On the one hand, existing literature on bias mitigation mostly treats PLM debiasing as a standalone problem which is separate from downstream tasks. Incorporating debiasing objectives into the fine-tuning procedure can be a viable solution but it is still less explored. On the other hand, one may expect to use existing debiasing methods to re-debias the fine-tuned PLM. However, due to catastrophic forgetting concern (Kirkpatrick et al., 2017), a sequential combination of fine-tuning and debiasing may worsen the downstream task performance (Goodfellow et al., 2013). Therefore, there is a research gap in unifying debiasing in PLMs and fine-tuning for building fair and accountable NLP services.

In this work, we propose **Causal-Debias**, a Causal Invariant Debiasing Model to unify the debiasing with downstream fine-tuning. In principle, we analyze the cause and propagation of biases and introduce the Structure Causal Model (SCM) (Pearl et al., 2000, 2016) to address the bias mitigation problem by exploiting the inherent causal mechanism in the downstream datasets. Specifically, Causal-Debias first exploits a causal intervention module to distinguish causal and non-causal factors. It then generates counterfactual sentences which have different non-causal factors but the same semantic meanings. The generated counterfactual sentences, along with the original sentences, are fed into an invariant optimization function to ensure a trade-off between the performance of downstream tasks and the effectiveness of debiasing. As illustrated in Fig. 1(d), Causal-Debias can preserve the PLM debiasing effect even after fine-tuning on the downstream dataset.

We evaluate performance of Causal-Debias in mitigating the gender and racial biases in several popular PLMs, e.g., BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and RoBERTa (Liu et al., 2019), on three GLUE (Wang et al., 2018) tasks (SST-2, CoLA, and QNLI). The results demonstrate Causal-Debias can significantly mitigate PLM biases after downstream fine-tuning while also maintaining the downstream task performance. We hope this work provides empirical evidence that stereotypical associations can reenter language models during the fine-tuning step. Moreover, we hope that debiasing with a causal perspective offers a more generalizable and reliable way for building fair and accountable NLP applications. We release the anonymous implementation of Causal-Debias at https://github.com/myZeratul/Causal-Debias.

2 Related Works

PLM Debiasing aims to remove biases, quantified as unwanted stereotypical associations, from pretrained language models. Existing works on PLM debiasing can be categorized into two lines based on whether downstream tasks are involved in the debiasing pipeline. (1) *Non-Task-Specific*: Counterfactual Data Augmentation (**CDA**) (Zmigrod et al., 2019) and **Dropout** (Webster et al., 2020) are two methods where debiasing happens in the pre-training stage (Meade et al., 2022). **Auto-Debias** (Guo et al., 2022), **Context-Debias** (Kaneko and Bollegala, 2021) and **MABEL** (He et al., 2022) remove biases in PLM by designing different bias-equalizing objectives. In this line of work, the parameters in the PLM are changed to meet the fairness criteria such as SEAT (May et al., 2019), and the ultimate goal is to remove bias associations from PLM, so that downstream tasks can benefit from the debiased models. However, we show that it is not this case. When the debiased models are fine-tuned on downstream tasks, the bias associations resurge, perhaps because the biases are not completely removed or maybe just covered up, or because downstream datasets are encoded with stereotypical associations. (2) *Task-Specific*: Existing works on this line, including **Sent-Debias** (Liang et al., 2020) and **FairFil** (Cheng et al., 2021), keep the parameters of PLMs untouched. Instead, they target on the sentence representations, and aim to remove bias associations from representations. Even though the sentence representations, which are the input for downstream tasks, are refined, downstream fine-tuning can still introduce new biases, as we show in experiments. In summary, our work differs from existing PLM debiasing in that we aim to unify fine-tuning procedure with debiasing so that the fine-tuned models are free from bias associations.

In addition, prior literature also challenges the effectiveness of bias mitigation. Gonen and Goldberg (2019) showed that debiasing methods only cover-up biases in word embeddings but do not remove them. Meade et al. (2022) found that existing debiasing methods for PLMs hurt the language modeling capability of PLMs and thus the practical utility of debiasing warrants attention. Our work follows the spirit of this line of works in that we empirically demonstrate the bias resurgence problem in fine-tuning and suggest to development debiasing techniques with downstream utility.

Causal Mechanism is crystallized in the *invariant learning*, suggesting that only associations invariant in the training set should be learned (Peters et al., 2016; Muandet et al., 2013). *Invariant Risk Minimization* (IRM) (Arjovsky et al., 2019) is a practical implementation of invariant learning, which is an optimization objective modifying the loss with a regularization term to enforce the invariant representations. Recent works have explored Structural Causal Model (SCM) (Schölkopf et al., 2012) to model auxiliary variables and show promising performance, ranging from domain generalization (Lv et al., 2022) in computer vision and intrinsic interpretability (Wu et al., 2022) in graph neural networks to factual knowledge (Li et al., 2022) and text classification (Qian et al., 2021) in NLP. In this work, we introduce SCM to PLM debiasing to discover the inherent causal relations between data and labels while achieving better debiasing performance. To our knowledge, this is the first work exploiting causality for debiasing PLMs.

3 Methodology

Our goal is to unify debiasing with downstream fine-tuning so that the fine-tuned language model can maintain solid performance with alleviated stereotypical associations. In other words, we want to prevent biases re-entering the language model during the fine-tuning. Unlike prior work that considers PLM debiasing as a standalone procedure, we mitigate language model biases *during* fine-tuning process. To this end, we propose Causal-Debias, a debiasing framework from a causal view. We first provide the basic formalism and proceed with the details of Causal-Debias.

Problem Definition: We denote a supervised NLP task with dataset (X, Y), we fine tune a pretrained language model to learn a mapping: $\mathcal{M}(X) \mapsto Y$. Our goal is to mitigate unwanted stereotypical associations in the fine-tuned model $\mathcal{M}$.

3.1 Biases from a Causal View

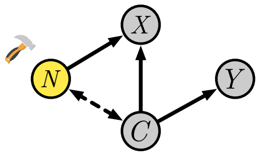


Figure 2: SCM of Causal-Debias. Each raw sentence of X is generated by a mix of causal factor C and non-causal factor N. Note that only the causal factor affects the ground truth label Y, while the **hammer** indicates the intervention on non-causal factor.

We use a Structure Causal Model (SCM) to characterize biases in the fine-tuning procedure. As shown in Fig. 2, there are four variables: input raw sentence X, downstream ground-truth label Y, causal factor C and non-causal factor N. Among those, causal and non-causal factor C and N are latent variables. Whether a factor is causal or non-causal depends on the specific downstream tasks. For example, in sentiment classification task, causal factors could be adjective sentiment words such as good or bad, and non-causal factors could be nouns or pronouns. In contrast, in coreference resolution task, pronouns could be causal factors while adjective words could be non-causal. Here we explain the diagram in detail:

- $C \rightarrow X \leftarrow N$. The input raw sentence X is a mix of two factors that are theoretically non-intersecting: causal factor C and non-causal factor N.
- $C \rightarrow Y$. From a causal view, the ground-truth label Y is only determined by causal factor C.
- $C \dashleftarrow\dashrightarrow N$. The dashed arrow delegates additional probabilistic dependencies (Pearl et al., 2016, 2000) between causal factor C and non-causal factor N – *cf.* Appendix A for examples.

However, $N \dashleftarrow\dashrightarrow C \rightarrow Y$ can create a spurious association between non-causal factor N and ground-truth label Y (denoted as $Y \not\perp\!\!\!\perp N$), so that C becomes a confounder between N and Y which opens a backdoor path $N \leftarrow C \rightarrow Y$. Hence, the unwanted stereotypical associations, which we assume are non-causal factors, may re-enter the fine-tuned PLM because of the unwanted associations between label Y and non-causal factors N.

To mitigate the bias propagation and avoid such spurious association that N affects Y through C, we make the feature induction assumption spurred by the Independent Causal Mechanisms (ICM) principle (Peters et al., 2017; Schölkopf et al., 2012), i.e., the intervention on non-causal factor N should be independent of the ground truth Y. According to causal inference (Pearl et al., 2016, 2000), there is a directed link from each of its parent variables $P(X)$ to X, if and only if the causal mechanism $X = f_X(P(X))$ exists for each variable X. That is, there exists a function from causal factor C to label Y without N's influence, making the causal association $C \rightarrow Y$ invariant across different N. More formal assumptions in terms of $N \dashleftarrow\dashrightarrow C$ are presented in Appendix A. In other words, ICM principle can assure the language model to not pick up the non-causal factors, i.e., unwanted associations, in the fine-tuning procedure.

Generally, only X and Y are observed during the fine-tuning, while neither the causal factor C nor the mapping from C to Y is available. We factorize the language model as a combination of two modules inspired by (Wu et al., 2022), i.e., $\mathcal{M} = m_Y \circ m_C$, where $m_C : X \rightarrow C$ discovers the causal factor from the observed X, and $m_Y : C \rightarrow Y$ outputs the prediction Y. Empirical risk minimization (ERM) is often used as the optimization strategy to train m_C and m_Y (Seo et al., 2022; Zhou et al., 2022):

$$\min_{m_C, m_Y} \mathcal{R}(m_Y \circ m_C(X), Y), \tag{1}$$

where $\mathcal{R}(\cdot, \cdot)$ can be any loss function (e.g., cross-entropy loss). However, ERM heavily relies on the statistical dependency between the input sentences and labels, ignoring the critical condition $Y \perp\!\!\!\perp N \mid C$ to guarantee the invariant causal association $C \rightarrow Y$ across different N, leading to the resurgence of undesired spurious associations.

Recent causal learning literature has proposed to use invariant risk minimization (IRM) objective to replace ERM (Arjovsky et al., 2019; Chang et al., 2020). The causal invariant learning encourages causal factor C to seek the patterns that are stable across different "enviornments", while abandoning the unstable non-causal patterns. We follow this line of literature, and propose to use causal invariant learning objective to mitigate fine-tuning biases:

$$\min_{m_C, m_Y} \mathcal{R}(m_Y \circ m_C(X), Y), \textit{s.t. } Y \perp\!\!\!\perp N \mid C, \tag{2}$$

where $N = X \setminus C$ is the non-causal factor. However, IRM often leverages multiple "enviornments" to facilitate causal learning. For example, Peyrard et al. (2021) use prior knowledge to partition the training set to form different environments. In NLP tasks such as sentiment classification, how to construct multiple "enviornments" in the context of language model fine-tuning is less studied. Next, we propose a causal intervention method to construct "enviornments" for causal learning.

3.2 Causal-Debias

Causal Intervention. The high-level idea for causal intervention is to create interventional distributions with respect to different demographic groups. Our interventional distribution is obtained by augmenting and expanding the original data distribution. First, let $\mathcal{W}_a$ and $\mathcal{W}_t$ denote *attribute* words and *target* words, respectively. In the case of gender bias, for instance, target words consist of gender-neutral words (e.g., *nurse, engineer, professor*), and attribute words are composed of the feminine (e.g., *she, woman, mother*) and masculine words (e.g. *he, man, father*) (Liang et al., 2020). Then we can obtain an augmented datasets X_d:

$$X_d = X_o \cup X_c, \tag{3}$$

where X_o denotes the original sentences from the downstream dataset containing any $\mathcal{W}_a$ or $\mathcal{W}_t$, and X_c represents the counterfactual sentences via performing attribute word counterfactual augmentation on X_o. For counterfactual augmentation, we create counterfactual sentences by replacing the attribute word to its corresponding pair (e.g., he->she). However, the augmented dataset X_d still has limitations as it is not sufficient to cover the diversity of demographic groups, which may cause debiasing performance degradation problems (*cf.* Section 4.3 for empirical study). To create sufficiently complex interventional distributions and obtain the most different demographic groups, we conduct the interventions by doing semantic matching between X_d with external corpora E, expanding X_o and X_c to $X_{\tilde{o}}$ and $X_{\tilde{c}}$ (see Table 1 for an example), respectively, as

$$X_{\tilde{d}} = X_{\tilde{o}} \cup X_{\tilde{c}} = \text{Top}_k(\text{sim}(X_d, E)) \cup X_d, \tag{4}$$

where $\text{sim}(\cdot)$ denotes the cosine similarity for semantic matching, and $\text{Top}_k(\cdot)$ selects the top-k semantic similar sentences. After obtaining the

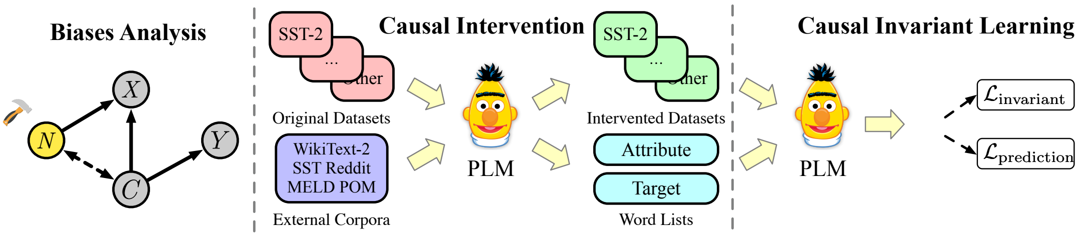


Figure 3: Overview of Causal-Debias. **Causal Intervention**: selecting top k semantic similar bias-related sentences from external corpora to cover the most different demographic groups. **Causal Invariant Learning**: fine-tuning the PLM with an invariant loss among different environments.

intervened sentences, we reconstruct the interventional distribution by combining $X_{\tilde{d}}$ and the rest bias-unrelated downstream dataset, both of which are applied for causal invariant learning. Table 1 shows an example of original sentence, and its corresponding counterfactual and expansion sentences.

Sentence Type	**Sentence Example**
Original	proves once again he hasn't lost his touch
Counterfactual	proves once again she hasn't lost her touch
Expansion	sachs is a guy with impressive intelligence and passion

Table 1: An example sentence in SST-2 dataset, and the corresponding counterfactual and expansion. The label is positive for all three sentences.

Causal Invariant Learning. Once we obtain the intervened dataset containing original and interventional data distributions, we proceed with causal invariant learning. Specifically, we do n-intervention $do(N = n)$, which removes all links from their parents $P(N)$ to the variable N while fixing N to the number of demographic n (e.g. $n = 2$ in the case of gender), to identify C whose relationship with Y is stable across different distributions:

$$\min \mathcal{L}_{\text{invariant}} = \mathbb{E}_n(\mathcal{R}) + \text{Var}_n(\mathcal{R}), \qquad (5)$$

where $\mathcal{R} = \mathcal{R}(\mathcal{M}(X), Y \mid do(N = n))$ computes the risk under the n-interventional distribution; $\mathbb{E}_n(\cdot)$ denotes the risks of different n-interventional distributions; Var$(\cdot)$ denotes the variance of risks over n-interventional distributions.

To calculate $\mathcal{L}_{\text{invariant}}$, the PLM is required to predict the same results on the sentences $X_{\tilde{o}}$ and $X_{\tilde{c}}$, which have equivalent semantics but different attribute words according to the IRM theory (Arjovsky et al., 2019). Thus in optimization, we have the interventional risk derived from Equation (5):

$$\mathcal{R}(\mathcal{M}(X_{\tilde{d}}), Y \mid do(N = \tilde{n})) = \mathbb{E}_{C=m_C(x), N=\tilde{n}} l(\tilde{y}, y), \qquad (6)$$

where $x \in X_{\tilde{d}}$ is a sentence instance with its prediction $\tilde{y}$ under the intervention $do(N = \tilde{n})$, and $y \in Y$ is ground-truth label, and $l(\cdot)$ denotes interventional loss function on a single sentence. Here we choose Wasserstein distance (Ramdas et al., 2017) as a loss function due to its ability to measure the agreement between the prediction of original and post-intervention sentences. The Wasserstein distance between $\tilde{y}$ and y is formalized as below:

$$D_{\text{Wasser}}(\tilde{y}, y) = \inf_{\gamma(\tilde{y}, y) \in \prod} \mathcal{E}_{(\tilde{y}, y) \sim \gamma} ||\tilde{y} - y||, \qquad (7)$$

PLM Fine-tuning. The above procedure leverages augmented datasets for causal invariant learning, and we can incorporate the invariant loss with specific downstream tasks to fine-tune a language model. This way, we can balance the trade-off between debiasing performance and downstream task performance (Meade et al., 2022). The overall objective of Causal-Debias is:

$$\min_{\tau} \mathcal{L}_{\text{prediction}} + \tau \mathcal{L}_{\text{invariant}}, \qquad (8)$$

where τ is the trade-off coefficiency, and $\mathcal{L}_{\text{prediction}}$ is the loss function of a specific downstream task, such as cross-entropy loss for classification and mean squared error loss for regression.

4 Experiments

4.1 Experimental Settings

Debiasing Benchmarks. We compare Causal-Debias with the following benchmarks. *Non-Task-Specific* methods including: **CDA**, **Dropout** (Webster et al., 2020), **Context-Debias** (Kaneko and Bollegala, 2021), **Auto-Debias** (Guo et al., 2022), and **MABEL** (He et al., 2022), and two *Task-Specific* methods including **Sent-Debias** (Liang et al., 2020) and **FairFil** (Cheng et al., 2021). In the *Non-Task-Specific* benchmarks, the debiasing stage is independent of fine-tuning downstream tasks.

Causal-Debias belongs to *Task-Specific* methods as downstream fine-tuning tasks are involved.

Pretrained Language Models. We use three representative PLMs as the backbone: BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and RoBERTa (Liu et al., 2019). Following (Guo et al., 2022), we implement them using the Huggingface Transformers library (Wolf et al., 2020).

Bias Word Lists. Following previous studies, we use human-being-created stereotype/attribute word lists to investigate and mitigate biases in PLMs. They are based on the research of the social science literature and other disciplines, which can reflect cultural or offensive biases. In particular, we consider the gender and race word lists used in (Kaneko and Bollegala, 2021) and (Manzini et al., 2019), respectively – *cf.* Appendix B for details.

External Corpora. For fair comparison, we exploit the same external corpora used in baselines, which are composed of 183,060 sentences from following sources: WikiText-2 (Merity et al., 2017), Standford Sentimente Treebank (Socher et al., 2013), Reddit, MELD (Poria et al., 2019) and POM (Park et al., 2014) – *cf.* Appendix C for details.

Evaluating Metrics: We evaluate biases in PLM embeddings with **SEAT** (May et al., 2019) and CrowS-Pair (Nangia et al., 2020). An ideally unbiased model should exhibit no difference in relative similarity. Following Guo et al. (2022); Liang et al. (2020); Kaneko and Bollegala (2021), we apply SEAT 6, 6b, 7, 7b, 8, and 8b tests to measure the gender bias, and use SEAT 3, 3b, 4, 5, 5b tests for racial bias evaluation. We report the effect size in the SEAT evaluation – the closer to 0, the lower bias a model has. More details about SEAT tests are presented in Appendix D. We also use Crowdsourced Stereotype Pairs (**CrowS-Pair**) (Nangia et al., 2020) as another metric to evaluate gender bias. CrowS-Pair is a dataset containing 1,508 examples covering different types of biases, where each example is a stereotype/anti-stereotype sentence pair with minimal semantics. The CrowS-Pair score closer to 50% is less stereotypical, indicating that the model assigns an equal probability to male and female sentences.

Other Details. As the related studies (Cheng et al., 2021; Liang et al., 2020), we conduct experiments on three downstream tasks, including a sentiment classification task **SST-2**, a grammatical acceptability judgment task **CoLA**, and a question-answering task **QNLI**. We follow the same PLMs as benchmarks: 1) BERT-base-uncased, ALBERT-large-v2, and RoBERTa-base in gender case; and 2) BERT-base-uncased and ALBERT-base-v2 in racial case. We trained Causal-Debias in 5 epochs with learning rate $2 \times e^{-5}$. The reported results are the average of 5 runs for all downstream tasks.

4.2 Results on Mitigating Gender Bias

SEAT Tests and Downstream Tasks Evaluation. Table 2 summarizes the debiasing results of models before and after fine-tuning on three downstream tasks , as well as the accuracy (Acc.) evaluations on downstream applications, from which we have the following **O**bservations.

(O1): Causal-Debias is more effective in mitigating gender bias than previous benchmarks, as it achieves the lowest average SEAT scores in all three downstream tasks. For example, Causal-Debias surpasses *Task-Specific* SOTA (FairFil) by 0.07, 0.01, and 0.07, and *Non-Task-Specific* SOTA (Auto-Debias) by 0.27, 0.21, and 0.09 on BERT, respectively. The excellent debiasing results can attribute to the following two characteristics of Causal-Debias: 1) combining debiasing PLMs and fine-tuning to avoid new biases, and 2) the causal intervention-based invariant learning that alleviates the impact of non-causal factors.

(O2): From the SEAT results after fine-tuning downstream tasks, the PLMs become more biased for almost all non-task-specific debiasing models. In particular, the latest debiasing methods Auto-Debias is greatly limited to the bias resurgence issue in all three tasks, although it achieved good performance on debiasing PLMs. These results verified not only the existence of bias resurgence but also the motivation of this study to attenuate the intrinsic bias of PLMs and fine-tuning bias jointly. Interestingly, the original PLMs do not suffer from this problem, as the bias can be mitigated after fine-tuning the downstream tasks in most cases. This result suggests that fine-tuning itself is an effective way of debiasing PLMs; combining with debiased models, however, will introduce extra bias. To deal with this dilemma, debiasing models should consider downstream tasks as a unity.

Two task-specific models, i.e., Sent-Debias and FairFil, cannot effectively alleviate the application bias even using downstream datasets for debiasing PLMs, because their focus is still intrinsic bias – by contrast, Causal-Debias unifies the two debiasing procedure and provides an systematic solution to

Methods	Before	SST-2		CoLA		QNLI	
		After	Acc.	After	Mcc.	After	Acc.
BERT	0.35	0.29 ↓0.06	92.7	0.18 ↓0.17	57.6	0.37 ↑0.02	91.3
+CDA	0.25	0.47 ↑0.22	81.3	0.29 ↑0.04	53.2	0.38 ↑0.13	89.1
+Dropout	0.42	0.48 ↑0.06	81.9	0.27 ↓0.15	52.2	0.44 ↑0.02	90.1
+Context-Debias	0.53	0.43 ↓0.10	91.9	0.57 ↑0.04	55.4	0.56 ↑0.03	89.9
+Auto-Debias	0.14	0.38 ↑0.24	92.1	0.32 ↑0.18	52.9	0.24 ↑0.10	91.1
+MABEL	0.50	0.55 ↑0.05	92.2	0.52 ↑0.02	57.8	0.54 ↑0.04	91.6
+Sent-Debias	0.26	0.21 ↓0.05	89.1	0.22 ↓0.04	55.4	0.32 ↑0.06	90.6
+FairFil	0.15	0.18 ↑0.03	91.6	0.12 ↓0.03	56.5	0.22 ↑0.07	90.8
+Causal-Debias (ours)	-	**0.11**	**92.9**	**0.11**	**58.1**	**0.15**	**91.6**
ALBERT	0.28	0.22 ↓0.06	92.6	0.24 ↓0.04	**58.5**	0.21 ↓0.07	91.3
+CDA	0.30	0.38 ↑0.18	92.4	**0.16** ↓0.14	53.1	0.31 ↑0.01	90.9
+Dropout	0.24	0.28 ↑0.04	90.4	0.25 ↑0.01	47.4	0.20 ↓0.04	**91.7**
+Context-Debias	0.33	0.11 ↓0.22	77.3	0.17 ↓0.16	55.4	0.20 ↓0.13	91.6
+Causal-Debias (ours)	-	**0.06**	**92.9**	**0.16**	57.1	**0.09**	91.6
RoBERTa	0.67	0.41 ↓0.26	**94.8**	0.41 ↓0.26	**57.6**	0.48 ↓0.19	92.8
+Context-Debias	1.09	0.26 ↓0.83	80.3	0.30 ↓0.79	55.4	0.37 ↓0.72	91.8
+Causal-Debias (ours)	-	**0.09**	93.9	**0.17**	54.1	**0.06**	**92.9**

Table 2: Gender debiasing results of average SEAT and application performance. ↓ and ↑ denote the improvement and reduction in debiasing performance in terms of SEAT, respectively. "-'" means that our model does not have an independent PLM debiasing process.

attenuates both biases simultaneously.

(O3): Previous benchmarks may significantly degrade the performance on downstream tasks after debiasing the PLMs. This is a natural result of existing debiasing models as they need to change the representations to mitigate bias and therefore inevitably decrease accuracy after debiasing (Liang et al., 2020). In addition to its superior debiasing effect, Causal-Debias achieves better performance on downstream tasks, e.g., exceeding all benchmarks on BERT in terms of performance. This result demonstrates the ability of Causal-Debias to minimize the disagreements among different demographic groups with identical semantic information, which is attained by intervening in downstream datasets to mitigate the bias recurrence.

Methods	Overall Score	
	Before (Dev.)	After (Dev.)
BERT	57.25 (7.25)	53.18 (3.18)
+CDA	56.11 (6.11)	58.42 (8.42)
+Dropout	55.34 (5.34)	44.56 (5.44)
+Context-Debias	58.01 (8.01)	58.89 (8.89)
+Auto-Debias	54.92 (4.92)	44.96 (5.04)
+MABEL	**50.76 (0.76)**	46.75 (3.25)
+Sent-Debias	52.29 (2.29)	55.04 (5.04)
+Causal-Debias (ours)	-	**48.94 (1.06)**

Table 3: CrowS-Pairs scores on SST-2. 'Dev.' denotes the value deviates from 50.

CrowS-Pairs. Table 3 reports CrowS-Pairs scores before and after fine-tuning on SST-2. Note that the deviation from 50 is usually used to measure the debiasing effect. Obviously, Causal-Debias achieves the lowest deviation among the fine-tuned debiasing models, which proves its best debiasing performance. In addition, the deviation values of previous methods increase after fine-tuning, which further confirms the existence of application bias ignored by existing methods. Again, BERT itself can alleviate the bias after fine-tuning, which is consistent with our observation **O2** from Table 2.

4.3 Ablation Study

To quantify the effect of the devised causal intervention and invariant risk learning, we build three variants of Causal-Debias:

- (V1) `w/o` E removes the intervention from external corpora in Eq. (4).
- (V2) `w/o` $E\&\mathcal{L}_{\text{invariant}}$ removes both interventions of external corpora in Eq. (4) and the invariant loss in Eq. (8) but maintains counterfactual augmentation of data in downstream tasks.
- (V3) `w/o` $\mathcal{L}_{\text{invariant}}$ only remains the prediction loss $\mathcal{L}_{\text{prediction}}$ in Eq. (8).

Fig. 4 compares the variants with full Causal-Debias on SST-2 task. The variant V1 performs worst on debiasing, indicating that external corpora contribute the most to model debiasing. However, external corpora also deteriorate the task perfor-

Methods	Before	SST-2		CoLA		QNLI	
		After	Acc.	After	Mcc.	After	Acc.
BERT	0.23	0.30 ↑0.07	92.7	0.16 ↓0.07	57.6	0.15 ↓0.08	91.3
+AUTO-DEBIAS	0.18	0.31 ↑0.13	92.1	0.20 ↑0.02	**59.6**	0.24 ↑0.06	91.1
+CAUSAL-DEBIAS (ours)	-	**0.11**	**92.9**	**0.06**	57.1	**0.11**	**91.6**
ALBERT	0.46	0.29 ↓0.06	**92.6**	0.19 ↓0.27	58.5	0.10 ↓0.36	92.2
+AUTO-DEBIAS	0.17	0.39 ↑0.22	86.8	0.18 ↑0.01	56.9	0.36 ↑0.09	91.1
+CAUSAL-DEBIAS (ours)	-	**0.13**	91.9	**0.16**	**59.6**	**0.01**	**92.5**

Table 4: Race debiasing results of average SEAT and application performance, and the original scores are from Meade et al. (2022).

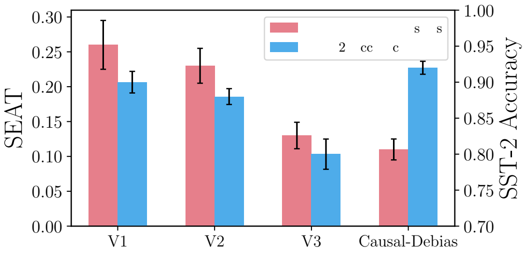


Figure 4: Ablation results. The lower SEAT score (red) and higher SST-2 accuracy (blue) are better.

mance when comparing the accuracy of V2 and V3, because the augmented sentences introduce extra semantics and obfuscate the PLMs. The adverse impact can be minimized by learning the invariant representations by the proposed causal interventions on downstream tasks via $\mathcal{L}_{\text{invariant}}$, which can be justified by the great discrepancy between Causal-Debias and V3 in terms of SST-2 accuracy.

4.4 Results on Mitigating Racial Bias

Racial debiasing refers to examining the association difference between European-American/African American names/terms and the stereotype words (pleasant vs. unpleasant) (Caliskan et al., 2017). Unlike gender debiasing, few prior studies investigated the racial debiasing problem, due to the difficulty of mitigating racial bias (Meade et al., 2022). A critical challenge is the potential word ambiguity (e.g., white, black) in various contexts (Guo et al., 2022). Table 4 reports the performance of Causal-Debias and Auto-Debias – the state-of-the-art racial debiasing model. Causal-Debias substantially decreases the racial biases on PLMs after fine-tuning, while obtaining comparable downstream performance. Auto-Debias, in contrast, still suffers from bias recurrence issue. Compared to Auto-Debias, Causal-Debias is more effective as it exploits downstream datasets for debiasing, which allows us to alleviate the influence of ambiguous words. Besides, the causal invariant learning in Causal-Debias encourages the model to learn consistent representations and clear meanings of ambiguous words so as to avoid bias-related associations.

5 Conclusion

In this paper, we propose a debiasing framework Causal-Debias to unify PLM bias mitigation with fine-tuning. Different from prior literature that treats PLM bias mitigation as a standalone task, Causal-Debias incorporates bias mitigation objective with downstream fine-tuning using causal invariant learning. Causal-Debias essentially “kills two birds with one stone”, because it prevents the unwanted stereotypical associations re-entering the fine-tuned model and also maintains favorable performance in downstream tasks. The experiment shows that fine-tuning existing debiased models will encode or even amplify unwanted bias associations (in gender and race). We also show that Causal-Debias can effectively reduce bias associations in fine-tuned language model without sacrificing downstream task performance. Our paper contributes to the NLP fairness fields by proposing a novel debiasing method from a causal invariant view. More importantly, we highlight the fact that biases can happen at any stage in PLM training pipeline, including the final fine-tuning steps. Even if unwanted stereotypical associations are removed, or covered up (Kirkpatrick et al., 2017), in the pretraining stage, the associations will re-surge in the downstream models. Hence, prior literature that focuses on pretrained model debiasing may be ineffective if used in a pretrain/fine-tune pipeline. We hope this study can shed light on mitigating biases for building fair and accountable NLP systems.

Acknowledgements

This work was supported by the Natural Science Foundation of Sichuan Province, China, under Grant 2022NSFSC0505, and National Natural Science Foundation of China under Grants 62176043 and 62072077.

Limitations

We now discuss limitations of Causal-Debias. In consideration of the fairness, we follow the prior bias mitigation work (Guo et al., 2022; Cheng et al., 2021; He et al., 2022) and use human-collected lists of gender and racial pairs for counterfactual data augmentation and intervened distribution generation. It is obvious that the bias word lists are inadequate to cover all the bias-related demographic groups, while we believe the general list is exhaustive. We consider there is a possible model improvement that leverages the perturbation augmentation on bias-related sentences along multiple demographic axes (Qian et al., 2022). Another possible improvement would be to generate bias words by using prompts to probe the biases that may lead to a bad effect.

Moreover, we also considered the use of external corpora. The external corpora have been significantly investigated in prior works (Liang et al., 2020; Cheng et al., 2021) and are utilized as an intervention corpora. Recently, He et al. (2022) used two natural language inference data (SNLI and MNLI with gender terms) to produce general-purpose debiased representations. There are several other corpora including News-commentary-v1 (Kaneko and Bollegala, 2021), Wikipedia (Zmigrod et al., 2019; Webster et al., 2020), and Wikitext-2 (Guo et al., 2022). A possible future direction of debiasing is how to mitigate the biases without heavily relying on any corpora and just using internal knowledge.

Moreover, in the paper we primarily focus on studying gender and racial bias mitigation. It is also worth exploring intersectional biases mitigation (Lalor et al., 2022) and domain-specific bias mitigation (Chuang and Yang, 2022; Abbasi et al., 2021).

We would also like to note that although Causal-Debias shows a satisfactory performance on SEAT tests and Crows-Pairs, these results should not be interpreted as a complete bias mitigation. Interestingly, He et al. (2022) expressed the same opinion. The main metrics (like CrowS-Pairs) are mainly against North American social biases and only reflect positive predictive power. They detect the presence of the biases but not their absence (Meade et al., 2022). He et al. (2022) did not use the SEAT tests and evaluated their model on various metrics. From the perspective of different usage scenarios, we need a more general and reliable debias metric for comparison between different models. The lack of universality and agreement in existing evaluation frameworks is a fundamental challenge in this field.

Ethics Statement

Regarding ethical concerns, we would like to note that our contributions are mainly about methodologies. The datasets and evaluation metrics in our work are also widely used in prior works. One ethical concern is the binarization of the genders and races, which is an over-simplification and is not proper to practical situations. Binarization is the common problem among most debiasing methods and we totally agree and support the development of more inclusive methodological tools, datasets, and evaluation methods.

Under our framework, we consider gender or race isolatedly and neglect the particular intersectional biases. It is apparent that the pretrained language model cannot be applied or operated in an ideal environment, and should be able to handle complex combinations of biases simultaneously.

Another ethical consideration is that Causal-Debias is entirely based on a English system. Such an assumption may not be a problem now but sooner or later it will be. The debias studies have to be situated on the high-resource languages while considering not only high-resource language systems but how to debias on the low-resource languages.

For instance, some languages such as Spanish, German, Chinese, or Japanese contain various words to describe masculine or feminine forms. The detection and removal of biases are greatly complicated by the need to consider both linguistic and social gender.

For the reasons above, practitioners should be very cautious when applying our framework to real-world use cases. In its current state, Causal-Debias should be seen not as a panacea for addressing biases in NLP, but rather as another initial effort to illuminate and undercut a critical, elusive, multifaceted problem.

References

Ahmed Abbasi, David Dobolyi, John P Lalor, Richard G Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *EMNLP*, pages 3748–3758.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv :1907.02893*.

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

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *ICML*, pages 1448–1458. PMLR.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *ICLR*.

Chengyu Chuang and Yi Yang. 2022. Buy tesla, sell ford: Assessing implicit stock market preference in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 100–105.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *ACL*, pages 1926–1940.

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. *arXiv preprint arXiv:1903.03862*.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *ACL*, pages 1012–1023.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. *arXiv:2210.14975*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *EACL*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Cheng-Jie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. How pre-trained language models capture factual knowledge? a causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1720–1732.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *ACL*, pages 5502–5515.

Paul Pu Liang, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Strong and simple baselines for multimodal utterance embeddings. In *NAACL*, pages 2599–2609.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality inspired representation learning for domain generalization. In *CVPR*, pages 8046–8056.

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.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *NAACL*, pages 622–628.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *ACL*, pages 1878–1898.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *ICLR*.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *ICML*, pages 10–18. PMLR.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *EMNLP*, pages 1953–1967.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *ICMI*, pages 50–57.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Maxime Peyrard, Sarvjeet Singh Ghotra, Martin Josifoski, Vidhan Agarwal, Barun Patra, Dean Carignan, Emre Kiciman, and Robert West. 2021. Invariant language modeling. *arXiv:2110.08413*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, pages 527–536.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *ACL*, pages 5434–5445.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. 2017. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M Mooij. 2012. On causal and anticausal learning. In *ICML*.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2022. Unsupervised learning of debiased representations with pseudo-attributes. In *CVPR*, pages 16742–16751.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP*, pages 353–355.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv:2010.06032*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022. Discovering invariant rationales for graph neural networks. *arXiv:2201.12872*.

Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, and Tong Zhang. 2022. Model agnostic sample reweighting for out-of-distribution learning. In *ICML*, pages 27203–27221. PMLR.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *ACL*, pages 1651–1661.

A Instantiated Causal Graphs

We instantiate causal graphs as shown in Fig. 2. Specifically, we use the example sentences, whose labels are determined by how words compose the meaning of the sentences. We use $C = 0, 1, 2$ to denote three different sentiment-related phrases, and use $N = 0, 1, 2$ to denote three different non-causal factors for simplicity.

- $C \perp\!\!\!\perp N$: The raw input sentences and bias-related parts are independently sampled and spliced.
- $C \to N$: The type of each causal part respects a given (static) probability distribution. The value of C, and the probability distribution of its causal part is given by:

$$P(N) = \begin{cases} 0.9 & \text{if } X = C, \\ 0.1 & \text{otherwise.} \end{cases} \quad (9)$$

- $N \to C$: Similar to the example for $C \to N$.
- $N \leftarrow V \to C$: There is a latent variable V takes continuous value from 0 to 1, and the probability distribution of N and C *s.t.*

$$N \sim \mathcal{B}(3, V), \quad C \sim \mathcal{B}(3, 1 - V), \quad (10)$$

where $\mathcal{B}$ stands for binomial distribution, i.e., for the variable C, if $C \sim \mathcal{B}(n, p)$ we have

$$P(C = k \mid p, n) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

B Bias Words List

We used the gender attribute words and target words lists proposed in (Kaneko and Bollegala, 2021), which is widely used in debiasing studies (Guo et al., 2022; Liang et al., 2020). In addition, we used the race attribute words and attribute words provided in (Manzini et al., 2019).

C External Corpora

Table 5 summarizes the five external datasets along with examples of the numerous templates occurring across various individuals, settings and in both written and spoken text. The external corpora E used to expand the downstream dataset is created from (Liang et al., 2020), including: 1) **WikiText-2** (Merity et al., 2017) – a dataset of formally written Wikipedia articles, where only the first 10% of WikiText-2 is used and verified sufficiently to capture formally written text); 2) **Stanford Sentiment Treebank** (Socher et al., 2013) is a collection of 10000 polarized written movie reviews; 3) **Reddit** data collected from discussion forums relating to politics, electronics, and relationships; 4) **MELD** (Poria et al., 2019) – a large-scale multimodal multi-party emotional dialog dataset collected from the TV-series Friends; and 5) **POM** (Park et al., 2014) – a dataset of spoken review videos collected across 1,000 individuals spanning multiple topics. These datasets have also been used in recent research in language understanding (Merity et al., 2017; Liu et al., 2019) and multimodal human language (Liang et al., 2019).

D SEAT Details

The WEAT metric measures the bias by comparing two sets of attribute words W_a (i.e., M and F) and two sets of target words W_t (i.e., A and B). In the case of gender, M denotes masculine words like "he", and F denotes feminine words like "she". Meanwhile, A and B are gender-neutral words (e.g., career or adjectives) whose embeddings should be equivalent between M and F. Formally, the bias degree of each word w is defined as:

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(w, a) - \frac{1}{|B|} \sum_{b \in B} \cos(w, b), \quad (11)$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity. Based on Equation (11), the WEAT effect size is:

$$d_{\text{WEAT}} = \frac{\mu(\{s(m, A, B)\}_{m \in M}) - \mu(\{s(f, A, B)\}_{f \in F})}{\sigma(\{s(t, A, B)\}_{t \in A \cup B}))}, \quad (12)$$

where μ and σ denote the mean and standard deviation, respectively. The SEAT metric generalizes the WEAT via replacing the word embeddings with a few simple sentence templates (e.g., "This is the <word>"). We can conclude from Equation (12) that the absolute SEAT effect size closer to 0 means lower biases. We list more details about the SEAT tests that are used in our experiments in Table 6, which are adapted from (Caliskan et al., 2017).

Dataset	Type	Topics	Formality	Length	Examples
Wikitext-2	written	everything	formal	24.0	"Ireland has made a large contribution to world literature in all its branches, particularly in the English language. Poetry in Irish is among the oldest vernacular poetry in *Europe/Africa*, with the earliest examples dating from the 6th century."
SST	written	movie reviews	informal	19.2	"*his/her* fans walked out muttering words like horrible and terrible, but had so much fun dissing the film that they didn't mind the ticket cost."
Reddit	written	politics, electronics, relationships	informal	13.6	"roommate cut my hair without my consent, ended up cutting *himself/herself* and is threatening to call the police on me"
MELD	spoken	comedy TV-series	informal	8.1	"that's the kind of strength that I want in the *man/woman* I love!"
POM	spoken	opinion videos	informal	16.0	"and *his/her* family is, like, incredibly confused"

Table 5: The five external corpora E used to expand the downstream dataset (Liang et al., 2020). Length represents the average length measured by the number of words in a sentence. Words in italics indicate the words used to intervene by casual invariant learning, e.g., (*man, woman*), (*Europe, Africa*). This table summarizes our expanded dataset in terms of topics, formality, and spoken/written text.

Bias Type	**Test**	**Demographic-specific words**	**Stereotype words**
Racial	SEAT-3	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-3b	European-American/African American terms	Pleasant vs. Unpleasant
	SEAT-4	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-5	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-5b	European-American/African American terms	Pleasant vs. Unpleasant
Gender	SEAT-6	Male vs. Female names	Career vs. Family
	SEAT-6b	Male vs. Female terms	Career vs. Family
	SEAT-7	Male vs. Female terms	Math vs. Arts
	SEAT-7b	Male vs. Female names	Math vs. Arts
	SEAT-8	Male vs. Female names	Science vs. Arts
	SEAT-9b	Male vs. Female terms	Science vs. Arts

Table 6: The SEAT test details extended from (Caliskan et al., 2017).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7,8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
4

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.