Conditional Supervised Contrastive Learning for Fair Text Classification

Jianfeng Chi
University of Virginia
jc6ub@virginia.edu

William Shand
University of Virginia
wss2ec@virginia.edu

Yaodong Yu
UC Berkeley
yyu@eecs.berkeley.edu

Kai-Wei Chang
UCLA
kwchang@cs.ucla.edu

Han Zhao
UIUC
hanzhao@illinois.edu

Yuan Tian
UCLA
yuant@ucla.edu

Abstract

Contrastive representation learning has gained much attention due to its superior performance in learning representations from both image and sequential data. However, the learned representations could potentially lead to performance disparities in downstream tasks, such as increased silencing of underrepresented groups in toxicity comment classification. In light of this challenge, in this work, we study learning fair representations that satisfy a notion of fairness known as equalized odds for text classification via contrastive learning. Specifically, we first theoretically analyze the connections between learning representations with a fairness constraint and *conditional supervised contrastive objectives*, and then propose to use conditional supervised contrastive objectives to learn fair representations for text classification. We conduct experiments on two text datasets to demonstrate the effectiveness of our approaches in balancing the trade-offs between task performance and bias mitigation among existing baselines for text classification. Furthermore, we also show that the proposed methods are stable in different hyperparameter settings.[1]

1 Introduction

Recent progress in natural language processing (NLP) has led to its increasing use in various domains, such as machine translation, virtual assistants, and social media monitoring. However, studies have demonstrated societal bias in existing NLP models (Bolukbasi et al., 2016; Zhao et al., 2017; May et al., 2019; Bordia and Bowman, 2019; Hutchinson et al., 2020; Webster et al., 2020; de Vassimon Manela et al., 2021; Sheng et al., 2021). In one major NLP application, text classification, bias is referred as the performance disparity of the trained classifiers over different demographic groups such as gender and ethnicity (Sun et al., 2019; Weidinger et al., 2021). Such bias poses potential risks: for example, toxicity classification models in online social media platforms show disparate performance in different social groups, leading to increased silencing of underserved groups (Dixon et al., 2018; Blodgett et al., 2020).

Meanwhile, an increasing line of work in contrastive learning (CL) has led to significant advances in representation learning (Hadsell et al., 2006; Logeswaran and Lee, 2018; He et al., 2020; Henaff, 2020; Chen et al., 2020; Khosla et al., 2020; Gao et al., 2021b). The general idea of contrastive learning in these works is to learn representations such that similar examples stay close to each other while dissimilar ones are far apart. Inspired by those works, recent works (Shen et al., 2021; Tsai et al., 2021, 2022) also propose to leverage contrastive learning to learn fair representations in classification. However, these works either lack theoretical justifications for the proposed approaches or adopt *demographic parity* (Dwork et al., 2012) as the fairness criterion, which eliminates the perfect classifier in the common scenario when the *base rates* differ among demographic groups (Hardt et al., 2016; Zhao and Gordon, 2019).

In this work, we aim to mitigate bias in text classification models via contrastive learning. In particular, we adopt the fairness notion, *equalized odds* (EO) (Hardt et al., 2016), which asks for equal true positive rates (TPRs) and false positive rates (FPRs) across different demographic groups (Zhao et al., 2019a). Based on information-theoretic concepts, we bridge the problem of learning fair representations with equalized odds constraint with contrastive learning objectives. We then propose an algorithm, called *conditional supervised contrastive learning*, to learn fair text classifiers.

Empirically, we conduct experiments on two text classification datasets (*e.g.*, toxic comment classification and biography classification) to show the proposed methods (1) can flexibly tune the

[1] Our code is publicly available at: https://github.com/JFChi/CSCL4FTC.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2736 - 2756

trade-offs between main task performance and the fairness constraint; (2) achieve the best trade-offs between main task performance and equalized odds compared to the existing bias mitigation approaches in text classification; (3) are stable to different hyperparameter settings, such as data augmentations, temperatures, and batch sizes. To the best of our knowledge, our work is the first to both theoretically and empirically study how to ensure the EO constraint via contrastive learning in text classification.

2 Background

We use $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ to denote the random variables for the input text and the categorical label for the main task, respectively. Furthermore, $A \in \mathcal{A}$ is the sensitive attribute (protected group) associated with the input text X (*e.g.*, the gender information in the occupation classification task). The corresponding lowercase letters denote the instantiation of the random variables. Given a text encoder $f : \mathcal{X} \to \mathcal{Z}$ (*e.g.*, BERT (Devlin et al., 2019)) and a classifier $g : \mathcal{Z} \to \mathcal{Y}$, we first transform the input text X into latent representation Z via f, and Z is used to give a prediction $\hat{Y}$ via g (*i.e.*, $X \xrightarrow{f} Z \xrightarrow{g} \hat{Y}$).

In the context of contrastive learning, data augmentation strategies have been widely adopted. Let $\mathcal{T}$ be a set of data augmentations and X' be the augmented input given the data augmentation $t(\cdot)$: $X' = t(X),\ t \sim \mathcal{T}$, where we assume that the augmentation t is sampled uniformly at random from $\mathcal{T}$. Similarly, we have $X' \xrightarrow{f} Z' \xrightarrow{g} \hat{Y}'$. Let H denote the entropy and I denote the mutual information, *e.g.*, $H(Z \mid Z', Y)$ is the conditional entropy of Z given Z' and Y, and $I(Z'; Z \mid Y)$ is the conditional mutual information of Z' and Z given Y. Due to the space limit, we refer readers to Cover (1999) for more background knowledge of the related notions (entropy and mutual information) in information theory.

We assume there is a joint distribution over X, Y, and A from which the data are sampled. Figure 1 shows the graphical model of the dependencies between input variables and outputs. We also assume that the sensitive attribute A is available only during model training, but it is not available during the testing phase. As a result, any post-processing methods that leverage sensitive attributes for bias mitigation during the testing phase are not feasible in our setting. In this work, we use

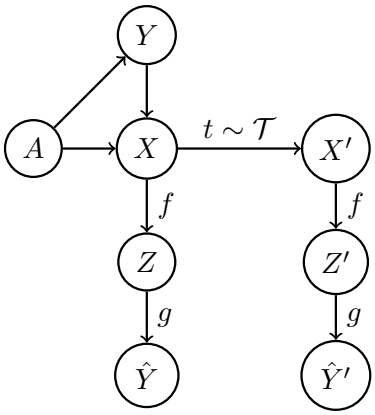

Figure 1: Graphical model of the dependencies between input variables and outputs. Note that we only assume there is a joint distribution over X, Y, and A from which the data are sampled, so the figure only shows one case of the dependencies over X, Y, and A.

equalized odds, a more refined fairness criterion for classification problems.

Definition 2.1 (Equalized Odds (Hardt et al., 2016))**.** A model satisfies equalized odds if $\hat{Y} \perp A \mid Y$.

At a high level, EO asks the model prediction to be independent of the sensitive attribute conditioned on the task label. If a model perfectly satisfies equalized odds, the differences of true positive rates and false positive rates across demographic groups will be 0. Equivalently, it also implies $I(\hat{Y}; A \mid Y) = 0$. Consider online comment toxicity classification as a real-world example to motivate the use of EO as a notion of fairness. In this case, false positive cases (benign text comments marked as toxic) can be seen as unintentional censoring, and false negative cases (toxic text comments marked as benign) might result in debates and discomforts (Baldini et al., 2021).

In contrast to another well-known group fairness definition, *i.e.*, demographic parity, EO does not require positive prediction rates to be the same across different demographic groups, which could possibly severely downgrade the model performance when the sensitive attribute is correlated to the task label (Hardt et al., 2016; Zhao and Gordon, 2019).

3 Our Method

In this section, we first theoretically connect learning fair representations with contrastive learning (Sec. 3.1). In particular, we first show that learning fair representations for equalized odds requires the minimization of $I(Z'; Z \mid Y)$ and the simultane-

ous maximization of $I(Z';Z \mid A,Y)$. To this end, we provide an upper bound of $I(Z';Z \mid Y)$ and a lower bound of $I(Z';Z \mid A,Y)$ to relax the original objective and then establish a relationship between the bounds and the (conditional) supervised contrastive learning objectives. Finally, inspired by our theoretical analysis, we design two practical methods for learning fair representations (Sec. 3.2). Due to the space limit, we defer all detailed proofs to Appendix A.

3.1 Connections between Contrastive Learning and Learning Fair Representations

In order to learn a model (text encoder followed by classifier) to satisfy equalized odds, we aim to learn a latent representation Z such that $Z \perp A \mid Y$. From an information-theoretic perspective, it suffices to minimize the conditional mutual information $I(Z;A \mid Y)$ to ensure EO due to the celebrated data-processing inequality. We identify a connection between contrastive learning and learning fair representations when the representations enjoy certain benign structures. Next, we formally state the assumptions to characterize such a structure.

Assumption 3.1. Let Z and Z' be the corresponding features from X and X', respectively. We assume that there exists a small positive constant $\epsilon > 0$, such that $H(Z \mid Z',Y) \leq \epsilon$.

At a high level, Assumption 3.1 says that the learned features from the contrastive learning procedure are well conditionally aligned (Wang and Isola, 2020). Specifically, given the label of a feature and its corresponding augmented feature, it is relatively easy to infer the corresponding positive pair used in the contrastive learning procedure. Note that the conditional entropy could be understood as the minimum inference error from this perspective (Farnia and Tse, 2016). Under Assumption 3.1, we provide the following lemma to characterize the relationship between Z, Z', A, and Y in terms of (conditional) mutual information.

Lemma 3.1. Under Assumption 3.1, given a set of data augmentations $\mathcal{T}$, let X' be the augmented input data where $X' = t(X)$, $t \sim \mathcal{T}$. Assuming the following Markov chains $X \xrightarrow{f} Z \xrightarrow{g} \hat{Y}$ and $X' \xrightarrow{f} Z' \xrightarrow{g} \hat{Y}'$ hold, we have

$$\begin{aligned} & I(Z';Z \mid Y) - I(Z';Z \mid A,Y) - \epsilon \\ \leq\ & I(Z;A \mid Y) \\ \leq\ & I(Z';Z \mid Y) - I(Z';Z \mid A,Y) + \epsilon. \end{aligned}$$

Lemma 3.1 indicates that we can minimize $I(Z;A \mid Y)$ via (1) minimizing $I(Z';Z \mid Y)$ and (2) maximizing $I(Z';Z \mid A,Y)$. In what follows, we will present an upper (lower) bound to minimize $I(Z';Z \mid Y)$ (maximize $I(Z';Z \mid A,Y)$) and connect the bounds with contrastive learning objectives. We first provide an upper bound of $I(Z';Z \mid Y)$.

Proposition 3.1. Given the assumptions in Lemma 3.1, we have

$$\begin{aligned} & I(Z';Z \mid Y) \\ \leq\ & -\mathbb{E}_{p(y)}\left[\mathbb{E}_{p(z'|y)}[\mathbb{E}_{p(z|y)}[\log p(z' \mid z,y)]]\right]. \end{aligned}$$

In order to better interpret the right side in Proposition 3.1, we define a similarity function $s(z',z;y)$ between z' and z for each y and assume $s(z',z;y) \propto p(z' \mid z,y)$ (*i.e.*, the more similar z' and z are in the latent space given task label y, the more likely z' is generated by z via data augmentation)[2]. With this assumption, the upper bound provided in Proposition 3.1 implies that $I(Z';Z \mid Y)$ can be minimized by encouraging similarity between any latent representations given the same task label, which is consistent with the goal of supervised contrastive loss (Khosla et al., 2020). Formally, given a batch of augmented examples $(x_i,y_i,a_i)_{i=1}^{2N}$ with size $2N$, where the last half examples of the batch are the augmented views of the first half and they share the same task labels (as well as the same sensitive attributes), *i.e.*, $x_{i+N} = t(x_i)$ for $i \in [N]$ and $t \sim \mathcal{T}$. Let N_{y_i} be the total number of examples in the batch that have the same task label as y_i, then supervised contrastive loss takes the following form:

$$L_{\text{sup}} = -\sum_{i=1}^{2N} \frac{1}{N_{y_i}-1} \sum_{j=1}^{2N} \mathbf{1}_{i \neq j, y_i = y_j} \log(\ell_{ij}), \tag{1}$$

[2] In the remainder of the paper, we let $s(\cdot,\cdot) = s(\cdot,\cdot;y)$, $\forall\, Y = y$ for the ease of practical implementations.

and ℓ_{ij} is defined as

$$\ell_{ij} = \frac{\exp\big(f(x_i)\cdot f(x_j)/\tau\big)}{\sum_{k=1}^{2N}\mathbf{1}_{i\neq k}\exp\big(f(x_i)\cdot f(x_k)/\tau\big)},$$

where τ is the temperature parameter, $\mathbf{1}_{i\neq k} = \mathbf{1}\{i \neq k\}$ and $\mathbf{1}\{\cdot\}$ is the indicator function, the similarity function is $s\big(f(x_i), f(x_j)\big) = \exp\big(f(x_i)\cdot f(x_j)/\tau\big)$, and the $\cdot$ symbol denotes the inner (dot) product. Supervised contrastive loss L_{sup} aims to encourage similarity between different examples with the same task label and discourage the ones having different labels. Thus, we minimize L_{sup} to approximately minimize $I(Z'; Z \mid Y)$. Next, we provide a lower bound of $I(Z'; Z \mid A, Y)$ for the maximization of $I(Z'; Z \mid A, Y)$.

Proposition 3.2. Given the assumptions in Lemma 3.1, define conditional supervised InfoNCE as CS-InfoNCE, *i.e.*,

$$\underbrace{\sup_s \mathbb{E}_{p(a,y)}\Big[\mathbb{E}_{p(z_i', z_i\mid a,y)^{\otimes N}}\big[\log \frac{\exp(s(z_i', z_i))}{\frac{1}{N}\sum_{j=1}^{N}\exp(s(z_i', z_j))}\big]\Big]}_{\text{CS-InfoNCE}},$$

where $p(\cdot)^{\otimes N}$ denotes the probability distributions of N independent examples and $s(\cdot,\cdot)$ is any similarity function that measure the similarity of z_i' and z_i. Then, we have

$$\text{CS-InfoNCE} \leq I(Z'; Z \mid A, Y).$$

Proposition 3.2 indicates the maximization of CS-InfoNCE leads to the maximization of $I(Z'; Z \mid A, Y)$. Given the examples that share the same task label and sensitive attribute, CS-InfoNCE encourages the similarity between different views of the same examples while discouraging others. Note that all positive and negative examples w.r.t. the anchoring example share the same task label and sensitive attribute. Given the same batch of examples $(x_i, y_i, a_i)_{i=1}^{2N}$ with size $2N$, we can formulate the contrastive objective as

$$L_{\text{CS-InfoNCE}} = -\sum_{i=1}^{2N}\frac{1}{N_{a_i,y_i}-1}\log(\ell_i), \quad (2)$$

and ℓ_i is defined as

$$\ell_i = \frac{\exp\big(f(x_i)\cdot f(x_i')/\tau\big)}{\sum_{k=1}^{2N}\mathbf{1}_{i\neq k, a_i=a_k, y_i=y_k}\exp\big(f(x_i)\cdot f(x_k)/\tau\big)},$$

where N_{a_i,y_i} is the total number of examples in the batch that have the same task label and sensitive attribute as y_i and a_i, and x_i and x_i' are the different views of the same example.

Interpretation of L_{sup} and $L_{\text{CS-InfoNCE}}$ in learning fair representations. In learning fair representations, the role of L_{sup} is to learn aligned and uniform representations (Wang and Isola, 2020) for each task label, while the role of $L_{\text{CS-InfoNCE}}$ is to encourage the dissimilarity of different examples that share the same task labels and sensitive attributes. In an ideal case where $L_{\text{sup}} = 0$, each data point that shares the same task label in the latent space collapse to a single point, and the perfect representations are learned. In this case, $L_{\text{CS-InfoNCE}} = 0$ as well. In practice, the overall combined effect of the L_{sup} and $L_{\text{CS-InfoNCE}}$ will encourage the similarity of examples having the same task label but belonging to different groups. Thus, our theory could also explain why other slightly different proposed contrastive objectives in a concurrent work (Park et al., 2022) could mitigate equalized odds. In Appendix C.2, we provide T-SNE visualization (Van der Maaten and Hinton, 2008) of the text embeddings using different training objectives to help better understand our methods.

3.2 Practical Implementations

The existing contrastive representation learning approaches fall into two categories: two-stage methods (Khosla et al., 2020; Chen et al., 2020) and one-stage methods (Gunel et al., 2021; Cui et al., 2021). Two-stage methods first pretrain the encoder in the first stage using the contrastive objective, then fix the encoder, and fine-tune the classifier using cross-entropy (CE) loss in the second stage. One-stage methods train both encoder and classifier using CE loss and contrastive loss end-to-end. Following the previous settings, we also implement our methods in these two ways. For the two-stage CL method, we first pretrain the text encoder using the following loss function in the first stage:

$$L_{\text{sup}} + \lambda \cdot L_{\text{CS-InfoNCE}}, \quad (3)$$

then we fix the pretrained encoder, and fine-tune classifier using CE loss. Note that $\lambda \geq 0$ controls the intensity of $L_{\text{CS-InfoNCE}}$. For the one-stage CL method, similar to Gunel et al. (2021), we formulate the loss function as:

$$(1-\gamma)\cdot L_{\text{CE}} + \gamma \cdot L_{\text{sup}} + \lambda \cdot L_{\text{CS-InfoNCE}}, \quad (4)$$

where $\gamma \in [0, 1]$ controls the relative weight of L_{sup} compared to L_{CE}. The major advantage of

our approach is that it can be directly substituted into existing NLP pipelines that use the "pretrain-and-finetune" paradigm popularized by large language models such as BERT. NLP practitioners can swap the fair CL finetuner into these pipelines to boost model fairness at low cost, with robust behavior against hyperparameter choices (see Sec. 4.2). Whereas large language models made it simple to build models with high performance, fair CL makes it simple to build models with high performance and fairness.

4 Experiments

In this section, we conduct experiments to investigate the following research questions:

RQ 1. How can we control the trade-offs between model classification performance and fairness via conditional supervised contrastive learning?

RQ 2. How do conditional supervised contrastive learning methods perform in terms of trade-offs between model performance and fairness compared to other in-processing bias mitigation methods in text classification?

RQ 3. Is conditional supervised contrastive learning sensitive to hyperparameter changes?

4.1 Experimental Setup

Datasets. We perform experiments using the following two datasets (see Appendix B for more details of the datasets and the data prepossessing pipelines):

- `Jigsaw-toxicity`[3] is a dataset for online comment toxicity classification. The main task of the dataset is to determine if the online comment is toxic, and we use "race and ethnicity" as the sensitive attribute (*e.g.*, whether "black" identity is mentioned in the comment text or not).
- `Biasbios` (De-Arteaga et al., 2019) is a dataset for occupation classification. The main task of the dataset is to determine the people's occupations given their biographies. The sensitive attribute is binary gender (*i.e.*, male and female).

[3]The dataset is publicly available at: https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification.

Evaluation Metrics. We evaluate our model based on model classification performance and EO fairness. We use the F1 score for model performance and True Positive Equality Difference + False Positive Equality Difference (Dixon et al., 2018) for EO fairness:

$$\Delta_{\text{TPR}} = \sum_a |\text{TPR}_a - \text{TPR}_{\text{overall}}|,$$
$$\Delta_{\text{FPR}} = \sum_a |\text{FPR}_a - \text{FPR}_{\text{overall}}|,$$

where Δ_{TPR} (Δ_{FPR}) is the true positive rate (false negative rate) for sensitive attribute a and $\text{TPR}_{\text{overall}}$ ($\text{FPR}_{\text{overall}}$) is the overall true positive rate (false negative rate). Following Pruksachatkun et al. (2021), we define equalized odds gap $\Delta_{\text{EO}} = \Delta_{\text{TPR}} + \Delta_{\text{FPR}}$, since equalized odds aligns with $\Delta_{\text{TPR}} + \Delta_{\text{FPR}}$, and when it is satisfied, $\Delta_{\text{TPR}} = \Delta_{\text{FPR}} = 0$ (Borkan et al., 2019). Note that when $|\mathcal{Y}| > 2$, Δ_{EO} will be summed over each value in $\mathcal{Y}$ since TPR and FPR are defined over each class (*i.e.*, $\Delta_{\text{EO}} = \sum_y \Delta^y_{\text{EO}}$).

Implementations and Baselines. In our experiments, we use BERT (Devlin et al., 2019) (`bert-base-uncased` as the text encoder followed by a two-layer MLP as the classifier)[4]. As suggested by previous works (Khosla et al., 2020; Gao et al., 2021b), the performance of contrastive learning is closely related to the choice of the following hyperparameters: (1) temperature, (2) (pre-training) batch size, and (3) data augmentation strategy. Thus, we conduct a grid-based hyperparameter search for temperature $\tau \in \{0.1, 0.5, 1.0, 2.0\}$, (pre-training) batch size $\text{bsz} \in \{32, 64, 128, 256\}$, and data augmentation strategy $t \in$ {EDA, back translation, CLM insert, CLM substitute} (see Appendix B for the detailed description of different augmentation strategies) for both two-stage CL and one-stage CL. We also conduct grid search of $\gamma \in \{0.1, 0.3, 0.7, 0.9\}$ in Eq. (4) for one-stage CL. In Appendix B, we provide the remaining hyperparameter details (*e.g.*, learning rate, training epochs, optimizer). Since it is not feasible to train large language models with large batch sizes via contrastive objectives given limited GPU memory, we use the gradient cache technique (Gao et al., 2021a) to adapt our implementations to limited GPU memory settings.

[4]We use the huggingface transformer implementation: https://github.com/huggingface/transformers.

We compare our methods with the following baselines, which have been empirically demonstrated effective for bias mitigation in text classification:

(1) Adversarial training (Elazar and Goldberg, 2018): Following the encoder + classifier setting, adversarial training leverages a discriminator to learn latent representations oblivious to the sensitive attribute. Note that the original adversarial training method is tailored for demographic parity and it is well known that demographic parity and equal odds are incompatible given different base rates (Kleinberg et al., 2017; Ball-Burack et al., 2021). To this end, we use the conditional learning techniques (Madras et al., 2018; Zhao et al., 2019a) to adapt adversarial training for equalized odds.

(2) Adversarial training with diverse adversaries (diverse adversaries) (Han et al., 2021): Adversarial training with diverse adversaries improves adversarial training by using an ensemble of discriminators and encourages the discriminator to learn orthogonal representations. Similar to adversarial training, we also apply the conditional learning techniques for learning the adversarial discriminators.

(3) Iterative null-space projection (INLP) (Ravfogel et al., 2020): Given a pretrained text encoder (we use CE loss to pretrain the text encoder and drop the prediction head using the validation set), INLP learns a linear guarding layer on top of the pretrained text encoder to filter the sensitive information and fine-tune the classifier given the pretrained text encoder and INLP. INLP learns the linear guarding layer by projecting the parameter matrices of linear classifiers (*e.g.*, SVM) to their null spaces iteratively. The training data of linear classifiers are the latent representations of input texts and sensitive attributes. In order to tailor INLP for equaled odds, Ravfogel et al. (2020) learns the linear classifier given the data from the same class each round.

We also use training using CE loss as a baseline. Except for INLP, all methods we test in our experiments train the text encoder (*e.g.*, BERT) directly, while INLP is a post-hoc debiasing method given a text encoder. In a sense, INLP is orthogonal to other methods since it tries to remove group-specific information after we learn the representations, while other methods learn the fair representations directly. We run each experiment with five different seeds and report the mean and standard deviation values for each evaluation metric.

4.2 Results and Analysis

RQ 1. In order to control the trade-offs between model classification performance and EO fairness, we vary the values of λ in Eq. (3) and Eq. (4). Figure 2 shows the classification performance and EO fairness of one-stage and two-stage CL when λ changes. Overall, as λ increases, the equalized odds gaps shrink at the cost of model classification performance. Compared to one-stage CL, two-stage CL achieves more flexible trade-offs in general. Given the same range of λ, the change of equalized odds gaps in two-stage CL is more significant than in one-stage CL. At the same time, the corresponding model classification performances are comparable or remain better.

RQ 2. We study the trade-offs between model performance and EO fairness of our proposed methods compared to the baselines. Figure 3 displays the performance and fairness of these methods under different hyperparameter settings for the `jigsaw` and `biasbios` datasets (trade-off parameters for all methods are described in more detail in Appendix B).

Among all methods, we find that two-stage CL and INLP achieve the best performance and fairness trade-offs. In the `biasbios` dataset, two-stage CL and INLP achieve similar performance and fairness trade-offs, and two-stage CL achieves more consistent results (*i.e.*, lower variance). In the `jigsaw` dataset, two-stage CL achieves more flexible performance and fairness trade-offs as it reaches the highest model performance. Besides, when F1 scores are around 0.58, two-stage CL also achieves more consistent results and a lower EO gap. Meanwhile, when F1 scores are between 0.62$\sim$0.64, INLP performs better. We note that the effectiveness of INLP highly depends on the pretrained encoder for INLP (see Appendix C.3 for the effects of different pre-training strategies for the text encoder in INLP), and a slight change in the text encoder could lead to a significant difference in the results, while CL-based methods target training the text encoder directly to ensure EO fairness and we demonstrate they are stable under hyperparame-

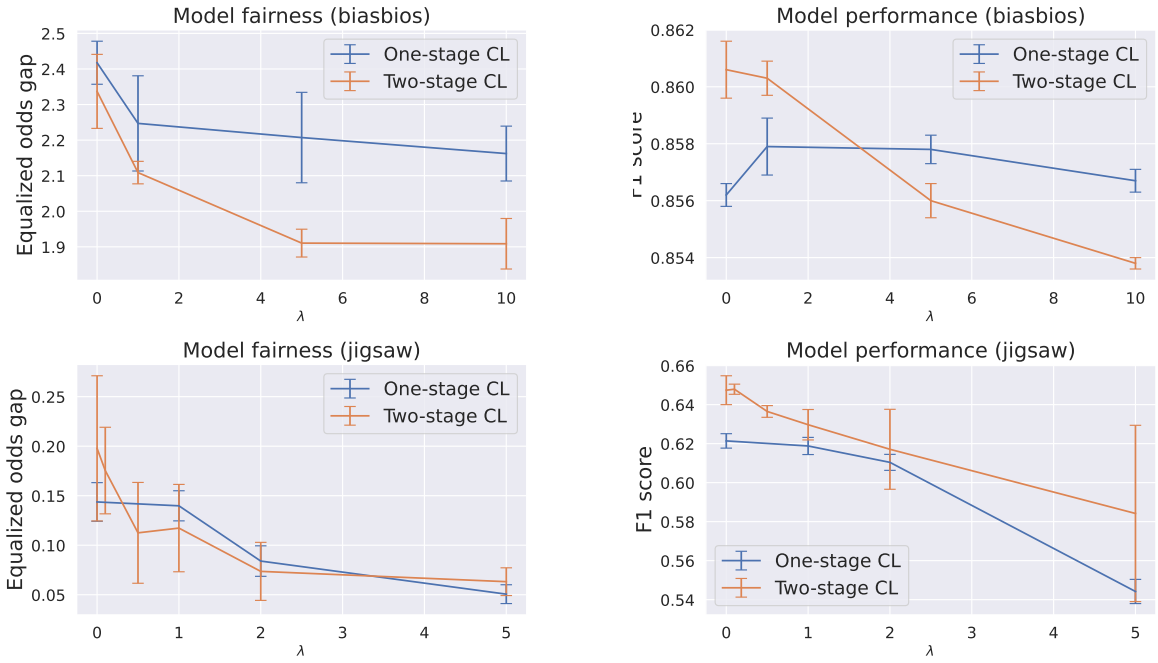

Figure 2: Classification performance and EO fairness of one-stage and two-stage CL when λ changes. The equalized odds gaps shrink at the cost of model classification performance as λ increases.

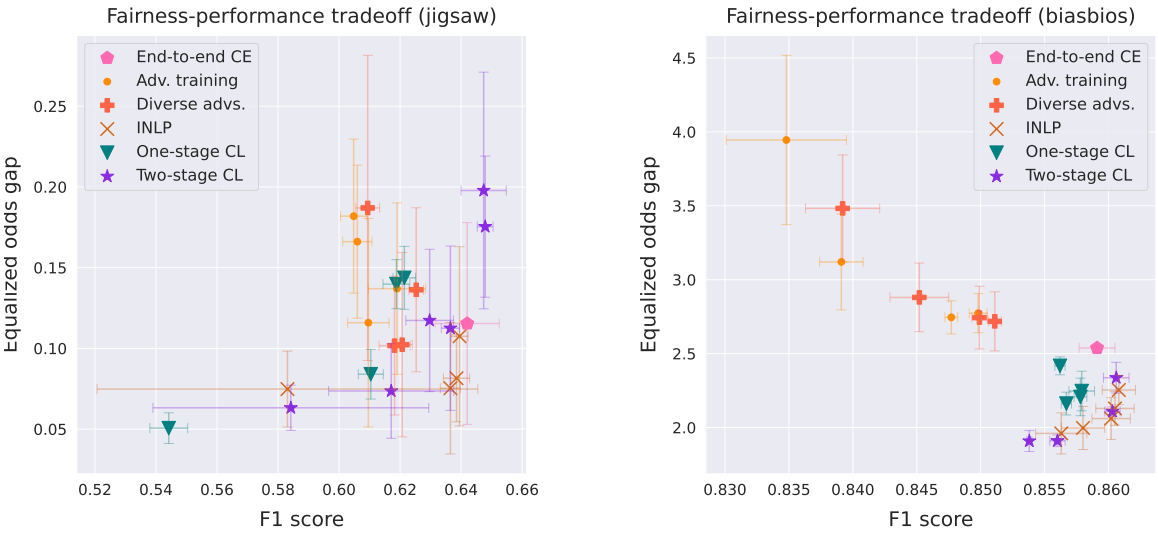

Figure 3: Classification performance and EO fairness and of our proposed methods compared against the baselines. Two-stage CL and INLP achieves the best performance and fairness trade-offs in general, and two-stage CL typically achieves more consistent results with lower variance.

ter changes (see RQ 3 below).

In comparison, the adversarial-training-based methods are relatively more unstable and consistently perform worse than CL-based methods and INLP, especially in the `biasbios` dataset. Furthermore, both adversarial-training-based methods and INLP introduce additional model components (*e.g.*, adversarial networks in adversarial-training-based methods and linear guarding layer in INLP) during training or inference, which complicates the actual implementation of the whole pipeline. In contrast, CL-based methods are well-suited to pre-training and fine-tuning paradigms in NLP applications.

RQ 3. We have shown that two-stage CL performs better than one-stage CL in RQ 1 and RQ 2. Thus, we choose two-stage CL to see if it is sensitive to key hyperparameter changes. As mentioned

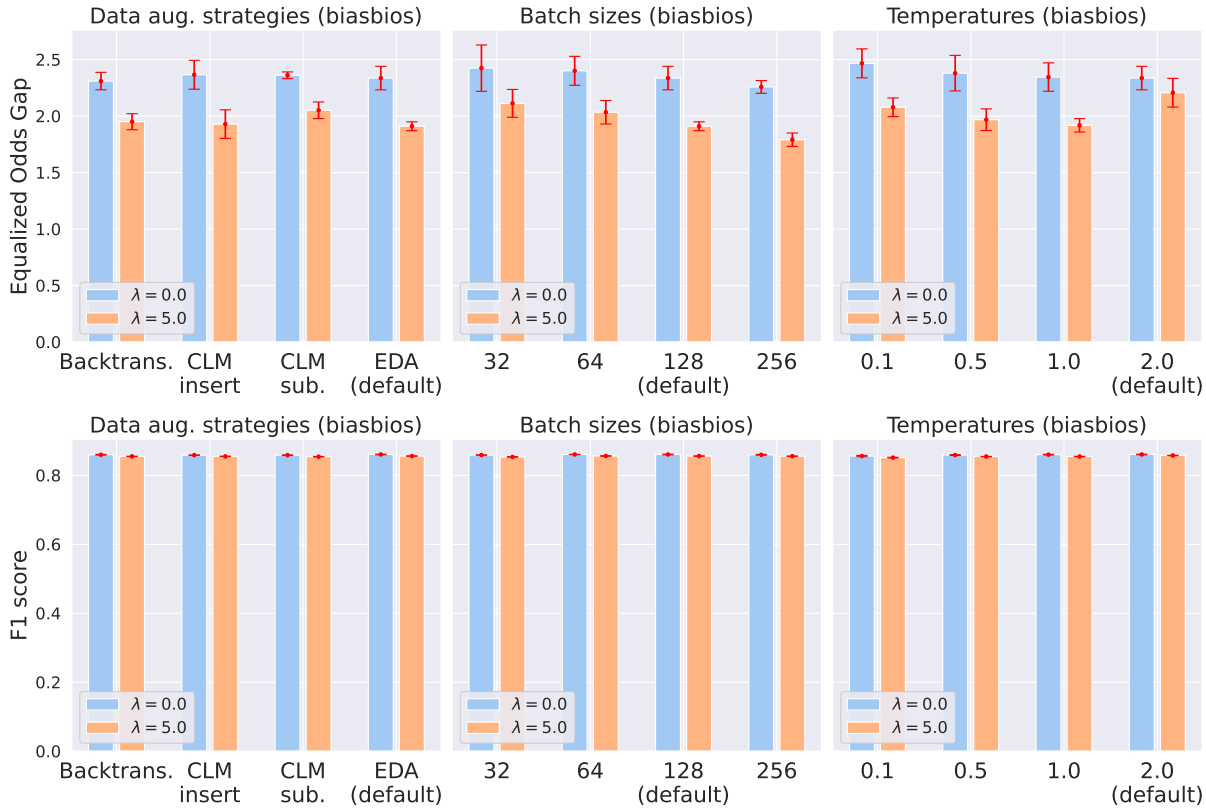

Figure 4: Sensitivity analysis of two-stage CL to key hyperparameter changes in (`biasbios`). "Default" in the X-axis indicates the default hyperparamter settings used in RQ 1 and RQ 2.

above, the performance of contrastive learning is closely related to temperature, (pre-training) batch size, and data augmentation strategy. Thus, we study whether the performance of two-stage CL is sensitive to these hyperparameters.

Figure 4 shows model performance and EO fairness of two-stage CL under different hyperparameter settings when $\lambda \in \{0.0, 5.0\}$ in the `biasbios` dataset (Figure 9 for the `jigsaw` in Appendix C.1). We see that two-stage CL are stable under a wide range of parameter settings: The equalized odds gaps are consistently decreasing when $\lambda = 5.0$ and the F1 scores are relatively high.

5 Related Work

Unintended social biases in NLP models have been identified in word/sentence embedding (Bolukbasi et al., 2016; May et al., 2019; Zhao et al., 2019b) and applications such as coreference resolution (Zhao et al., 2018; Rudinger et al., 2018; Cao and Daumé III, 2020), language modeling (Bordia and Bowman, 2019), machine translation (Stanovsky et al., 2019), and text classification (Ball-Burack et al., 2021; Baldini et al., 2022).

In the literature, there are some recent works that aim to learn fair representations via contrastive learning (Cheng et al., 2021; Shen et al., 2021; Tsai et al., 2021, 2022). Among these works, Cheng et al. (2021) propose contrastive objectives to learn debiased sentence embeddings that minimize the correlation between embedded sentences and biased words. In classification tasks, Tsai et al. (2021, 2022) proposed contrastive objectives to remove sensitive information; Shen et al. (2021) proposed a similar contrastive objective to achieve a similar goal. According to Tsai et al. (2021), all those proposed contrastive objectives target demographic parity in principle.

Our theoretical results involve key notions (*e.g.*, entropy and mutual information) in information theory (Cheng et al., 2020; Colombo et al., 2021). Information-theoretic-based methods have been used for representation learning for NLP applications. For example, Colombo et al. (2021) proposed a variational upper bound of mutual information to learn disentangled textual representations for fair classification and style transfer.

Compared to the previous work, our work uses equalized odds as the fairness criterion. To the best of our knowledge, our work is the first to connect

the problem of learning fair representations with contrastive learning to ensure the EO constraint and explore its effectiveness for bias mitigation in text classification in large language models (*e.g.*, BERT).

6 Conclusion

In this paper, we theoretically and empirically study how to leverage contrastive learning for fair text classification. Inspired by our theoretical results, we propose conditional supervised contrastive objectives to learn aligned and uniform representations while mixing the representation of different examples that share the same sensitive attribute for every task label. We conduct experiments to demonstrate the effectiveness of our algorithms in learning fair representations for text classification and show that our methods are stable in different hyperparameter settings. In the future, we plan to extend our algorithms to the settings of intersectional bias (Kearns et al., 2018; Yang et al., 2020).

Limitations

Like most prior work (Ravfogel et al., 2020; Tsai et al., 2022), we conduct experiments on the binary sensitive attribute. Especially, we acknowledge that due to the limitation of the dataset, our analysis of gender bias only considers binary gender, which is not ideal (Dev et al., 2021). One interesting future direction is to extend our method to ensure fairness for intersectional groups (Kearns et al., 2018; Yang et al., 2020). In principle, our theory also holds in intersectional bias. However, the disproportionate distributions of the intersectional sensitive attributes might pose challenges in sampling negative examples in $L_{\text{CS-InfoNCE}}$. One possible solution is to use a memory bank to sample negative examples (Wu et al., 2018). We leave this analysis as future work as it is an important question that warrants an independent study.

Ethics Statement

This work aims for bias mitigation for text classification. Like other bias mitigation methods, it could help increase people's trust in NLP models. For example, our methods could help reduce unintentional censoring (false positive cases) and debates or discomforts (false negative cases) in online comment toxicity classification (see Section 2 for more details). Our study targets equalized odds and do not capture all notions of bias (*e.g.*, individual fairness) in text classification. These issues are universal to bias mitigation techniques and not particular to our use case.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. Jianfeng Chi, William Shand, and Yuan Tian acknowledge support from NSF #1829004, #1920462, #2002985, a Facebook Faculty Fellowship, and a Google Research Scholar Award. Yaodong Yu acknowledges support from the joint Simons Foundation-NSF DMS grant #2031899. Han Zhao would like to thank the support from a Facebook research award. Kai-Wei Chang acknowledge support from NSF #1927554 and a Sloan Research Fellow.

References

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland. Association for Computational Linguistics.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. 2021. Your fairness may vary: group fairness of pretrained language models in toxic text classification. *arXiv preprint arXiv:2108.01250*.

Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 116–128, New York, NY, USA. Association for Computing Machinery.

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

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

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *NAACL-HLT (Student Research Workshop)*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020. Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7530–7541, Online. Association for Computational Linguistics.

Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2021. A novel estimator of mutual information for learning to disentangle textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6539–6550, Online. Association for Computational Linguistics.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Farzan Farnia and David Tse. 2016. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems*, 29.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021a. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33.

Jon M. Kleinberg, Sendhil Mullainathan, and M. Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *ArXiv*, abs/1609.05807.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019*, pages 622–628. Association for Computational Linguistics (ACL).

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2022. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR.

Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3320–3331, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Tianqin Li, Martin Q. Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2022. Conditional contrastive learning with kernel. In *International Conference on Learning Representations*.

Yao-Hung Hubert Tsai, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Conditional contrastive learning: Removing undesirable information in self-supervised representations. *arXiv preprint arXiv:2106.02866*.

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

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

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 preprint arXiv:2010.06032*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. 2020. Fairness with overlapping groups; a probabilistic perspective. In *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078. Curran Associates, Inc.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. 2019a. Conditional learning of fair representations. In *International Conference on Learning Representations*.

Han Zhao and Geoff Gordon. 2019. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019b. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

A Omitted Proofs

A.1 Proof of Lemma 3.1

Proof. By the definition of conditional mutual information:

$$
\begin{aligned}
&I(Z';Z \mid Y) - I(Z';Z \mid A,Y) \\
&= \big(H(Z \mid Y) - H(Z \mid Z',Y)\big) \\
&\quad - \big(H(Z \mid A,Y) - H(Z \mid Z',A,Y)\big) \\
&= \big(H(Z \mid Y) - H(Z \mid A,Y)\big) \\
&\quad + \big(H(Z \mid Z',A,Y) - H(Z \mid Z',Y)\big) \\
&= I(Z;A \mid Y) \\
&\qquad + \big(H(Z \mid Z',A,Y) - H(Z \mid Z',Y)\big) \\
&\leq I(Z;A \mid Y) + H(Z \mid Z',A,Y) \\
&\leq I(Z;A \mid Y) + H(Z \mid Z',Y) \\
&\leq I(Z;A \mid Y) + \epsilon.
\end{aligned}
$$

Next, we prove the opposite side,

$$
\begin{aligned}
&I(Z';Z \mid Y) - I(Z';Z \mid A,Y) \\
&= \big(H(Z \mid Y) - H(Z \mid Z',Y)\big) \\
&\quad - \big(H(Z \mid A,Y) - H(Z \mid Z',A,Y)\big) \\
&= \big(H(Z \mid Y) - H(Z \mid A,Y)\big) \\
&\quad + \big(H(Z \mid Z',A,Y) - H(Z \mid Z',Y)\big) \\
&= I(Z;A \mid Y) \\
&\qquad + \big(H(Z \mid Z',A,Y) - H(Z \mid Z',Y)\big) \\
&\geq I(Z;A \mid Y) - H(Z \mid Z',Y) \\
&\geq I(Z;A \mid Y) - \epsilon,
\end{aligned}
$$

which completes the proof. □

A.2 Proof of Proposition 3.1

Proof.

$$
\begin{aligned}
&I(Z';Z \mid Y) = -H(Z' \mid Z,Y) + H(Z' \mid Y) \\
&= \mathbb{E}_{p(y)}\big[\mathbb{E}_{p(z',z|y)}[\log p(z' \mid z,y)] \\
&\quad - \mathbb{E}_{p(z'|y)}[\log p(z' \mid y)]\big] \\
&= \mathbb{E}_{p(y)}\big[\mathbb{E}_{p(z',z|y)}[\log p(z' \mid z,y)] \\
&\quad - \mathbb{E}_{p(z'|y)}[\log \mathbb{E}_{p(z|y)}[p(z' \mid z,y)]]\big] \\
&\leq \mathbb{E}_{p(y)}\big[\mathbb{E}_{p(z',z|y)}[\log p(z' \mid z,y)] \\
&\quad - \mathbb{E}_{p(z'|y)}[\mathbb{E}_{p(z|y)}[\log p(z' \mid z,y)]]\big] \\
&\leq -\mathbb{E}_{p(y)}\big[\mathbb{E}_{p(z'|y)}[\mathbb{E}_{p(z|y)}[\log p(z' \mid z,y)]]\big]
\end{aligned}
$$

where the second line follows the marginal of a joint distribution can be expressed as the expectation of the corresponding conditional distribution, and the third line follows Jensen's Inequality.

□

A.3 Proof of Proposition 3.2

The proof techniques used in Proposition 3.2 follow in Proposition 2.4 in Tsai et al. (2021), which could be dated back to Oord et al. (2018); Poole et al. (2019). To make the paper self-contained, we include all the details of the lemmas to get the final results.

The proof of Proposition 3.2 is dependent on the Lemmas A.1-A.5 showed in Figures 5 and 6 as well as Proposition A.1 in Figure 7. Finally, we present the proof of Proposition 3.2 in Figure 8.

B Experimental Details

B.1 Data prepossessing pipelines

Jigsaw. Our first dataset, which we refer to as `jigsaw`, is a corpus of comments from an online forum associated with a toxicity rating. `jigsaw`'s main task is binary classification: given a "toxicity" score in the range $[0, 1]$ that has been assigned to each comment, we determine whether the "toxicity" score is greater or equal to 0.5. Each comment is also annotated with some "identity" labels, indicating whether some identities belonging to specific demographic groups are mentioned in the comment. We focus on the identity labels related to "race or ethnicity" and binaries the identity labels into black and non-black. Note that there are other sensitive attributes in the *Jigsaw-Toxicity* dataset, and we constrain the scope of our study to the "race" attributes present in text classification datasets. We follow (Koh et al., 2021) to perform the train/val/test splits. The data with "race or ethnicity" identity labels are split into training, validation, and test sets, summarized in Table 1.

Bias-in-Bios. To measure model fairness and performance in the multi-class classification setting, we use the professional biographies dataset of (De-Arteaga et al., 2019), which we refer to as the `biasbios` dataset. The data consist of nearly 400,000 online biographies collected from the Common Crawl corpus. These biographies are annotated with one of the 28 professions to which their subject belongs. The data are mapped to a binary gender based on the occurrence of gendered pronouns and are scrubbed to exclude the authors' names and pronouns. It is worth noting that mapping gender to binary labels is a strong simplified assumption to map data to a demographic label cleanly; it ignores people who do not identify as female or male, as well as the complexity of gender identity more generally. We refer readers to the original work (De-Arteaga et al., 2019) for further discussion of these issues. For our experiments, we attempt to predict the profession as our task label while protecting against the gender attribute. We replicate the splits of `biasbios` used by (Ravfogel et al., 2020), which are summarized in Table 2.

B.2 Detailed Implementations and Hyperparameter Settings

In this section, we provide more details on our implementations and give the hyperparameters we use in our experiments. We first detail how we tune each method's performance and fairness trade-offs.

- **One-stage / Two-stage CL.** for one- and two-stage CL, once we determine the best classification performance by conducting a grid search on temperature, (pre-training) batch size, and data augmentation strategies (as well as γ in one-stage CL) , we only tune the parameter λ described in Sec. 3.2, which affects the trade-offs between supervised contrastive loss L_{sup} and the conditional supervised InfoNCE loss $L_{\text{CS-InfoNCE}}$. For two-stage CL, we set the pre-training epochs to be 15 and 25 for `jigsaw` and `biasbios`, respectively, and early stop the pre-training if there is no improvement on the validation set for three consecutive epochs.

- **Diverse adversarial training.** Following Han et al. (2021), we use an ensemble of three adversarial discriminators and the same adversarial network architecture. There are two hyperparameters of interest: λ_{diff} and λ_{adv}. λ_{diff} is a difference loss hyperparameter that encourages discriminators to learn orthogonal representations. λ_{adv} affects the trade-offs between task performance and fairness. We first do a grid search on $\lambda_{diff} = \{0, 100, 1000, 5000\}$ and vary the values of λ_{adv} to determine the best hyperparameter configurations.

- **Adversarial training.** The implementation is nearly identical to diverse adversarial training, except that there is just one adversarial discriminator.

- **INLP.** Following Ravfogel et al. (2020), we use the weights of an SVM classifier as the parameters of the linear guarding layer and

Lemma A.1 ((Nguyen et al., 2010)). Let $\mathcal{Z}$ be the sample space for Z' and Z, $s : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be any function, and $\mathcal{P}$ and $\mathcal{Q}$ be the probability measures over $\mathcal{Z} \times \mathcal{Z}$. We have

$$D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q}) = \sup_s \mathbb{E}_{(z',z)\sim\mathcal{P}}[s(z',z)] - \mathbb{E}_{(z',z)\sim\mathcal{Q}}[\exp(s(z',z))] + 1$$

Proof. We first get the second-order functional derivative of the objective: $-\exp(s(z',z)) \cdot d\mathcal{Q}$, which is negative and it implies there is a supreme value for the objective. Next, we set the first-order functional derivative of the objective to be zero:

$$d\mathcal{P} - \exp(s(z',z)) \cdot d\mathcal{Q} = 0.$$

Reorganizing the equation above we get the optimal similarity function $s^*(z',z) = \log(\frac{d\mathcal{P}}{d\mathcal{Q}})$. Plugging it into the original objective, we have

$$\mathbb{E}_{\mathcal{P}}[s^*(z',z)] - \mathbb{E}_{\mathcal{Q}}[\exp(s^*(z',z))] + 1 = \mathbb{E}_{\mathcal{P}}[\log(\frac{d\mathcal{P}}{d\mathcal{Q}})] = D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q}).$$

□

Lemma A.2 (Four-variable variant of Lemma A.1). Let $\mathcal{Z}$ be the sample space for Z' and Z, $\mathcal{Y}$ be the sample space for Y, $\mathcal{A}$ be the sample space for A, $s : \mathcal{Z} \times \mathcal{Z} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ be any function, and $\mathcal{P}$ and $\mathcal{Q}$ be the probability measures over $\mathcal{Z} \times \mathcal{Z} \times \mathcal{Y} \times \mathcal{A}$. We have

$$D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q}) = \sup_s \mathbb{E}_{(z',z,y,a)\sim\mathcal{P}}[s(z',z,y,a)] - \mathbb{E}_{(z',z,y,a)\sim\mathcal{Q}}[\exp(s(z',z,y,a))] + 1$$

Proof. The proof technique is identical to the proof of Lemma A.1 and the only difference is that the similarity function takes four variables as input. □

Lemma A.3. $\sup_s \mathbb{E}_{(z',z_1)\sim\mathcal{P},(z',z_{2:N})\sim\mathcal{Q}^{\otimes N-1}}\big[\log \frac{\exp(s(z',z_1))}{\frac{1}{N}\sum_{j=1}^N \exp(s(z',z_j))}\big] \le D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q})$

Proof.

$$\begin{aligned}
D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q}) &= \mathbb{E}_{(z',z_{2:N})\sim\mathcal{Q}^{\otimes N-1}}\Big[D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q})\Big] \\
&\ge \mathbb{E}_{(z',z_{2:N})\sim\mathcal{Q}^{\otimes N-1}}\left[\mathbb{E}_{\mathcal{P}}[\log \frac{\exp(s^*(z',z))}{\frac{1}{N}\sum_{j=1}^N \exp(s(z',z_j))}] - \mathbb{E}_{\mathcal{Q}}[\frac{\exp(s^*(z',z))}{\frac{1}{N}\sum_{j=1}^N \exp(s(z',z_j))}] + 1\right] \\
&= \mathbb{E}_{(z',z_{2:N})\sim\mathcal{Q}^{\otimes N-1}}\left[\mathbb{E}_{\mathcal{P}}[\log \frac{\exp(s^*(z',z))}{\frac{1}{N}\sum_{j=1}^N \exp(s(z',z_j))}] - 1 + 1\right] \\
&= \mathbb{E}_{(z',z_1)\sim\mathcal{P},(z',z_{2:N})\sim\mathcal{Q}^{\otimes N-1}}\big[\log \frac{\exp(s(z',z_1))}{\frac{1}{N}\sum_{j=1}^N \exp(s(z',z_j))}\big],
\end{aligned}$$

where the first line follows the fact that $D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q})$ is a constant, the second line follows Lemma A.1, the third line follows the fact that (z', z_1) and $(z', z_{2:N})$ are interchangeable when sampling from $\mathcal{Q}$. Thus, for any similarity function s, we have

$$\sup_s \mathbb{E}_{(z',z_1)\sim\mathcal{P},(z',z_{2:N})\sim\mathcal{Q}^{\otimes N-1}}\big[\log \frac{\exp(s(z',z_1))}{\frac{1}{N}\sum_{j=1}^N \exp(s(z',z_j))}\big] \le D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q})$$

□

Figure 5: Lemmas required for the proof of Proposition 3.2.

Lemma A.4.

$$\begin{aligned}&D_{\mathrm{KL}}(P_{Z',Z} \,\|\, \mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}])\\ =\ &\sup_s \mathbb{E}_{(z',z)\sim P_{Z',Z}}[s(z',z)] - \mathbb{E}_{(z',z)\sim \mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}]}[\exp(s(z',z))] + 1.\end{aligned}$$

Proof. We use Lemma A.1 and substitute $\mathcal{P}$ and $\mathcal{Q}$ with $P_{Z',Z}$ and $\mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}]$, respectively. □

Lemma A.5.

$$\begin{aligned}&I(Z';Z \mid A,Y)\\ =\ &D_{\mathrm{KL}}(P_{Z',Z,A,Y} \,\|\, P_{A,Y}P_{Z'|A,Y}P_{Z|A,Y})\\ =\ &\sup_s \mathbb{E}_{(z',z,a,y)\sim P_{Z',Z,A,Y}}[s(z',z,a,y)]\\ &\qquad - \mathbb{E}_{(z',z,a,y)\sim P_{A,Y}P_{Z'|A,Y}P_{Z|A,Y}}[\exp(s(z',z,a,y))] + 1.\end{aligned}$$

Proof. We use Lemma A.2 and substitute $\mathcal{P}$ and $\mathcal{Q}$ with $P_{Z',Z,A,Y}$ and $P_{A,Y}P_{Z'|A,Y}P_{Z|A,Y}$, respectively. □

Figure 6: Lemmas required for the proof of Proposition 3.2.

Proposition A.1.

$$D_{\mathrm{KL}}(P_{Z',Z} \,\|\, \mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}]) \le I(Z';Z \mid A,Y)$$

Proof. We have

$$\begin{aligned}&D_{\mathrm{KL}}(P_{Z',Z} \,\|\, \mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}])\\ =\ &\sup_s \mathbb{E}_{(z',z)\sim P_{Z',Z}}[s(z',z)] - \mathbb{E}_{(z',z)\sim \mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}]}[\exp(s(z',z))] + 1\\ =\ &\sup_s \mathbb{E}_{(z',z,a,y)\sim P_{Z',Z,A,Y}}[s(z',z)] - \mathbb{E}_{(z',z,a,y)\sim P_{A,Y}P_{Z'|A,Y}P_{Z|A,Y}}[\exp(s(z',z))] + 1,\end{aligned}$$

where the first equation follows Lemma A.4. Let $s^*(z',z)$ be the function when the supreme value is achieved and let $\hat{s}^*(z',z,a,y) = s^*(z',z), \forall\ (a,y) \in P_{A,Y}$, and we have

$$\begin{aligned}&D_{\mathrm{KL}}(P_{Z',Z} \,\|\, \mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}])\\ =\ &\sup_s \mathbb{E}_{(z',z,a,y)\sim P_{Z',Z,A,Y}}[s(z',z)] - \mathbb{E}_{(z',z,a,y)\sim P_{A,Y}P_{Z'|A,Y}P_{Z|A,Y}}[\exp(s(z',z))] + 1\\ =\ &\mathbb{E}_{(z',z,a,y)\sim P_{Z',Z,A,Y}}[\hat{s}^*(z',z,a,y)] - \mathbb{E}_{(z',z,a,y)\sim P_{A,Y}P_{Z'|A,Y}P_{Z|A,Y}}[\exp(\hat{s}^*(z',z,a,y)] + 1\\ \le\ &\sup_{\hat{s}} \mathbb{E}_{(z',z,a,y)\sim P_{Z',Z,A,Y}}[\hat{s}(z',z,a,y)] - \mathbb{E}_{(z',z,a,y)\sim P_{A,Y}P_{Z'|A,Y}P_{Z|A,Y}}[\exp(\hat{s}(z',z,a,y)] + 1\\ =\ &I(Z';Z \mid A,Y),\end{aligned}$$

where the last equation follows Lemma A.5. □

Figure 7: Proposition A.1 and its proof.

Proof. Define two probability measures $\mathcal{P} = P_{Z',Z}$ and $\mathcal{Q} = \mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}]$, we have

$$\begin{aligned}
&\mathbb{E}_{p(a,y)}\left[\mathbb{E}_{p(z_i',z_i|a,y)^{\otimes N}}\left[\log \frac{\exp(s(z_i', z_i))}{\frac{1}{N}\sum_{j=1}^{N}\exp(s(z_i', z_j))}\right]\right] \\
&= \mathbb{E}_{(z',z_1)\sim\mathcal{P},(z',z_{2:N})\sim\mathcal{Q}^{\otimes N-1}}\left[\log \frac{\exp(s(z', z_1))}{\frac{1}{N}\sum_{j=1}^{N}\exp(s(z', z_j))}\right] \\
&\leq D_{\mathrm{KL}}(\mathcal{P}\|\mathcal{Q}) \\
&= D_{\mathrm{KL}}(P_{Z',Z} \parallel \mathbb{E}_{P_{A,Y}}[P_{Z'|A,Y}P_{Z|A,Y}]) \\
&\leq I(Z'; Z \mid A, Y).
\end{aligned}$$

where the second equation follows Lemma A.3 and the last equation follows Proposition A.1. □

Figure 8: Proof of Proposition 3.2.

Table 1: Summary of training, validation, and test splits for the `jigsaw` dataset.

Data split	**Samples**	**Protected attribute** (NB = non-black, B = black)	**Task label average**
Training	25,954 (60.5%)	61.9% (NB), 38.1% (B)	0.2822
Validation	4,390 (10.2%)	62.4% (NB), 37.6% (B)	0.2897
Test	12,562 (29.3%)	61.2% (NB), 38.8% (B)	0.2873
Total	42,906	61.7% (NB), 38.3% (B)	0.2844

follow the same hyperparameters in training the linear guarding layer. The trade-off hyperparameter that we tune for INLP is N_{clf}, which is the number of classifiers trained by INLP (*i.e.*, the number of rounds).

Table 3 contains the trade-off hyperparameters used for our experiments on the `jigsaw` dataset, while Table 4 summarizes the trade-off hyperparameter choices for the `biasbios` dataset. The remaining hyperparamters for all methods are listed in Table 5.

B.3 Data Augmentation Strategies

In this section, we provide a description of the data augmentation strategies used in CL-based methods[5].

- **Easy data augmentation (EDA)(Wei and Zou, 2019).** EDA consists of four simple operations: synonym replacement, random insertion, random swap, and random deletion. Following the suggestions provided by the original paper, we choose the augmentation ratio to be 0.1 and create four augmented examples per example.
- **Back translation (Edunov et al., 2018).** It first translates the input example to another language and back to English. We use the machine translation model `wmt19-en-de` in our experiment.
- **Word replacement using contextual language model (CLM insert) (Kobayashi, 2018).** It replaces words based on a language model that leverages contextual word embeddings to find the most similar word for augmentation. We use the RoBERTa-base language model and choose the augmentation rate of 0.1.
- **Word insertion using contextual language model (CLM insert) (Kobayashi, 2018).** It inserts words based on a language model that leverages contextual word embeddings to find the most similar word for augmentation. We use the RoBERTa-base language model and choose the augmentation rate of 0.1.

[5]We use the implementations of data augmentation at: https://github.com/makcedward/nlpaug.

Table 2: Summary of the training, validation, and test splits of the biasbios dataset.

Data split	**Samples**	**Protected attribute** (F = female, M = male)
Training	255,710 (65.0%)	46.0% (F), 54.0% (M)
Validation	39,369 (10.0%)	47.8% (F), 52.2% (M)
Test	98,344 (25.0%)	46.5% (F), 53.5% (M)
Total	393,423	46.3% (F), 53.7% (M)

Table 3: Trade-off hyperparameters tested for RQ2 (Figure 3) for the jigsaw dataset.

Method	**Hyperparameters tested**
Adversarial training	$\lambda_{adv} \in \{0.1, 0.5, 1, 2\}$
Diverse adversarial training ($N_{adv} = 3$, $\lambda_{diff} = 100$)	$\lambda_{adv} \in \{0.1, 0.5, 1, 2\}$
INLP	$N_{clf} \in \{20, 50, 80, 100, 150\}$
One-stage CL	$\lambda \in \{0, 1, 2, 5\}$
Two-stage CL	$\lambda \in \{0, 0.1, 0.5, 1, 2, 5\}$

Table 4: Trade-off hyperparameters tested for RQ2 (Figure 3) for the biasbios dataset.

Method	**Hyperparameters tested**
Adversarial training	$\lambda_{adv} \in \{0.1, 0.2, 0.5, 1\}$
Diverse adversarial training ($\lambda_{diff} = 5000$)	$\lambda_{adv} \in \{0.1, 0.2, 0.5, 1\}$
INLP	$N_{clf} \in \{20, 50, 100, 300, 400\}$
One-stage CL	$\lambda \in \{0, 1, 5, 10\}$
Two-stage CL	$\lambda \in \{0, 1, 5, 10\}$

Table 5: Additional hyperparameters used for experiments.

Hyperparameter	(jigsaw)	(biasbios)
Batch size (fine-tuning)	32	32
Learning rate	2e-5	2e-5
Epochs	10	7
Optimizer	Adam	Adam

C Additional Experimental Results

C.1 More Comments for CL-based Methods

Our method achieves highly consistent results w.r.t. fairness and performance compared to the baseline methods. Figure 9 visualizes the model performance and EO gaps of two-stage CL under different hyperparameter settings when $\lambda \in \{0.0, 2.0\}$ in the jigsaw dataset.

C.2 Visualization of the BERT Embeddings using Different Objectives

In Figure 10, we show the T-SNE visualization (Van der Maaten and Hinton, 2008) of text embeddings learned with different training objectives. We can see that both CE-trained and CL-trained embeddings capture the class information well (points with the same markers form their own clusters). However, points with the same sensitive attributes (the same colors) within the same class are more likely to form small clusters. When we introduce $L_{\text{CS-InfoNCE}}$, those points tend to be more aligned.

C.3 How Different Pretrained Text Encoders affects the Performance of INLP?

To provide the clearest comparison between our proposed methods and the baselines, we used the

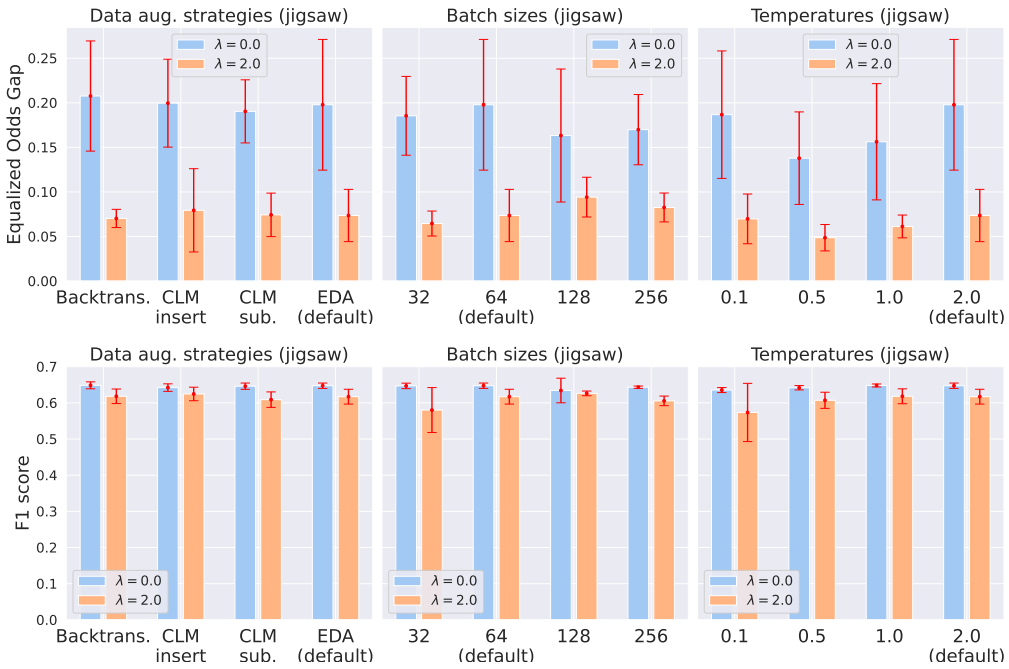

Figure 9: Sensitivity analysis of two-stage CL to key hyperparameter changes in (`Jigsaw`).

best settings for the baseline methods we could attain. Nonetheless, we observed that the performance of INLP was highly sensitive to the encoder training settings, which could be an important practical consideration for practitioners selecting between different ways of improving model fairness. Figure 11 compares the performance and fairness of INLP using different encoder pre-training strategies. We see that in both datasets, the classification and fairness performance of INLP changes drastically even with the same values of trade-off parameter. Even training with the same objectives (CE loss), the text encoders obtained in different epochs after convergence greatly affect its performance. For example, the CE-trained encoder obtained in the last epoch of training nearly shows no effects on bias mitigation. If we do not train the text encoder using our datasets and directly use the parameters of the `bert-base-uncased` (this is the experimental setting of the previous work (Ravfogel et al., 2020)), the model performances drastically decrease as the training iterations of INLP increase. Lastly, INLP does not perform well when using supervised contrastive loss to train the text encoder. In comparison, our methods are more robust to hyperparameter changes.

Figure 10: T-SNE visualization of text embeddings using different training objectives (zoom in for better visualization) in the `biasbios` dataset. Different colors indicate different sensitive attributes (e.g., red for males and green for females), and different markers indicate different classes. CE-trained and CL-trained embedding capture the class information well (points with the same markers form their own clusters). However, points with the same sensitive attributes within the same class are more likely to form small clusters. When we introduce $L_{\text{CS-InfoNCE}}$, those points tend to be more aligned.

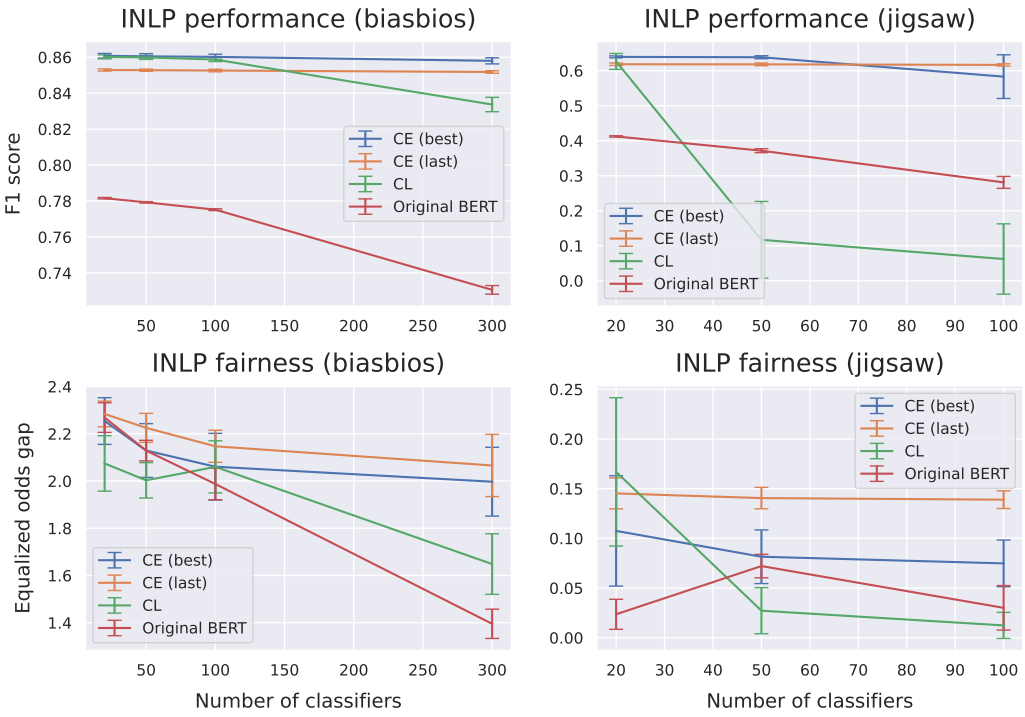

Figure 11: Comparison of INLP performance and fairness under different pretrained encoders. CE (best) indicates that we train the text encoder using CE loss and save the encoder in the epoch that achieves the best validation loss for INLP. CE (last) indicates that we train the text encoder using CE loss and save the encoder in the last epoch for INLP. CL indicates that we train the text encoder using supervised contrastive loss. Original BERT indicates that we use the `bert-based-uncased` as the text encoder.