Does Representational Fairness Imply Empirical Fairness?

Aili Shen[♠*] **Xudong Han**[♡] **Trevor Cohn**[♡] **Timothy Baldwin**[♡◇] **Lea Frermann**[♡]
♠ Amazon Alexa AI, Australia
♡ School of Computing and Information Systems, The University of Melbourne
◇ Department of Natural Language Processing, MBZUAI
aili.shen@amazon.com, xudongh1@student.unimelb.edu.au
{t.cohn,tbaldwin,lfrermann}@unimelb.edu.au

Abstract

NLP technologies can cause unintended harms if learned representations encode sensitive attributes of the author, or predictions systematically vary in quality across groups. Popular debiasing approaches, like adversarial training, remove sensitive information from representations in order to reduce disparate performance, however the relation between representational fairness and empirical (performance) fairness has not been systematically studied. This paper fills this gap, and proposes a novel debiasing method building on contrastive learning to encourage a latent space that separates instances based on target label, while mixing instances that share protected attributes. Our results show the effectiveness of our new method and, more importantly, show across a set of diverse debiasing methods that *representational fairness does not imply empirical fairness*. This work highlights the importance of aligning and understanding the relation of the optimization objective and final fairness target. *Our code is available at:* https://github.com/AiliAili/contrastive_learning_repo.

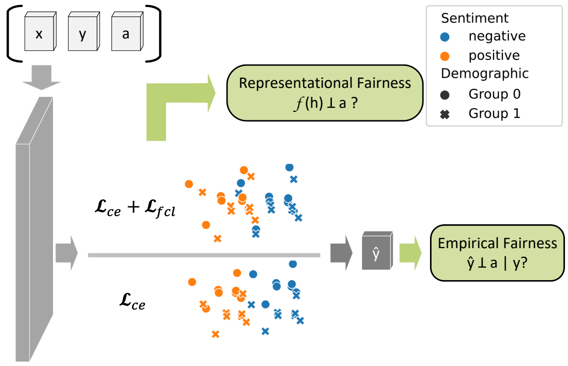

Figure 1: Illustration of our proposed method in the context of sentiment classification, where inputs (x) are mapped to hidden representations, which will then be used to make predictions $\hat{y}$. The points represent the instances in the latent space learned by a given model, marked with respect to sentiment and demographic labels. On the top and bottom of the gray line are hidden representations from our proposed method and a naively trained model. Representational fairness is measured based on the extent to which an attacker (f) can reconstruct protected attributes (a) from hidden representations (h). Empirical fairness measures performance disparities, and measures whether model predictions are independent of protected attributes.

1 Introduction

Neural methods have achieved great success for text classification tasks. However, they have been trained on datasets which embody cultural and societal stereotypes from the real world, captured in spurious correlations between target labels and protected attributes. This can result in biased predictions violating *empirical fairness*, i.e., models perform unequally for different sub-groups. A related, but different problem occurs if *representational fairness* is violated which means that learned representations encode potentially sensitive author information (such as demographic information), which can be recovered by an adversarial attacker. Addressing and reducing such cases of model bias has attracted substantial research interest across tasks including Twitter sentiment analysis (Blodgett et al., 2016; Han et al., 2021b), part-of-speech tagging (Hovy and Søgaard, 2015; Li et al., 2018), and image activity recognition (Wang et al., 2019; Zhao et al., 2017).

One line of work attempts to achieve empirical fairness through learning fair representations – removing authorship-related sensitive information from learned representations – under the assumption that fair representations will naturally lead to fairer models (Li et al., 2018; Ravfogel et al., 2020; Han et al., 2021a). For example, adversarial training is a popular method which directly aims to prevent a discriminator from reverse-engineering protected attribute information from learned rep-

[*] This work was done when Aili Shen was at The University of Melbourne.

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 81–95
November 20–23, 2022.

resentations (Elazar and Goldberg, 2018; Resheff et al., 2019; Han et al., 2021b,a; Li et al., 2018). Similarly, null-space projection approaches remove protected information from hidden representations by projecting learned text representations to the null-space of linear protected attribute discriminators (Ravfogel et al., 2020, 2022).

In this paper, we systematically explore the interaction between fair representations and empirical fairness, both via three classes of existing approaches, as well as in considering the application of contrastive learning (Oord et al., 2018; Li et al., 2021a; Tian et al., 2020; Henaff, 2020; Bui et al., 2021; Li et al., 2021b; Chen et al., 2020b) to fairness. Contrastive learning is a natural and flexible choice of approach for representational fairness, in explicitly differentiating representations between different classes. Representational fairness is achieved by learning a space which simultaneously separates instances according to their labels, while mixing instances with different protected attributes (like gender or race), either globally (Section 3.2) or per class (Section 3.3).

Our contributions in this work are:

1. We present two debiasing methods based on contrastive learning, with loss components that capture different fairness criteria;
2. Based on experimental results over Twitter sentiment analysis and profession classification, we show that our proposed method achieves the best representational fairness, where most baseline methods fail;
3. We show that there is no correlation between representational and empirical fairness, debunking previous assumptions about the empirical value of fair representations.

2 Related Work

We review relevant research on fairness criteria, debiasing methods, and contrastive learning.

2.1 Fairness Criteria

Various types of fairness have been proposed, such as group fairness (Hardt et al., 2016; Zafar et al., 2017a; Cho et al., 2020), individual fairness (Sharifi-Malvajerdi et al., 2019; Yurochkin et al., 2020; Dwork et al., 2012), and causality-based fairness (Garg et al., 2019; Wu et al., 2019; Zhang et al., 2018; Zhang and Bareinboim, 2018). In this work, we focus on group fairness relative to the demographic variables available in our datasets.

To quantify how the performance of models varies across different demographic subgroups, there are three widely used fairness criteria. *Demographic parity* (Feldman et al., 2015; Zafar et al., 2017b; Cho et al., 2020) measures whether the model achieves equal positive prediction rates towards each demographic subgroup, without taking the main task label into consideration. *Equal opportunity* (Hardt et al., 2016; Madras et al., 2018a) (Cho et al., 2020; Hardt et al., 2016; Madras et al., 2018a) requires equal true positive rates for instances from each subgroup conditioned on the main task label, while *equalised odds* requires equal true positive and false positive rates for instances from each subgroup and with the same main task label. The definition of these three criteria is limited to binary classification, whereas we extend the measurement of fairness to each main task label, such that bias is measurable in multi-class classification settings.

2.2 Achieving Empirical Fairness

To optimize towards group fairness, prior debiasing methods fall into three categories. *Pre-processing* manipulates the training data e.g., by balancing the input, followed by re-training the model on a fairer dataset (Badjatiya et al., 2019; Elazar and Goldberg, 2018) but is computationally prohibitive for large datasets and models, and insufficient to ensure fairness (De-Arteaga et al., 2019; Wang et al., 2019). *Post-processing* methods "bleach" sensitive information from learned representations after main task training (Ravfogel et al., 2020). *In-processing* approaches augment the original training objective, to encourage the model to learn representations that are oblivious to protected attributes, aiming to achieve empirical fairness through representational fairness. For example, adversarial models (Beutel et al., 2017; Li et al., 2018; Barrett et al., 2019; Han et al., 2021b) encourage the main model to learn representations that are indistinguishable wrt the protected attributes by a jointly trained discriminator. Our contrastive learning methods also introduce an augmented objective, but unlike adversarial methods, do not require modification of the model architecture, and hence do not add model parameters. Tsai et al. (2021) proposed a similar approach in a self-supervised learning setting.

Other methods directly optimize fairness measures during training (Madras et al., 2018b; Zhao et al., 2020a; Cho et al., 2020). For example, Cho et al. (2020) use kernel density estimation to approximate equalised odds during training, but tailored to binary classification, leading to poor performance–fairness tradeoffs in high-dimensional settings. We introduce two variants of the contrastive losses which directly optimize fairness for demographic parity or equal opportunity, respectively.

Various recent studies (Ravfogel et al., 2020; Han et al., 2021b; Chi et al., 2022; Zhao et al., 2020b; Chowdhury et al., 2021; Tsai et al., 2021; Zhao and Gordon, 2019) claimed to generate fair representations, while exclusively evaluating their methods in terms of empirical fairness. Other work has used metrics like representation leakage to quantify how much protected attribute information can be recovered from learned representations (Han et al., 2021b; Elazar and Goldberg, 2018; Li et al., 2018; Wang et al., 2019). However, it has not been systematically studied whether fair representations lead to fair predictions, which is one contribution of this paper.

2.3 Contrastive Learning

Contrastive learning aims to pull similar instances together and push dissimilar instances apart by maximizing the similarities of similar instances and minimizing those of dissimilar pairs within the unit feature space (Oord et al., 2018; Tian et al., 2020; Li et al., 2021a; Grill et al., 2020; Chen et al., 2020a; Henaff, 2020). Its success hinges on an appropriate definition of similarity. Originating in computer vision, in vanilla contrastive learning positive (similar) instance image pairs are generated via data augmentation (i.e., meaning-invariant manipulation of an input image such as cropping or blurring (Chen et al., 2020a; Fang et al., 2020; Cubuk et al., 2019)), and negative (dissimilar) instance pairs correspond to distinct items in the original data. More recently, supervised contrastive learning (SCL) was proposed in the context of classification, where positive instances belong to the same class, and negative instances belong to different classes (Khosla et al., 2020). When combined with a cross entropy loss, it has been shown to improve model robustness to noise and data sparsity (Gunel et al., 2021), as well as adversarial attacks (Bui et al., 2021). We leverage the ability of SCL to explicitly constrain class-based positioning of instances in feature space, to enforce representational fairness. We present evidence of its effectiveness, and use it to systematically study the relationship between representational and empirical fairness.

The most relevant work to our proposed method is Gupta et al. (2021), whose training objective consists of three parts: (1) cross-entropy loss, which is identical to vanilla training; (2) upper bound for the mutual information between inputs and hidden representations, which relies on a manually-defined prior over the hidden representations to calculate a KL divergence loss; and (3) lower bound estimator for the conditional mutual information, similar to $\mathsf{Con}_{\mathrm{eo}}$ in our paper (see Equation (3)). Although Gupta et al. (2021) have the same cross-entropy objective and lower-bound estimation as the equal opportunity variant of our proposed method, its second objective (upper bound estimator) focuses on learning task-agnostic representations while ours learns task-specific representations. Moreover, in this paper, we also show that the demographic parity variant consistently outperforms the equal opportunity variant.

2.4 Intrinsic Fairness

Intrinsic bias refers to biases in the geometry of text representations in upstream pre-trained language models (prior to any task-specific fine-tuning). Such representations are agnostic to downstream tasks, and common metrics for intrinsic biases rely on predefined templates, e.g., gendered word pairs for word embedding association test (Caliskan et al., 2017) and masked sentences (Kurita et al., 2019).

There is a broad range of studies on the correlation between intrinsic and extrinsic bias (Goldfarb-Tarrant et al., 2021; Cao et al., 2022). Jin et al. (2020) show that debiasing the intrinsic bias leads to less extrinsic bias, but conversely, Steed et al. (2022) argue that extrinsic bias is better explained by bias in downstream datasets rather than intrinsic bias in upstream text representations. Similar to this paper, Orgad et al. (2022) examine the influence of downstream task debiasing on representations. However, it also focuses exclusively on intrinsic bias rather than representational fairness. In summary, most previous work is aimed at measuring and mitigating task-agnostic *intrinsic* bias.

In contrast, the leakage metric for representational fairness in this paper is task-specific, and measures the predictability of protected information from the task-specific representations that are learned as part of fine-tuning. Given that both leakage (intrinsic) and empirical fairness (extrinsic) are defined in a task-specific way, we expect a stronger correlation between the two. This expectation is at the core of common debiasing approaches, such as adversarial methods. To the best of our knowledge, this paper is the first to explore this correlation.

3 Fair & Supervised Contrastive Learning

Our method augments the objective of supervised contrastive learning to simultaneously encourage data separation in terms of the main class labels, and discourage the differentiation of data points on the basis of their protected attributes. While the method is compatible with different classifier architectures, here we use the following setup:

1. An *embedding module*, $\boldsymbol{e} = \texttt{Embed}(\boldsymbol{x})$, which maps an input instance $\boldsymbol{x}$ (e.g., a document) to a vector representation $\boldsymbol{e}$, which is in turn used as input to the encoder network;
2. An *encoder network*, $\boldsymbol{h} = \texttt{Enc}(\boldsymbol{e})$, which maps the input representation to the final hidden representation;
3. An *aggregated objective* ($\mathcal{L}_*$), which is a weighted combination of a cross-entropy loss, contrastive loss based on main task labels, and contrastive loss based on protected attribute labels, as described next.

3.1 Contrastive Loss

Given a mini-batch with a set of N randomly sampled instances, positive instance pairs (representing the same concept) and negative instance pairs (representing different concepts) are formed. These pairs can be created based on either their main task label or their protected attribute, as described below. Assuming a batch of positive and negative pairs, the contrastive loss is computed as,

$$\mathcal{L}_{\text{scl}} = \sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_p / \tau)}{\sum_{q \in Q(i)} \exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_q / \tau)},$$

where $i{=}1 \dots N$ is the index of an instance in the mini-batch; $Q(i) \equiv \{1 \dots N\} \setminus \{i\}$; $\boldsymbol{h}_i = l_2(\texttt{Enc}(\texttt{Embed}(\boldsymbol{x}_\texttt{i})))$ is the normalised representation; and $\tau{>}0$ is a scalar temperature parameter controlling smoothness. $P(i)$ is the set of instances that result in positive pairs for the ith instance, and $|P(i)|$ is its cardinality. We next describe how positive/negative pairs are created.

For ease of illustration, we overload the definition of $\mathcal{L}_{\text{scl}}$ as an function, i.e.,

$$\mathcal{L}_{\text{scl}} = \mathcal{L}_{\text{scl}}(\boldsymbol{h}; \tau; P(\cdot); Q(\cdot)), \tag{1}$$

where $P(\cdot)$ is the set of indices of positive samples, and $Q(\cdot)$ is the set of sample indices that are considered in the contrastive loss.

$\mathcal{L}_{\text{scl}}$ is computed on positive and negative samples constructed based on main task labels (e.g., POS vs. NEG sentiment), where instances in the mini-batch belonging to the same main task class are used to construct positive samples; otherwise, they are used to form negative samples. The intuition behind this loss component is that representations that are well-separated for the main task are more desirable.

3.2 Fair Contrastive Learning for Demographic Parity

Demographic parity is satisfied if predictions are independent of protected attributes, i.e., $\Pr(\hat{y}{=}1|a{=}0) = \Pr(\hat{y}{=}1|a{=}1)\ \forall y{\in}Y, a{\in}A$, where Y is the main task label set and A is the protected attribute value set. With fair contrastive learning, the training objective for demographic parity ($\mathcal{L}_{\text{fcl-dp}}$) is to infer latent representations which are oblivious to the protected attribute of an instance. We create samples with respect to protected attribute labels (e.g., $a =$ MALE vs. $a =$ FEMALE), where instances of the same protected attribute class form positive samples; otherwise, they constitute negative samples:

$$\mathcal{L}_{\text{fcl-dp}} = -1 \times \mathcal{L}_{\text{scl}}(\boldsymbol{h}; \tau; P_{\text{fcl-dp}}(\cdot); Q(\cdot)),$$

where $P_{\text{fcl-dp}}(i) \equiv \{p \in Q(i) : a_p = a_i\}$ constructs positive samples based on protected attributes rather than target classes in supervised contrastive learning (Equation (1)). Importantly, the -1 changes the sign of supervised contrastive loss, enforcing representations of instances with different protected attribute values to mix together by discouraging the model from effectively contrasting those instances.

The final classifier objective produces task-indicative and protected-attribute-agnostic representations, as the weighted sum of standard cross-entropy loss $\mathcal{L}_{\text{ce}}$, and contrastive loss terms $\mathcal{L}_{\text{scl}}$, and $\mathcal{L}_{\text{fcl-dp}}$,

$$\mathcal{L}_{\text{dp}} = \mathcal{L}_{\text{ce}} + \alpha \mathcal{L}_{\text{scl}} + \beta \mathcal{L}_{\text{fcl-dp}} \tag{2}$$

where α and β are hyperparameters that control the relative importance of the cross entropy and contrastive learning terms. We refer to the contrastive classifier based on the loss in Equation (2) as $\mathsf{Con}_{\mathrm{dp}}$.

3.3 Fair Contrastive Learning for Equal Opportunity

A model is fair wrt equal opportunity (Hardt et al., 2016) if instances from different groups *within the same class* are treated equally, i.e., $\Pr(\hat{y} = y|Y = y, a{=}0) = \Pr(\hat{y}{=}y|Y{=}y, a{=}1)\ \forall y{\in}Y, a{\in}A$, connecting directly to the widely-used fairness metric GAP (see Section 4.2).

Accordingly, we construct samples in terms of protected attribute labels conditioned on the main task labels, and compute $\mathcal{L}_{\text{fcl-eo}}$ as the average loss over labels,

$$\mathcal{L}_{\text{fcl-eo}} = \frac{-1}{|Y|}\sum_{y\in Y}\mathcal{L}_{\text{scl}}(\boldsymbol{h};\tau;P_{\text{fcl-eo}}(\cdot);Q_{\text{fcl-eo}}(\cdot)),$$

where $Q_{\text{fcl-eo}}(i,y)) \equiv \{q|q \in 1,\ldots,N, y_q = y, \text{and } q \neq i\}$ ensures that contrastive losses are calculated per class, and $P_{\text{fcl-eo}}(i,y) \equiv \{p \in Q_{\text{fcl-eo}}(i,y) : a_p = a_i\}$ constructs positive samples based on protected attributes from a particular main task class y. Optimizing for $\mathcal{L}_{\text{fcl-eo}}$ minimizes mutual information between instances from different protected groups within each target class.

Analogous to Equation (2), we define a fair classifier objective wrt equal opportunity as,

$$\mathcal{L}_{\text{eo}} = \mathcal{L}_{\text{ce}} + \alpha\mathcal{L}_{\text{scl}} + \beta\mathcal{L}_{\text{fcl-eo}}. \tag{3}$$

We refer to contrastive classifiers based on the loss in Equation (2) as $\mathsf{Con}_{\mathrm{eo}}$.

3.4 Remarks

Non-binary protected attributes: Our $\mathcal{L}_{\text{fcl-dp}}$ and $\mathcal{L}_{\text{fcl-eo}}$ extend to non-binary protected attributes by sampling negative instances at random from any alternative subgroup.

Loss component weights: The same value is adopted for α and β for both $\mathcal{L}_{\text{scl}}$ and $\mathcal{L}_{\text{fcl-dp}}$/$\mathcal{L}_{\text{fcl-eo}}$ as they are similar in concept and magnitude, and weighting them equally balances performance with bias reduction, as confirmed in extensive preliminary experiments.

Relation to mutual information: Optimizing contrastive loss is equivalent to maximizing mutual information between classes (Oord et al., 2018; Khosla et al., 2020). Conversely, in representational fairness, representations $\boldsymbol{h}$ should be independent of protected attributes a, i.e., minimise mutual information between $\boldsymbol{h}$ and a. $\mathcal{L}_{\text{fcl-dp}}$ and $\mathcal{L}_{\text{fcl-eo}}$ intuitively satisfy this by flipping the sign of the contrastive objective.

4 Experiments

In this section, we report experimental results for bias mitigation. All experiments are conducted with the *fairlib* library (Han et al., 2022b), and full experimental details are provided in Appendix D.

4.1 Comparison Models

We evaluate the utility of contrastive fairness, and systematically study the relation between representational and empirical fairness. To do so, we include competitive debiasing methods covering *pre-*, *in-*, and *post-processing*:

1. CE: train $\texttt{Enc}(\cdot)$ with cross-entropy loss. No bias mitigation.
2. INLP: train $\texttt{Enc}(\cdot)$ with cross-entropy loss, and apply iterative null-space projection (Ravfogel et al., 2020) to the learned representations. Specifically, a linear discriminator is iteratively trained over the protected attribute to project the representation onto the discriminator's null-space, thereby reducing protected attribute information from the representations.
3. Adv: jointly train $\texttt{Enc}(\cdot)$ with cross-entropy loss and an ensemble of 3 adversarial discriminators over the protected attribute, with an orthogonality constraint applied to each pair of sub-discriminators to encourage them to learn different aspects of the representations (Han et al., 2021b). The $\texttt{Enc}(\cdot)$ is trained to prevent protected attributes from being identified, and thus results in fairer representations.
4. FairBatch: formulate the model training as a bi-level optimization problem, which minimises prediction disparities through adjusting resampling probabilities (Roh et al., 2021).
5. $\mathsf{EO}_{\mathrm{GLB}}$: optimize equal opportunity through proxy objective functions based on group-specific cross-entropy, which essentially adjusts instances weights in training (Shen et al., 2022).
6. Gate: use demographic information to make predictions, with balanced training as regularizers in training to avoid learning spurious correlations (Han et al., 2022a). Unlike the afore-

mentioned models, which aim to reduce both representational and empirical bias, Gate is expected to be high in representational bias and low in empirical bias.

In summary, we incorporate three types of baselines: (1) INLP and Adv remove protected information from hidden representations to mitigate representational bias, which is similar to our contrastive learning methods; (2) FairBatch and EO_{GLB} mitigate empirical bias based on model predictions, without considering representational fairness; and (3) Gate uses protected information explicitly to make fair predictions, explicitly violating representational fairness.

4.2 Evaluation Metrics

Following Ravfogel et al. (2020), we adopt **Accuracy** for both the binary and multi-classification datasets to evaluate the performance of models on the main task, and measure empirical fairness based on equal opportunity in terms of the model predictions. To measure representational fairness, we follow Elazar and Goldberg (2018) in measuring protected attribute leakage in text representations.

To measure **empirical fairness**, we adopt equal opportunity, which measures the difference in true positive rate (TPR) between binary protected attribute a and $\neg a$ (such as FEMALE vs. MALE) for each main task class. It is defined as $\text{GAP}^{\text{TPR}}_{a,y} = |\text{TPR}_{a,y} - \text{TPR}_{\neg a,y}|, y \in Y$, where $\text{TPR}_{a,y} = \mathbb{P}\{\hat{y} = y|y, a\}$. Here $\hat{y}$ and y are the predicted and gold-standard main task labels; and Y is the set of main task labels. $\text{TPR}_{a,y}$ measures the percentage of correct predictions among instances with main task label y and protected attribute a. $\text{GAP}^{\text{TPR}}_{a,y}$ measures the absolute difference between the two different groups represented by the protected attribute, given the main task class y. To take all target classes into consideration, we follow De-Arteaga et al. (2019) and Ravfogel et al. (2020) in calculating the root mean square of $\text{GAP}^{\text{TPR}}_{a,y}$ over all classes $y \in Y$, to get a single score:

$$\text{GAP} = \sqrt{\frac{1}{|Y|}\sum_{y \in Y}(\text{GAP}^{\text{TPR}}_{a,y})^2}$$

A difference of 0 indicates a fair model, as the prediction $\hat{y}$ is conditionally independent of protected attribute a. For ease of exposition, we report the equal opportunity fairness (Fairness) as $1 - \text{GAP}$, where larger is better and a perfectly fair model will achieve a fairness score of 1.

Distance to the optimum (DTO) has been used to simplify model comparisons in previous work (Marler and Arora, 2004; Han et al., 2022a), which measures the Euclidean distance from a particular model to the optimum point (aka "Utopia" point), usually set to 100% accuracy and 100% equal opportunity fairness, denoting the best possible values. While the dimensions of the space are performance and fairness, DTO explicitly reflects the performance-fairness trade-off of a model. We calculate DTO based on empirical fairness, and perform model selections based the smallest DTO over the development set (Han et al., 2022a).

Representational Fairness is evaluated through **Leakage** as the ability of an attacker to recover the protected attribute from a model's final hidden representations. We train one attacker (i.e., neural network) for each model, to extract information of protected attributes from a model's final-layer hidden representations (Wang et al., 2019; Han et al., 2021b). We fix the attacker architecture across models, so that attackers are not guaranteed to be optimal and leakage estimators should be interpreted as lower bounds.[1]

4.3 Experiment 1: Sentiment Analysis

4.3.1 Task and Dataset

The task is to predict the binary sentiment for a given English tweet, based on the dataset of Blodgett et al. (2016) (**Moji** hereafter), where each tweet is also annotated with a binary private attribute indirectly capturing the ethnicity of the tweet author as either African American English (AAE) or Standard American English (SAE). Following previous studies (Ravfogel et al., 2020; Han et al., 2021b), the training dataset is balanced with respect to both sentiment and ethnicity but skewed in terms of sentiment–ethnicity combinations (40% HAPPY-AAE, 10% HAPPY-SAE, 10% SAD-AAE, and 40% SAD-SAE, respectively).[2] The dataset contains 100K/8K/8K train/dev/test instances.

4.3.2 Implementation Details

Following previous work (Elazar and Goldberg, 2018; Ravfogel et al., 2020; Han et al., 2021b), we

[1]Preliminary analyses revealed that non-linear attackers outperform linear ones in recovering protected attributes, and attackers with different non-linear architectures have similar capacity to recover protected attribute information from representations. We use non-linear MLPs as our attacker. Further details are in Appendix A.

[2]Note that the dev and test set are balanced in terms of sentiment–ethnicity combinations.

Model	Accuracy ↑	Fairness ↑	DTO↓	Leakage ↓
CE	72.3±0.5	61.2±1.4	47.7	87.9±3.3
INLP	73.3±0.0	85.6±0.0	30.3	86.7±0.6
Adv	75.6±0.4	90.4±1.1	26.3	78.8±6.0
Gate	**76.2±0.3**	90.1±1.5	**25.8**	100.0±0.0
FairBatch	75.1±0.6	**90.6±0.5**	26.7	88.4±0.4
EO_{GLB}	75.2±0.2	90.1±0.4	26.7	85.7±1.2
Con_{dp}	75.8±0.3	88.1±0.6	26.9	**54.2±0.9**
Con_{eo}	74.1±0.7	84.1±3.0	30.3	80.1±4.2

Table 1: Experimental results on **Moji** (averaged over 5 runs). The best result for each metric is indicated in **bold**. Here, ↑ and ↓ indicate that higher and lower performance, resp., is better for the given metric.

Model	Accuracy ↑	Fairness ↑	DTO↓	Leakage ↓
CE	82.3±0.2	85.1±0.8	23.2	98.0±0.0
INLP	82.3±0.0	88.6±0.0	21.0	97.6±0.1
Adv	81.9±0.2	**90.6±0.5**	20.4	88.6±4.6
Gate	**83.7±0.2**	90.4±0.9	**18.9**	100.0±0.0
FairBatch	82.2±0.1	89.5±1.3	20.6	98.0±0.3
EO_{GLB}	81.7±0.4	88.4±1.0	21.7	97.2±0.5
Con_{dp}	82.1±0.2	84.3±0.8	23.9	**76.3±1.5**
Con_{eo}	81.8±0.3	85.2±0.4	23.5	84.9±3.4

Table 2: Experimental results on **Bios** (averaged over 5 runs).

use DeepMoji (Felbo et al., 2017), a model pretrained over 1.2 billion English tweets, as `Embed`$(\cdot)$ to obtain text representations. The parameters of DeepMoji are fixed in our experiments.

4.3.3 Results

Table 1 presents the results. Our proposed methods achieve competitive empirical fairness results with other debiasing methods, all of which improve over CE. Adv, Gate, FairBatch, and EO_{GLB} achieve the best performance in terms of Fairness, while our proposed method Con_{dp} achieves the best performance in terms of Leakage. Specifically, none of the baselines reduce leakage substantially except for Adv. The reason that Adv can reduce Leakage is that the architecture of Adv is the closest one to the leakage estimation framework, which also employs attackers to extract protected attributes and unlearns attackers in training. However, Con_{dp} still outperforms Adv, highlighting the effectiveness of our proposed method in improving representational fairness. The ineffectiveness of INLP, Gate, FairBatch, and EO_{GLB} in reducing Leakage is due to different reasons: INLP is due to the fact that it relies on linear projections to remove protected attribute information and is ineffective at removing nonlinear correlations; Gate is due to the fact that it employs a gate mechanism to augment text representations with protected information, and as a result, achieves 100% Leakage; and both FairBatch and EO_{GLB} are due to the fact that these two methods are optimized to directly mitigate empirical bias without considering representational bias. This indicates that the relationship between representational fairness and empirical fairness is not as simple as suggested in previous work (Elazar and Goldberg, 2018; Ravfogel et al., 2020; Han et al., 2021b)

Con_{eo}, which is proposed to ensure conditional representational fairness within each class, achieves similar prediction fairness to Con_{dp}, but with much worse leakage. This further shows that representational fairness cannot be directly linked to prediction fairness. It is encouraging to see that incorporating debiasing techniques can contribute to improvement on the main task. We hypothesise that incorporating debiasing techniques (either in the form of adversarial training or contrastive loss) acts as a form of regularisation, leading to greater robustness over the training dataset skew relative to the unbiased test set.

4.4 Experiment 2: Profession Classification

4.4.1 Task and Dataset

The task is to predict a person's profession given their biography, based on the dataset of De-Arteaga et al. (2019) (**Bios** hereafter), consisting of short online biographies which have been labelled with one of 28 professions (main task label) and binary gender (protected attribute). We use the dataset split of (De-Arteaga et al., 2019; Ravfogel et al., 2020), consisting of 257K/40K/99K train/dev/test instances.[3]

4.4.2 Implementation Details

Following the work of Ravfogel et al. (2020), we use the `[CLS]` token representation of the pretrained uncased BERT-base (Devlin et al., 2019) as `Embed`$(\cdot)$, without any further finetuning.

4.4.3 Results

Table 2 shows the results on the test set. In terms of prediction fairness, baseline methods achieve similar results, however, both Con_{dp} and Con_{eo} are less effective for improving prediction fairness. We hypothesise that this is because of the multi-class setting (28 classes), where the large number

[3]There are slight differences between our dataset and that used by De-Arteaga et al. (2019) and Ravfogel et al. (2020) as a small number of biographies were no longer available on the web when we scraped them.

of main task classes impedes the ability of contrastive learning to learn representations that jointly maximize mutual information for main task classes and minimize mutual information for demographic labels. In Section 4.5, we conduct ablation studies to analyse their robustness to the number of classes, affirming our explanation. In terms of the representational fairness, consistent with the results over **Moji**, $\mathsf{Con}_{\mathrm{dp}}$ and $\mathsf{Con}_{\mathrm{eo}}$ substantially reduce Leakage, where most baselines fail.

Overall, the trend for these three types of methods over the **Bios** dataset is consistent with that over the **Moji** dataset: (1) INLP and Adv, which focus on representational fairness, result in empirical fairness improvements and marginal gain in Leakage; (2) FairBatch and $\mathsf{EO}_{\mathrm{GLB}}$, which target for empirical fairness, lead to fairer predictions but no benefit to Leakage; and (3) Gate, which augments representations with protected information, also improves empirical fairness while suffering from 100% Leakage. Based on the consistent trend over two benchmark datasets, we argue that it cannot be assumed that empirical fairness is associated with representational fairness, with the fact that $\mathsf{Con}_{\mathrm{dp}}$ and $\mathsf{Con}_{\mathrm{eo}}$ achieve the best representational fairness but lowest empirical fairness further adding weight to this argument.

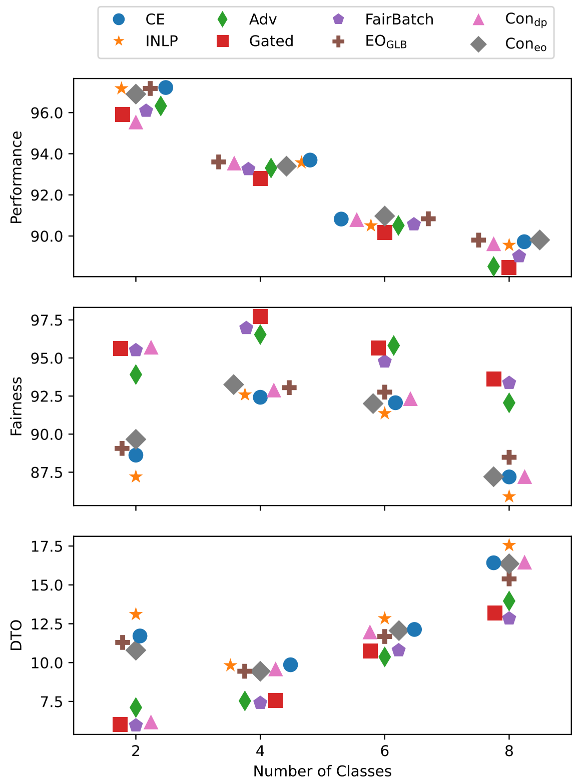

Figure 2: Varying the number of classes in the **Bios** dataset. We treat the number of classes as a categorical variable, and draw categorical scatter plots with non-overlapping points.

4.5 Analysis

Robustness to the Number of Classes Our proposed methods are quite effective over **Moji** but not competitive over **Bios** in terms of Fairness. We hypothesize that this is due to contrastive loss struggling with a larger number of classes. To verify this, we construct 4 synthetic datasets from **Bios** by selecting a subset of classes from 2 to 8.[4]

Figure 2 presents Accuracy, empirical Fairness, and DTO with respect to 2, 4, 6, and 8 target classes. Although the scores with respect to different numbers of classes are not directly comparable as we also have to vary the number of classes in the test set, resulting in different test sets, it is reasonable to compare the trend of changes in the rank of debiasing methods.

Overall, increasing the number of classes leads to a decrease in Accuracy while Fairness is almost unchanged. As a result, the trade-off between Accuracy and Fairness (DTO) drops. In terms of Accuracy, $\mathsf{Con}_{\mathrm{dp}}$ and $\mathsf{Con}_{\mathrm{eo}}$ achieve competitive performance with other debiasing methods, all of which are slightly worse than CE.

Looking at empirical Fairness, $\mathsf{Con}_{\mathrm{dp}}$ achieve quite competitive performance when the number of target classes is 2, while $\mathsf{Con}_{\mathrm{eo}}$ is unable to significantly improve Fairness. This is consistent the results over the binary classification dataset (**Moji**). For other settings (4, 6, and 8 target classes), $\mathsf{Con}_{\mathrm{eo}}$ shows better trade-offs than $\mathsf{Con}_{\mathrm{dp}}$. However, both $\mathsf{Con}_{\mathrm{dp}}$ and $\mathsf{Con}_{\mathrm{eo}}$ only achieve slight improvements in Fairness, and are not as good as some other debiasing methods.

To conclude, the changes in DTO confirm our hypothesis that contrastive loss struggles with a larger number of classes: contrastive loss achieves one of the best DTO for 2 classes, competitive results with other debiasing methods for 4 and 6 classes, and the worst DTO for 8 classes.

Correlation between Representational and Empirical Fairness Although we have discussed the connection between representational and empirical fairness for individual methods, it is still not clear how they are correlated.

[4] In Appendix C.1, we provide the full details of the synthetic datasets.

For each method, we have 5 random runs, and in total, there are 5 groups of methods: (1) CE; (2) INLP and Adv; (3) FairBatch and EO_{GLB}; (4) Gate; and (5) Con_{dp} and Con_{eo}. To treat each group of methods equally, we fit a bivariate Gaussian distribution to each method over the 5 runs, and draw 20k random samples from each group for a given dataset.

Based on the random samples, the Pearson correlation coefficients between representational and empirical fairness over **Moji** and **Bios** are 0.072 and -0.222, respectively. Clearly, both correlation coefficients are not substantially better than 0, indicating that there is little to no linear dependency between representational fairness and empirical fairness. Even more damningly, the negative sign for the **Bios** suggests that worse representational fairness may result in higher empirical fairness.

Clearly further work is required to examine the theoretical difference/connection between representational and empirical fairness, which we leave to future work.

5 Conclusion

Biased representations and predictions can reinforce existing societal biases and stereotypes. While previous work has assumed a direct link between biases in the representations learned by models and performance disparities in model predictions, there has not been a systematic study of the relationship between the two. We have explored the relationship wrt both a range of existing methods and two newly-proposed methods based on supervised contrastive learning. The contrastive learning methods are based on the intuition that similar instances belonging to the same main task class should be pulled together and similar instances belonging to the same protected attribute class should be pushed apart in the representation space, based on which we proposed to combine cross-entropy loss with two contrastive loss components in optimizing neural networks in two different ways, incorporating demographic parity and equal opportunity respectively. Experimental results over two tasks demonstrate the effectiveness of the proposed methods in terms of representation fairness, but further analysis showed no meaningful correlation between representational fairness and empirical fairness, contradicting a common assumption made in prior research, and motivating future work on approaches that achieve both representational and empirical fairness.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback and suggestions. This work was funded by the Australian Research Council, Discovery grant DP200102519. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

Limitations

A limitation of our proposed methods is that we focus on learning fair representations for the main task, where the protected attribute is explicitly present in the dataset. The mitigation of biases present only implicitly, such as protected information revealed in the text rather than indicated by demographics, as studied by Lahoti et al. (2020), is out scope of our work. For main tasks other than classification, such as generation tasks, adoption of contrastive learning for generating fairer text is not trivial, which is one direction for future work. In our work, $\texttt{Embed}(\cdot)$ is not learned or finetuned together with $\texttt{Enc}(\cdot)$ and the classification layer in an end-to-end fashion. However, finetuning the $\texttt{Embed}(\cdot)$ has the potential for better task-specific or semantic-preserving representations of text, which may further remove biases encoded in pretrained models. One simplifying assumption in our work is that we focus exclusively on binary protected attributes, implying the adoption of an oversimplified binary notion of gender. Exploring attributes of higher arity, and more complex and realistic bias dimensions, is an important direction for future work.

Ethical Considerations

We propose Con_{dp} and Con_{eo} to prevent text classifiers from encoding protected information. However, there is a possibility that multiple protected attributes, such as gender, age, and ethnicity, are encoded in text and the dataset is annotated only wrt one of the protected attribute. Therefore, a method designed to alleviate a specific type of bias is not guaranteed to be bias-free. The usage of our fair classifiers in the real world should be carefully monitored with the aid of domain experts.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6330–6335.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Anh Bui, Trung Le, He Zhao, Paul Montague, Seyit Camtepe, and Dinh Phung. 2021. Understanding and achieving efficient robustness with adversarial supervised contrastive learning. *arXiv preprint arXiv:2101.10027*.

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

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*.

Jianfeng Chi, William Shand, Yaodong Yu, Kai-Wei Chang, Han Zhao, and Yuan Tian. 2022. Conditional supervised contrastive learning for fair text classification. *arXiv preprint arXiv:2205.11485*.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. 2020. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems*.

Somnath Basu Roy Chowdhury, Sayan Ghosh, Yiyuan Li, Junier Oliva, Shashank Srivastava, and Snigdha Chaturvedi. 2021. Adversarial scrubbing of demographic information for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 550–562.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123.

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*, pages 120–128.

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.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012*, pages 214–226.

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.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

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

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*.

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*.

Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. 2021. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7610–7619.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021a. Decoupling adversarial training for fair NLP. In *Findings of the Association for Computational Linguistics*, pages 471–477.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021b. 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.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022a. Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*. To appear.

Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. 2022b. fairlib: A unified framework for assessing and improving classification fairness. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022) Demo Session*. To appear.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.

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

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 2: Short papers)*, pages 483–488.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2020. On transferability of bias mitigation effects in language model fine-tuning. *arXiv preprint arXiv:2010.12864*.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

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

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021a. Prototypical contrastive learning of unsupervised representations. In *Proceedings of the 9th International Conference on Learning Representations*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 25–30.

Yunfan Li, Peng Hu, Jerry Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021b. Contrastive clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 8547–8555.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018a. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3381–3390.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018b. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3381–3390.

R Timothy Marler and Jasbir S Arora. 2004. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6):369–395.

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

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. 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.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. *arXiv preprint arXiv:2201.12091*.

Yehezkel S. Resheff, Yanai Elazar, Moni Shahar, and Oren Sar Shalom. 2019. Privacy and fairness in recommender systems via adversarial training of user representations. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, pages 476–482.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. Fairbatch: Batch selection for model fairness. In *Proceedings of 9th International Conference on Learning Representations*.

Saeed Sharifi-Malvajerdi, Michael J. Kearns, and Aaron Roth. 2019. Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing Systems*, pages 8240–8249.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Optimising equal opportunity fairness in model training. *arXiv preprint arXiv:2205.02393*.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Proceeding of the 16th European Conference on Computer Vision*, pages 776–794.

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*.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319.

Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1438–1444.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2020. Training individually fair ML models with sensitive subspace robustness. In *Proceedings of the 8th International Conference on Learning Representations*.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017b. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 962–970.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making - the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2037–2045.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. 2020a. Conditional learning of fair representations. In *International Conference on Learning Representations*.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. 2020b. Conditional learning of fair representations. In *International Conference on Learning Representations*.

Han Zhao and Geoffrey J. Gordon. 2019. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, pages 15649–15659.

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.

# L	D	AF	Moji	Bios
1	–	–	84.80±0.54	96.63±0.03
2	100	Tanh	87.12±0.51	97.91±0.03
2	100	ReLU	87.03±0.34	97.92±0.04
2	300	Tanh	87.37±0.13	98.00±0.03
2	300	ReLU	87.89±0.34	97.96±0.05
4	100	Tanh	87.21±0.57	97.84±0.10
4	100	ReLU	87.38±0.70	97.82±0.06
4	300	Tanh	87.42±0.45	97.90±0.05
4	300	ReLU	87.50±0.29	97.86±0.04

Table 3: Leakage estimations over **Moji** and **Bios** with respect to different attacker architectures. **# L**, **D**, and **AF** denote number of hidden layers, hidden dimensions, and activation functions, respectively. Leakage estimation statistics (mean and standard deviation) are calculated over 5 runs.

A Robustness to Leakage Estimation

To analyse the robustness of leakage estimations, we vary attacker architectures and compare estimated leakage of the CE model. Table 3 summaries results over the **Moji** and **Bios** datasets

Overall, leakage estimations are robust to different architectures, except the results of linear attackers (i.e., 1 layer), which are consistently worse over both datasets.

In terms of the standard deviation, the training set of **Bios** is larger than that of **Moji** (205k v.s. 100k), resulting in a smaller standard deviation for leakage estimations over **Bios** than **Moji**.

B Adv Settings

Each sub-discriminator consists of two MLP layers with a hidden size of 256, where the first layer is accompanied with a LeakyReLU activation function. The final classifier layer is used to predict the protected attribute. Sub-discriminators are optimized for at most 100 epochs after each epoch of $\texttt{Enc}(\cdot)$ training, leading to extra training time.

C Bios Distribution

Table 4 shows the number of instances of each profession, the number of male and female individuals of each profession, and the ratio of female individuals for each profession in the **Bios** training dataset.

Profession	Total	Male	Female	Ratio
professor	76748	42130	34618	0.451
physician	26648	13492	13156	0.494
attorney	21169	13064	8105	0.383
photographer	15773	10141	5632	0.357
journalist	12960	6545	6415	0.495
nurse	12316	1127	11189	0.908
psychologist	11945	4530	7415	0.621
teacher	10531	4188	6343	0.602
dentist	9479	6133	3346	0.353
surgeon	8829	7521	1308	0.148
architect	6568	5014	1554	0.237
painter	5025	2727	2298	0.457
model	4867	840	4027	0.827
poet	4558	2323	2235	0.490
filmmaker	4545	3048	1497	0.329
software_engineer	4492	3783	709	0.158
accountant	3660	2317	1343	0.367
composer	3637	3042	595	0.164
dietitian	2567	183	2384	0.929
comedian	1824	1439	385	0.211
chiropractor	1725	1271	454	0.263
pastor	1638	1245	393	0.240
paralegal	1146	173	973	0.849
yoga_teacher	1076	166	910	0.846
dj	964	828	136	0.141
interior_designer	949	182	767	0.808
personal_trainer	928	505	423	0.456
rapper	911	823	88	0.097

Table 4: Statistics of the **Bios** training dataset.

C.1 Synthetic Dataset Construction

We follow Subramanian et al. (2021) in constructing the binary classification version of the **Bios** dataset based on the two professions of *nurse* and *surgeon*. For the additional classes in the synthetic experiments, we further select pairs of professions that are both large in size and biased in gender skew, resulting in *photographer* + *teacher*, *dentist* + *psychologist*, and *software engineer* + *model*. The resulting training dataset sizes are 21145, 47449, 68873, and 78232 for 2, 4, 6, and 8 classes, respectively.

D Hyperparameter Settings

We vary the architecture of $\texttt{Embed}(\cdot)$ across different tasks, and do not finetune it during training. The architecture of $\texttt{Enc}(\cdot)$ consists of two fully-connected layers with a hidden size of 300. All models are trained and evaluated on the same dataset splits, and models are selected based on their performance on the development set. For fair comparison, we first finetune the learning rate and batch size using grid search, then finetune hyperparameters introduced by the corresponding debiasing methods for each model on each dataset. For all experiments, we use the Adam optimizer (Kingma and Ba, 2015) and early stopping with a patience of 10.

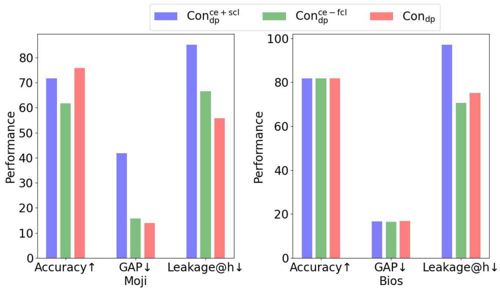

Figure 3: Effects of contrastive loss components for Con$_{\mathrm{dp}}$.

D.1 Twitter Sentiment Analysis

For CE, the learning rate is 0.001, and the batch size is 1024. For INLP, following Ravfogel et al. (2020), we use 300 linear SVM classifiers, each of which is trained over a subset of instances with the same target class. For Adv, the number of sub-discriminators is 3, λ_{adv} is 1.0, and λ_{diff} is 0.01. For Gate, all hyperparameters are the same as CE, except the hidden layers of MLP are replaced by a hyperparameter-free augmentation layer. For FairBatch, the objective is equal opportunity, and the adjustment rate for resampling probabilities is 0.19952623149688797. For EO$_{\mathrm{GLB}}$, the strength of the additional difference loss is 0.3981071705534973. For Con$_{\mathrm{dp}}$, $\tau = 0.01$, and $\alpha = \beta = 0.0199526231496888$. For Con$_{\mathrm{eo}}$, all hyperparameters are the same as Con$_{\mathrm{dp}}$, except for $\alpha = \beta = 0.7943282347242822$.

D.2 Occupation Classification

For CE, the learning rate is 0.003, and the batch size is 2048. For INLP, each classifier is trained over a subset of instances with same target class. For Adv, the number of sub-discriminators is 3, λ_{adv} is 1.0, and λ_{diff} is 0.01. For Gate, all hyperparameters are the same as CE, except for the hidden layers of MLP are replaced hyperparameter-free augmentation layer. For FairBatch, the objective is equal opportunity, and the adjustment rate for resampling probabilities is 0.05011872336272725. For EO$_{\mathrm{GLB}}$, the strength of the additional difference loss is 0.00707945784384138. For Con$_{\mathrm{dp}}$, $\tau = 0.01$, and $\alpha = \beta = 0.00011885022274370189$. For Con$_{\mathrm{eo}}$, all hyperparameters are the same as Con$_{\mathrm{dp}}$, except for $\alpha = \beta = 0.00016788040181225607$.

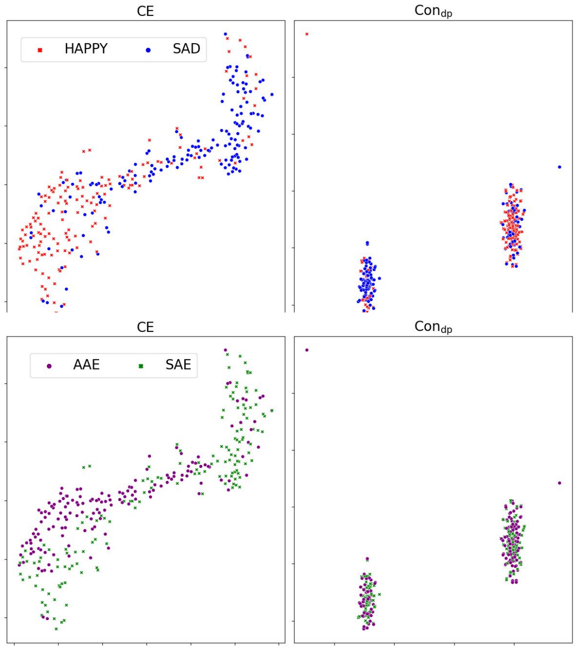

Figure 4: t-SNE scatter plots of learned representations of CE and Con$_{\mathrm{dp}}$ over the **Moji** dataset (based on 150 random samples from each main task class; best viewed in colour). Red and blue colours indicate that they have different sentiment (main task) labels: red → HAPPY and blue → SAD. Green and purple colours indicate that they have different ethnic groups (protected attribute): purple → AAE and green → SAE.

D.3 Analysis

D.3.1 Effect of Loss Components

See Figure 3 for a breakdown of results for each loss component of Con$_{\mathrm{dp}}$ over **Moji** and **Bios**.

D.3.2 Visualising Representations

See Figure 4 for t-SNE plots of learned representations for CE vs. Con$_{\mathrm{dp}}$ over **Moji**.