Optimising Equal Opportunity Fairness in Model Training

Aili Shen[♠] **Xudong Han**[♠] **Trevor Cohn**[♠] **Timothy Baldwin**[♠♡] **Lea Frermann**[♠]

♠ School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
♡ Department of Natural Language Processing, MBZUAI

{aili.shen,t.cohn,tbaldwin,lfrermann}@unimelb.edu.au, xudongh1@student.unimelb.edu.au

Abstract

Real-world datasets often encode stereotypes and societal biases. Such biases can be implicitly captured by trained models, leading to biased predictions and exacerbating existing societal preconceptions. Existing debiasing methods, such as adversarial training and removing protected information from representations, have been shown to reduce bias. However, a disconnect between fairness criteria and training objectives makes it difficult to reason theoretically about the effectiveness of different techniques. In this work, we propose two novel training objectives which directly optimise for the widely-used criterion of *equal opportunity*, and show that they are effective in reducing bias while maintaining high performance over two classification tasks.

1 Introduction and Background

Modern neural machine learning has achieved great success across a range of classification tasks. However, when applied over real-world data, especially in high-stakes settings such as hiring processes and loan approvals, care must be taken to assess the fairness of models. This is because real-world datasets generally encode societal preconceptions and stereotypes, thereby leading to models trained on such datasets to amplify existing bias and make biased predictions (i.e., models perform unequally towards different subgroups of individuals). This kind of unfairness has been reported over various NLP tasks, such as part-of-speech tagging (Hovy and Søgaard, 2015; Li et al., 2018; Han et al., 2021b), sentiment analysis (Blodgett et al., 2016; Shen et al., 2021), and image activity recognition (Wang et al., 2019; Zhao et al., 2017).

Various methods have been proposed to mitigate bias, including adversarial training, and pre- and post-processing strategies. Adversarial training aims to make it difficult for a discriminator to predict protected attribute values from learned representations (Han et al., 2021c; Elazar and Goldberg, 2018; Madras et al., 2018). Pre- and post-processing strategies vary greatly in approach, including transforming the original dataset to reduce protected attribute discrimination while retaining dataset utility (du Pin Calmon et al., 2017), iteratively removing protected attribute information from (fixed) learned representations (Ravfogel et al., 2020), or reducing bias amplification by injecting corpus-level constraints during inference (Zhao et al., 2017).

However, training strategies and optimisation objectives are generally disconnected from fairness metrics which directly measure the extent to which different groups are treated (in)equitably. This makes it difficult to understand the effectiveness of previous debiasing methods from a theoretical perspective. In this work, we propose to explicitly incorporate equal opportunity into our training objective, thereby achieving bias reduction. This paper makes the following contributions:

1. We are the first to propose a weighted training objective that directly implements fairness metrics.
2. Observing that model performance for different classes can vary greatly, we further propose a variant of our method, taking both bias reduction among protected attribute groups and bias reduction among different classes into consideration.
3. Experimental results over two tasks show that both proposed methods are effective at achieving fairer predictions, while maintaining performance.

Our code is available at: https://github.com/AiliAili/Difference_Mean_Fair_Models.

[*] Equal contributors to this work.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4073 - 4084
July 10-15, 2022 ©2022 Association for Computational Linguistics

2 Related Work

2.1 Fairness Criteria

Various criteria have been proposed to capture different types of discrimination, such as group fairness (Hardt et al., 2016; Zafar et al., 2017a; Cho et al., 2020; Zhao et al., 2020), individual fairness (Sharifi-Malvajerdi et al., 2019; Yurochkin et al., 2020; Dwork et al., 2012), and causality-based fairness (Wu et al., 2019; Zhang and Bareinboim, 2018a,b). In this work, we focus on group fairness, whereby a model should perform equally across different demographic subgroups.

To quantify how predictions vary across different demographic subgroups, demographic parity (Feldman et al., 2015; Zafar et al., 2017b; Cho et al., 2020), equal opportunity (Hardt et al., 2016; Madras et al., 2018), and equalised odds (Cho et al., 2020; Hardt et al., 2016; Madras et al., 2018) are widely used to measure fairness. *Demographic parity* ensures that models achieve the same positive rate for each demographic subgroup, oblivious the ground-truth target label. *Equal opportunity* requires that a model achieves the same true positive rate (TPR) across different subgroups, considering only instances with a positive label. *Equalised odds* goes one step further in requiring not only the same TPR but also the same false positive rate (FPR) across groups.

Demographic parity, equal opportunity, and equalised odds only focus on the prediction outcome for one specific target label (i.e. a "positive" class) in a binary classification setting, but does not apply fairness directly to multi-class settings, when fairness for different subgroups across all classes is required. Equal opportunity can be generalised by extending the "positive" class to each target class, as we do in our work.

2.2 Debiasing Methods

A broad range of methods has been proposed to learn fair models. Based on where debiasing occurs, in terms of dataset processing, model training, and inference, we follow Cho et al. (2020) in categorising methods into: (1) pre-processing, (2) post-processing, and (3) in-processing.

Pre-processing methods manipulate the original dataset to mitigate discrimination (Wang et al., 2019; Xu et al., 2018; Feldman et al., 2015; du Pin Calmon et al., 2017; De-Arteaga et al., 2019). For example, du Pin Calmon et al. (2017) transform the original dataset to reduce discrimination while retaining dataset utility. Class imbalance methods used in bias reduction, such as dataset sampling (Kubat and Matwin, 1997; Wallace et al., 2011), instance reweighting (Cui et al., 2019; Li et al., 2020; Lin et al., 2017), and weighted max-margin (Cao et al., 2019), also belong to this category. For example, Lahoti et al. (2020), Subramanian et al. (2021b), and Han et al. (2021a) reweight instances by taking the (inverse of) joint distribution of the protected attribute classes and main task classes into consideration. Wang et al. (2019) and Han et al. (2021a) down-sample the majority protected attribute group within each target class, and train on the resulting balanced dataset.

Post-processing methods calibrate the prediction outcome or learned representations of models to achieve fair predictions (Hardt et al., 2016; Pleiss et al., 2017; Zhao et al., 2017; Ravfogel et al., 2020). For example, Zhao et al. (2017) enforce a corpus-level constraint during inference to reduce bias. Ravfogel et al. (2020) iteratively remove protected attribute information from representations generated by an fixed encoder, by iteratively training a discriminator over the projected attribute and projecting the representation into the discriminator's null space.

In-processing methods learn fair models during model training. One family of approaches is based on constrained optimisation, incorporating fairness measures as regularisation terms or constraints (Zafar et al., 2017b; Subramanian et al., 2021a; Donini et al., 2018; Narasimhan, 2018; Cho et al., 2020). For example, Zafar et al. (2017a) translate equalised odds into constraints on FPR and FNR across groups, and solve using constraint programming. Cho et al. (2020) adopt kernel density estimation to quantify demographic parity and equalised odds, but in a manner which is limited to low-dimensional data and binary classification tasks. Another line of work is to use adversarial training to obtain fair models, in jointly training an encoder and discriminator(s) over the encoded representations such that the discriminator(s) are ineffective at predicting the protected attribute values from learned representations (Han et al., 2021c; Elazar and Goldberg, 2018; Madras et al., 2018; Zhang et al., 2018; Agarwal et al., 2018; Roh et al., 2020). Elsewhere, Shen et al. (2021) use contrastive learning to learn fair models by simultaneously pushing instances belonging to the same target class closer and pulling instances belonging to the same protected attribute class further apart.

The most relevant work to ours is *FairBatch* (Roh et al., 2021). It proposes to formulate the original task as a bi-level optimisation problem, where the inner optimiser is the standard training algorithm and the outer optimiser is responsible for adaptively adjusting the sampling probabilities of instances with a given target class and protected attribute value, based on the equal opportunity metric achieved by the intermediate inner model. That is, they adaptively adjust the instance *resampling* probability during training to reduce bias. However, different from FairBatch, whose resampling strategy is bound by the sampling probability $[0, 1]$, our proposed method achieves bias reduction by *reweighting* instances during training, where the reweighting range is unbounded, leading to greater flexibility in trading off performance and fairness.

3 Methodology

3.1 Preliminaries

Suppose we have some data $X \in \mathbb{R}^n$, target labels $Y \in C$, and protected attribute values $A = \{0, 1\}$, where C is the number of target classes for a given task.

Equal opportunity A classifier is said to satisfy equal opportunity if its prediction is conditionally independent of the protected attribute A given the target label Y, $\{P(\hat{y} = y|Y = y, A = 0) = P(\hat{y} = y|Y = y, A = 1)\}$ for $\forall y \in Y$. Here, $\hat{y}$ is a prediction outcome, $y \in Y$ and $a \in A$. As mentioned above, we slightly modify the definition of equal opportunity by allowing y to be each candidate target class, accommodating multi-class settings. We explicitly address the fairness criterion across *all* target classes by promoting comparable true positive rates across protected classes.

3.2 Optimising Equal Opportunity

Instead of using a fairness proxy (Zafar et al., 2017b) or kernel density estimation to quantify fairness (Cho et al., 2020), we propose to optimise equal opportunity by directly minimising the absolute difference in loss between different subsets of instances belonging to the same target label but with different protected attribute classes,

$$\mathcal{L}_{\text{eo}}^{\text{class}} = \mathcal{L}_{ce} + \lambda \sum_{y=\in C} \sum_{a \in A} |\mathcal{L}_{ce}^{y,a} - \mathcal{L}_{ce}^{y}| \quad (1)$$

Here, $\mathcal{L}_{ce}$ denotes the average cross-entropy loss based on instances in the batch; $\mathcal{L}_{ce}^{y,a}$ denotes the average cross-entropy loss computed over instances with the target label y and the protected attribute label a; and $\mathcal{L}_{ce}^{y}$ denotes the average cross-entropy loss computed over all instances with target label y. Our proposed loss $\mathcal{L}_{\text{eo}}^{\text{class}}$ is the weighted sum of the overall cross-entropy and the sum of the cross-entropy difference for each target label overall and that conditioned on the target label, thereby capturing both performance and fairness. This method is denoted as EO_{CLA}, as it captures $\underline{\text{cla}}$ss-wise equal opportunity.

3.3 Equal Opportunity across Classes

One drawback of EO_{CLA} is that it only focuses on optimising equal opportunity, ignoring the fact that the performance for different classes can vary greatly, especially when the dataset is skewed. To learn fair models not only towards demographic subgroups but also across target classes, we propose a variant of Equation 1, by introducing one additional constraint on top of equal opportunity to encourage the label-wise cross entropy loss terms to align. Formally: $\mathcal{L}_{ce}^{y_1} \approx \mathcal{L}_{ce}^{y_2}$, where $y_1 \neq y_2$, and $y_1 \in Y$, $y_2 \in Y$. This objective encourages equal opportunity not only for demographic subgroups but also across different target classes:

$$\mathcal{L}_{\text{eo}}^{\text{global}} = \mathcal{L}_{ce} + \lambda \sum_{y=\in C} \sum_{a \in A} |\mathcal{L}_{ce}^{y,a} - \mathcal{L}_{ce}| \quad (2)$$

This method is denoted as EO_{GLB}, short for $\underline{\text{gl}}$o$\underline{\text{b}}$al equal opportunity.

3.4 Theory

In this section, we show how our training objective is related to equal opportunity in the binary classification and binary protected attribute setting. Note that our proof naturally extends to cases where the numbers of target classes and/or protected attribute values are greater than two as described in Equations 1 and 2.

Let $m_{y,a}$ be the number of training instances with target label y and protected attribute a in a batch. For example, $m_{1,0}$ denotes the number of instances with target label 1 and protected attribute 0 in the batch. Let $\mathcal{L}^{y,a}$ be the average loss for instances with target label y and protected attribute a. For example, $\mathcal{L}^{1,0}$ is the average loss for instances with target label 1 and protected attribute 0.

3.4.1 Cross-Entropy Loss

The vanilla cross-entropy loss is computed as:

$$\frac{1}{N}(m_{0,0}\mathcal{L}^{0,0} + m_{0,1}\mathcal{L}^{0,1} + m_{1,0}\mathcal{L}^{1,0} + m_{1,1}\mathcal{L}^{1,1}) \quad (3)$$

which is the average loss over different subsets of instances with a given target label and protected attribute class.

3.4.2 Difference Loss

The EO_{CLA} method defined in Equation 1 can be written as:

$$\begin{aligned}\mathcal{L}^{\text{class}}_{\text{eo}} &= \mathcal{L}_{ce} + \lambda \sum_{y,a} |\mathcal{L}^{y,a}_{ce} - \mathcal{L}^{y}_{ce}| \\ &= \sum_{y,a} \Big[\frac{m_{y,a}}{N}\mathcal{L}^{y,a}_{ce} + \lambda\, \text{sign}_{y,a}\, (\mathcal{L}^{y,a}_{ce} - \mathcal{L}^{y}_{ce})\Big], \\ &= \sum_{y,a} \Big[(\frac{m_{y,a}}{N} + \lambda\, \text{sign}_{y,a})\mathcal{L}^{y,a}_{ce} \\ &\qquad - \lambda\, \text{sign}_{y,a}\, \mathcal{L}^{y}_{ce}\Big],\end{aligned} \tag{4}$$

where sign is a sign function, and $\text{sign}_{y,a} = \text{sign}\,(\mathcal{L}^{y,a}_{ce} - \mathcal{L}^{y}_{ce})$. Noting that for binary protected attributes, $\text{sign}_{y,a} = -\,\text{sign}_{y,\neg a}$, and $\sum_{y,a} \mathcal{L}^{y}_{ce} = 0, \forall y$ in this case:

$$\mathcal{L}^{\text{class}}_{\text{eo}} = \sum_{y,a} (\frac{m_{y,a}}{N} + \lambda\, \text{sign}_{y,a})\mathcal{L}^{y,a}_{ce} \tag{5}$$

By comparing Equations 3 and 5, we can see that for target label y, our method dynamically increases the weight for poorly-performing subsets (i.e. $\text{sign}_{y,g} = 1$) by λ, and decreases the weight for well-performing subsets ($\text{sign}_{y,g} = -1$) by λ, thereby leading to fairer predictions by adjusting the weight for instances with different protected attribute classes conditioned on a given target label.

3.4.3 From Binary Cross-Entropy to True Positive Rate

Using the definition of binary cross-entropy

$$-[y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))],$$

the loss for a certain subset, e.g., the subset of instances with target label 1 and protected attribute class 0, can be simplified as:

$$\begin{aligned}\mathcal{L}^{1,0} =& -\frac{1}{m_{1,0}} \sum_{j=1}^{m_{1,0}} \Big(y_j \cdot \log(p(\hat{y}_j)) \\ &+ (1 - y_j) \cdot \log(1 - p(\hat{y}_j))\Big) \\ =& -\frac{1}{m_{1,0}} \sum_{j=1}^{m_{1,0}} \log(p(\hat{y}_j))\end{aligned} \tag{6}$$

Notice that $p(\hat{y}_j)$ is equivalent to $p(\hat{y}_j = 1)$, making $\mathcal{L}^{1,0} = -\frac{1}{m_{1,0}} \sum_{j=1}^{m_{1,0}} \log(p(\hat{y}_j))$ an unbiased estimator of $-\log p(\hat{y} = 1|y = 1)$, which approximates $-\log$ TPR.

Minimising the expectation of the absolute difference between $\mathcal{L}^{1,0}$ and $\mathcal{L}^{1,1}$ can approximate the true positive rate difference between two groups with the same target label 1:

$$\begin{aligned}&\text{argmin}_\theta\, \mathbb{E}(|\mathcal{L}^{1,0} - \mathcal{L}^{1,1}|) \\ &= \text{argmin}\, |-\log p(\hat{y} = 1|y = 1, g = 0) \\ &\qquad - (-\log p(\hat{y} = 1|y = 1, g = 1))| \\ &\approx \text{argmin}\, |\log \frac{TPR_{1,0}}{TPR_{1,1}}| \\ &= \text{argmin}\, |TPR_{1,0} - TPR_{1,1}|\end{aligned}$$

This demonstrates that minimising the absolute difference between $\mathcal{L}^{1,0}$ and $\mathcal{L}^{1,1}$ is roughly equivalent to minimising the TPR difference between two groups with the same target label, which is precisely the formulation of equal opportunity, as described in Section 4.3. Therefore, the second term in our proposed method (Equation 1) is optimising directly for equal opportunity.

4 Experiments

Our experiments compare the performance and fairness of our methods against various competitive baselines, and across two classification tasks.

4.1 Baselines

We compare our proposed methods EO_{CLA} and EO_{GLB} against the following seven baselines:

1. CE: train the model with cross-entropy loss and no explicit bias mitigation.
2. INLP: first train the model with cross-entropy loss to obtain dense representations, and iteratively apply null-space projection to the learned representations to remove protected attribute information (Ravfogel et al., 2020). The resulting representations are used to make predictions.
3. Adv: jointly train the model with cross-entropy loss and an ensemble of three adversarial discriminators for the projected attribute, with an orthogonality constraint applied to the discriminators to encourage diversity (Han et al., 2021c).
4. DS: downsample the dataset corresponding to the protected attribute conditioned on a given target label (Han et al., 2021a).
5. RW: reweight instances based on the (inverse) joint distribution of the protected attribute classes and target classes (Han et al., 2021a).

6. Constrained: formulate the task as a constrained optimisation problem, where equal opportunity is incorporated as constraints (Subramanian et al., 2021a).
7. FairBatch: formulate the model training as a bi-level optimisation problem, as described in Section 2.2 (Roh et al., 2021).

4.2 Experiment Setup

For each task, we first obtain document representations from their corresponding pretrained models, which are not finetuned during training. Then document representations are fed into two fully-connected layers with a hidden size of 300d. For all experiments, we use the Adam optimiser (Kingma and Ba, 2015) to optimise the model for at most 60 epochs with early stopping and a patience of 5. All models are trained and evaluated on the same dataset splits, and models are selected based on their performance on the development set. We finetune the learning rate, batch size, and extra hyperparameters introduced by the corresponding debiasing methods for each model on each dataset (see the Appendix for details). Noting the complexity of model selection given the multi-objective accuracy–fairness tradeoff and the absence of a standardised method for selecting models based on both criteria in fairness research, we determine the best-achievable accuracy for a given model, and select the hyperparameter settings that reduce bias while maintaining accuracy as close as possible to the best-achievable value (all based on the dev set). We leave the development of a fair and robust model selection method to future work.

4.3 Evaluation Metrics

To evaluate the performance of models on the main task, we adopt $\mathrm{F}_1^{\text{micro}}$ and $\mathrm{F}_1^{\text{macro}}$ for all our datasets, taking class imbalance into consideration, especially in the multi-class setting.

To evaluate fairness, we follow previous work (De-Arteaga et al., 2019; Ravfogel et al., 2020) and adopt root mean square TPR gap over all classes, which is defined as

$$\mathrm{GAP} = \sqrt{\frac{1}{|C|}\sum_{y\in Y}(\mathrm{GAP}_y^{\mathrm{TPR}})^2},$$

where $\mathrm{GAP}_y^{\mathrm{TPR}} = |\mathrm{TPR}_{y,a} - \mathrm{TPR}_{y,\neg a}|,\ \ y \in Y$, and $\mathrm{TPR}_{a,y} = \mathbb{P}(\hat{y} = y|y,a)$, indicating the proportion of correct predictions among instances with target label y and protected attribute label a. $\mathrm{GAP}_y^{\mathrm{TPR}}$ measures the absolute performance difference between demographic subgroups conditioned on target label y, and a value of 0 indicates that the model makes predictions independent of the protected attribute.

4.4 Twitter Sentiment Analysis

4.4.1 Task and Dataset

For our first dataset, the task is to predict the binary sentiment for a given English tweet, where each tweet is also annotated with a binary protected 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., 2021c; Shen et al., 2021), we adopt the dataset of Blodgett et al. (2016) (**Moji** hereafter), where 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). The number of instances in the training, dev, and test sets are 100K, 8K, and 8K, respectively. The dev and test set are balanced in terms of sentiment–ethnicity combinations.

4.4.2 Implementation Details

Following previous work (Elazar and Goldberg, 2018; Ravfogel et al., 2020; Han et al., 2021c), we use DeepMoji (Felbo et al., 2017), a model pretrained over 1.2 billion English tweets, as the encoder to obtain text representations. The parameters of DeepMoji are fixed in training. Hyperparameter settings are provided in Appendix A.2.

4.4.3 Experimental Results

Table 1 presents the results over the **Moji** test set. Compared to CE, INLP and Adv moderately reduce model bias while simultaneously improving model performance. Surprisingly, both DS and RW reduce GAP substantially and achieve the joint best $\mathrm{F}_1^{\text{micro}}$, indicating that the biased prediction is mainly due to the imbalanced distribution of protected attribute classes conditioned on a given target label, and the imbalanced distribution of sentiment–ethnicity combinations.[1] However,

[1] However, it does not hold the other way around as demonstrated by previous studies (Wang et al., 2019), indicating that a balanced dataset either in terms of target label and protected attribute combination, or in terms of protected attribute class distribution conditioned on target classes, can still lead to biased predictions.

Model	F_1^{micro} ↑	GAP ↓
CE	72.09±0.65	40.21±1.23
INLP	72.81±0.01	36.81±3.49
Adv	74.47±0.68	30.59±2.94
DS	76.16±0.28	14.96±1.08
RW	**76.21**±0.16†	14.70±0.86
Constrained	75.22±0.20	15.92±4.86
FairBatch	75.81±0.17	15.36±3.07
EO_{CLA}	75.03±0.25	**10.83**±1.40†
EO_{GLB}	75.20±0.20	11.49±1.07

Table 1: Experimental results on the **Moji** test set (averaged over 10 runs); **Bold** = Best Performance; ↑= the higher the better; ↓= the lower the better. The best result is marked with "†" if the difference over the next-best method is statistically significant (based on a one-tailed Wilcoxon signed-rank test; $p < 0.05$), noting that if the best method is one of our methods, we compare it to the next-best method which is not our own.

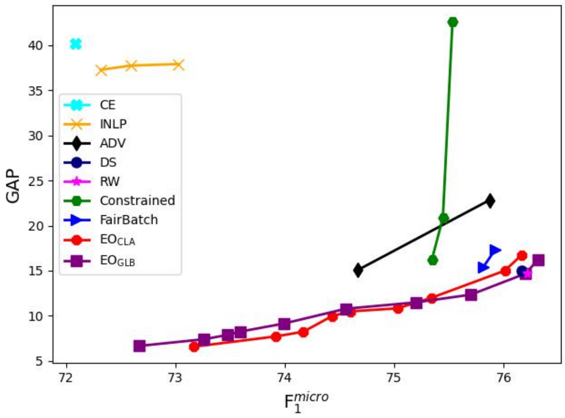

Figure 1: F_1^{micro} vs. GAP of different models on the **Moji** test set, as we vary the most sensitive hyperparameter for each model.

the drawback of dataset imbalance methods is that they lack the flexibility to control the performance–fairness tradeoff. Both Constrained and FairBatch also effectively reduce bias and achieve improved performance. Both of our methods, EO_{CLA} and EO_{GLB}, achieve competitive performance on the main task with the largest bias reduction. For all models except INLP, we can see that incorporating debiasing techniques leads to improved performance on the main task. We hypothesise that incorporating debiasing techniques (either in the form of adversarial training, data imbalance methods, or optimising towards equal opportunity) acts as a form of regularisation, thereby reducing the learned correlation between the protected attribute and main task label, and encouraging models to learn task-specific representations.

Performance–Fairness tradeoff. We plot the tradeoff between F_1^{micro} and GAP for all models on the **Moji** test set in Figure 1. In this, we vary the most-sensitive hyperparameter for each model: the number of iterations for INLP, the λ weight for adversarial loss for Adv, the step size of adjusting resampling probability for FairBatch, and the weight for minimising the loss difference for EO_{CLA} and EO_{GLB}.[2] As we can see, INLP has limited capacity to reduce bias, and the performance for the main task is slightly worse than the other methods. Compared with Adv, Constrained, and FairBatch, our proposed methods EO_{CLA} and EO_{GLB} achieve fairer predictions while maintaining competitive performance (bottom right). Another advantage of our methods is that they allow for greater variability in the performance–fairness tradeoff, demonstrating the effectiveness and superiority of our proposed method. Note that only the pareto points for each model are plotted. For example, for Adv, we experimented with 7 values of λ, but the results are captured by only two pareto points.

4.5 Profession Classification

4.5.1 Task and Dataset

For our second dataset, the task is to predict a person's occupation given their biography (De-Arteaga et al., 2019), where each short online biography is labelled with one of 28 occupations (main task label) and binary gender (protected attribute). Following previous work (De-Arteaga et al., 2019; Ravfogel et al., 2020), the number of instances in the training, dev, and test sets are 257K, 40K, and 99K, respectively.[3]

4.5.2 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) to obtain text representations, and keep BERT fixed during training. Hyperparameter settings for all models are provided in Appendix A.3.

[2] For CE, DS, and RW, there is no hyperparameter that controls the tradeoff between model performance and bias reduction.

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

Model	$\text{F}_1^{\text{macro}}$ ↑	$\text{F}_1^{\text{micro}}$ ↑	GAP ↓
CE	**75.95**±0.10†	**82.19**±0.04 †	16.68±0.46
INLP	71.44±0.40	79.54±0.18	13.52±1.54
Adv	70.88±2.31	79.72±1.02	16.78±0.87
DS	67.73±0.26	78.48±0.10	9.17±0.41
RW	69.21±0.36	76.18±0.32	**8.58**±0.49†
FairBatch	75.14±0.28	81.82±0.07	10.80±1.04
EO_{CLA}	72.07±0.18	81.52±0.06	12.80±0.42
EO_{GLB}	75.11±0.18	81.74±0.07	12.72±0.51

Table 2: Experimental results on the **Bios** test set (averaged over 10 runs). The best result is marked with "†" if the difference over the next-best method is statistically significant (based on a one-tailed Wilcoxon signed-rank test; $p < 0.05$), noting that if the best method is one of our methods, we compare it to the next-best method which is not our own.

4.5.3 Experimental Results

Table 2 shows the results on the **Bios** test set.[4] We can see that Adv is unable to reduce GAP even at the cost of performance in terms of $\text{F}_1^{\text{micro}}$ and $\text{F}_1^{\text{macro}}$. Both DS and RW reduce bias in terms of GAP, at the cost of a drop in performance, in terms of $\text{F}_1^{\text{micro}}$ and $\text{F}_1^{\text{macro}}$. We attribute this to the dramatic decrease in the number of training instances for DS, and the myopia of RW in only taking the ratio of occupation–gender combinations into consideration but not the difficulty of each target class. Among INLP, FairBatch, EO_{CLA}, and EO_{GLB}, we can see that FairBatch achieves a reasonable bias reduction with the least performance drop. This is due to it dynamically adjusting the resampling probability during training. Comparing EO_{CLA} and EO_{GLB}, we can see that EO_{GLB} is better able to deal with the dataset class imbalance (reflected in $\text{F}_1^{\text{macro}}$), while reducing bias.

Performance–Fairness tradeoff. Figure 2 shows the $\text{F}_1^{\text{micro}}$–GAP tradeoff plot for the **Bios** test set. We can see that INLP and Adv reduce bias at the cost of performance, as do DS and RW. Compared with FairBatch, EO_{CLA} and EO_{GLB} provide greater control in terms of performance–fairness tradeoff, such as achieving a smaller GAP with a slight decrease of $\text{F}_1^{\text{micro}}$. A similar trend is also observed for the $\text{F}_1^{\text{macro}}$–GAP tradeoff as shown in Figure 3. Although EO_{CLA} is outperformed by FairBatch, EO_{GLB} provides greater control in terms of performance–fairness tradeoff, suggesting an advantage of EO_{GLB} in enforcing fairness across target classes, especially for the imbalanced dataset.

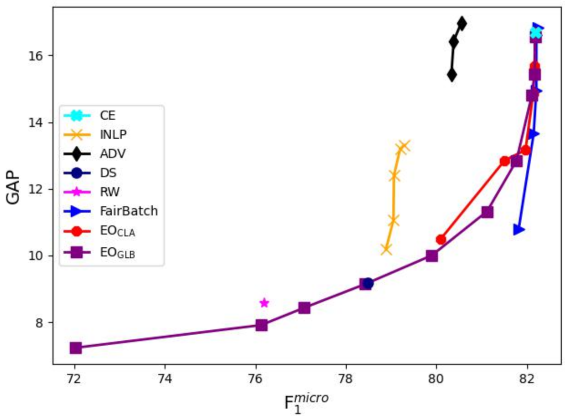

Figure 2: $\text{F}_1^{\text{micro}}$ vs. GAP of different models on the **Bios** test set, as we vary the most sensitive hyperparameter for each model.

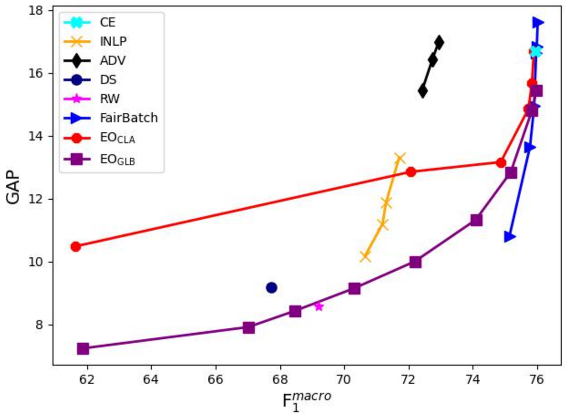

Figure 3: $\text{F}_1^{\text{macro}}$ vs. GAP of different models on the **Bios** test set, as we vary the most sensitive hyperparameter for each model.

5 Analysis

To better understand the effectiveness of our proposed methods, we perform two sets of experiments: (1) an ablation study, and (2) an analysis of training efficiency.

5.1 Ablation Study

EO_{CLA} can be reformulated as $\mathcal{L}_{ce} + \lambda \sum_{y \in C}\{\max(\mathcal{L}_{ce}^{y,a}, \mathcal{L}_{ce}^{y,\neg a}) - \min(\mathcal{L}_{ce}^{y,a}, \mathcal{L}_{ce}^{y,\neg a})\}$, effectively assigning more weight to worse-performing instances (argmax loss) and less weight to better-performing instances (argmin loss). To explore the impact of adjusting weights on model performance,

[4] We omit results for Constrained as it did not converge on this data set, presumably because of its brittleness over multi-class classification tasks.

we experiment with two versions: (1) $\mathcal{L}_{ce} + \lambda \sum_{y \in C} \max(\mathcal{L}_{ce}^{y,a}, \mathcal{L}_{ce}^{y,\neg a})$, denoted as $\text{EO}_{\text{CLA}}^{\text{max}}$, where we assign higher weights to worse-performing instances without changing the weights assigned to better-performing instances; and (2) $\mathcal{L}_{ce} - \lambda \sum_{y \in C} \min(\mathcal{L}_{ce}^{y,a}, \mathcal{L}_{ce}^{y,\neg a})$, denoted as $\text{EO}_{\text{CLA}}^{\text{min}}$, where we assign smaller weights to better-performing instances without changing the weights assigned to worse-performing instances. Correspondingly, for EO_{GLB}, we have $\text{EO}_{\text{GLB}}^{\text{max}}$ and $\text{EO}_{\text{GLB}}^{\text{min}}$. Hyperparameter settings for each model can be found in Appendix B.1.

Tables 3 and 4 show the results for the different models on **Moji** and **Bios**. We can see that the full EO_{CLA} and EO_{GLB} both achieve better bias reduction than ablated *min* and *max* counterparts on **Moji**, while maintaining similar levels of performance in terms of $\text{F}_1^{\text{micro}}$.[5] On **Bios**, we can see that $\text{EO}_{\text{CLA}}^{\text{max}}$ outperforms EO_{CLA} in bias reduction and model performance except for $\text{F}_1^{\text{micro}}$, indicating that it is beneficial for bias reduction to increase the weight for worse-performing instances. On the other hand, $\text{EO}_{\text{CLA}}^{\text{min}}$ is inferior to EO_{CLA} in terms of both bias reduction and performance. We conjecture that reducing the weights for better-performing instances is harmful for model performance (especially for minority classes) over datasets with imbalanced distributions, as is the case for **Bios**.[6] Among the three variants of EO_{GLB}, $\text{EO}_{\text{GLB}}^{\text{max}}$ slightly improves performance on the main task and maintains the same level of bias reduction as EO_{GLB}, while $\text{EO}_{\text{GLB}}^{\text{min}}$ improves performance on the main task but does not reduce bias. Overall, these results show that our two methods perform best in their original formulations.

Model	$\text{F}_1^{\text{micro}}$ ↑	GAP ↓
EO_{CLA}	75.03±0.25	10.83±1.40
$\text{EO}_{\text{CLA}}^{\text{max}}$	75.92±0.10	13.79±1.64
$\text{EO}_{\text{CLA}}^{\text{min}}$	75.33±0.19	14.50±1.78
EO_{GLB}	75.20±0.20	11.49±1.07
$\text{EO}_{\text{GLB}}^{\text{max}}$	76.31±0.10	16.47±0.90
$\text{EO}_{\text{GLB}}^{\text{min}}$	76.27±0.13	18.01±0.40

Table 3: Ablation results over **Moji** test set (averaged over 10 runs).

Model	$\text{F}_1^{\text{macro}}$ ↑	$\text{F}_1^{\text{micro}}$	GAP ↓
EO_{CLA}	72.07±0.18	81.52±0.06	12.80±0.42
$\text{EO}_{\text{CLA}}^{\text{max}}$	72.09±0.19	79.95±0.12	8.98±0.43
$\text{EO}_{\text{CLA}}^{\text{min}}$	53.17±0.53	76.66±0.23	19.22±1.68
EO_{GLB}	75.11±0.18	81.74±0.07	12.72±0.51
$\text{EO}_{\text{GLB}}^{\text{max}}$	75.37±0.06	81.89±0.03	12.47±0.51
$\text{EO}_{\text{GLB}}^{\text{min}}$	75.95±0.12	82.19±0.05	16.74±0.42

Table 4: Ablation results over **Bios** test set (averaged over 10 runs).

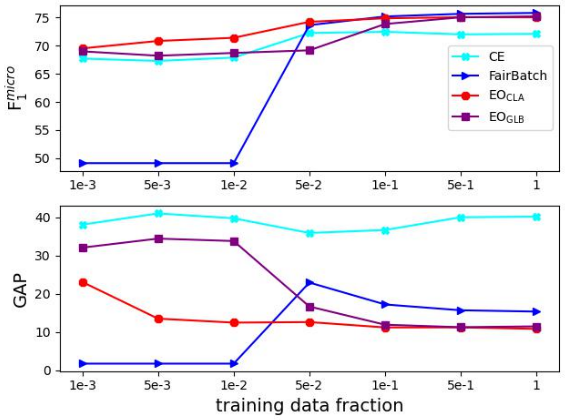

Figure 4: $\text{F}_1^{\text{micro}}$ vs. GAP of different models on the **Moji** test set. The full training set is 100K instances.

5.2 Training Efficiency

To understand the training efficiency of the different models, we perform experiments with varying training data sizes on both **Moji** and **Bios**. Based on results from Tables 1 and 2, we provide results for CE, FairBatch, EO_{CLA}, and EO_{GLB}.

Figure 4 presents the results for **Moji**. When the proportion of training data is no larger than 1K, FairBatch is unable to learn a decent model, while both EO_{CLA} and EO_{GLB} are still effective. As we increase the number of training instances, improved performance on the main task can been observed for all models, and larger bias reduction is achieved for all models except CE. Overall, EO_{CLA} and EO_{GLB} perform well in low-resource settings and achieve better bias reduction for larger volumes of training instances, demonstrating their superiority.

Figure 5 presents the results for **Bios**. We see that FairBatch outperforms EO_{CLA} and EO_{GLB}, especially in terms of $\text{F}_1^{\text{macro}}$ and GAP. Our explanation is that FairBatch adopts a resampling strategy, while our method adopts a reweighting strategy. Although statistically equivalent, resampling outper-

[5] For the max and min versions of both EO_{CLA} and EO_{GLB}, we finetune with the corresponding best-performing λ, respectively. A smaller GAP value cannot be achieved by further adjusting/increasing the value of λ.

[6] This is in line with previous research (Swayamdipta et al., 2020), which shows that easy-to-learn instances are important in optimising models.

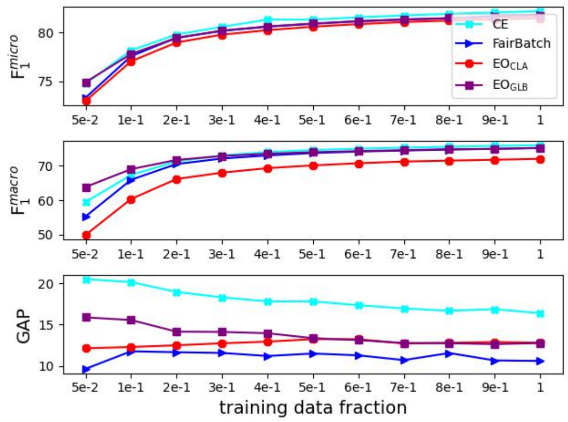

Figure 5: $\mathrm{F}_1^{\mathrm{micro}}$ and $\mathrm{F}_1^{\mathrm{macro}}$ vs. GAP of different models on the **Bios** test set. The full training set is 257K.

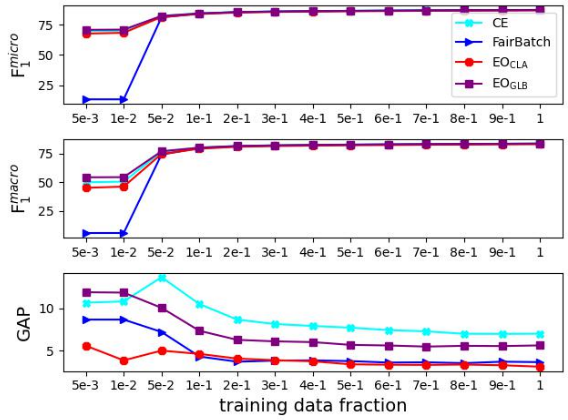

Figure 6: $\mathrm{F}_1^{\mathrm{micro}}$, $\mathrm{F}_1^{\mathrm{macro}}$, vs. GAP of different models on the subset of **Bios** data. Here, instances are from the top-8 most common classes, whose proportion is greater than 4% in the original dataset, resulting into a full training dataset size of 188K.

forms reweighting when combined with stochastic gradient algorithms (An et al., 2021). The data imbalance in **Bios** exacerbates this effect. To verify this, we generated a version of **Bios** with only instances belonging to the top-8 most common classes, whose ratio in the original training set is bigger than 4%. Figure 6 presents results with the subset of dataset consisting of the top-8 most common classes. The plots show a similar trend as observed for the **Moji** dataset on this relatively balanced dataset. Specifically, when the training dataset is small, FairBatch is unable to learn a decent model, while both $\mathsf{EO}_{\mathrm{CLA}}$ and $\mathsf{EO}_{\mathrm{GLB}}$ are still effective.

5.3 Limitations

Consistent with previous work, we did not finetune the underlying pretrained models in obtaining document representations in this work. Finetuning may further remove biases encoded in the pretrained models, which we leave to future work. This work focused only on datasets with binary protected attributes, and future experiments should explore the methods' generalization to higher-arity attributes. For both INLP and Adv, we follow experimental setup from the original papers, noting that the `fairlib` (Han et al., 2022) debiasing framework[7] — which was developed after this work was done — recently showed that both models can obtain better performance and fairness scores with a larger budget for hyperparameter finetuning.

6 Conclusion

We proposed to incorporate fairness criteria into model training, in explicitly optimising for equal opportunity by minimising the loss difference over different subgroups conditioned on the target label. To deal with data imbalance based on the target-label, we proposed a variant of our method which promotes fairness across all target labels. Experimental results over Twitter sentiment analysis and profession classification tasks show the effectiveness and flexibility of our proposed methods.

Ethical Considerations

Our works aims to achieve fairer models, contributing to equal treatment for different demographic subgroups. However, its usage in the real world should be carefully calibrated/auditioned as debiasing for one projected attribute does not guarantee fairness for other protected attributes. In this work, due to the limitations of the dataset, we treat gender as binary, which is not perfectly aligned with the real world.

Acknowledgements

We thank anonymous reviewers for their helpful feedback. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. This work was funded in part by the Australian Research Council.

[7] https://pypi.org/project/fairlib/

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69.

Jing An, Lexing Ying, and Yuhua Zhu. 2021. Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients. In *Proceedings of the 9th International Conference on Learning Representations*.

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.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1565–1576.

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

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277.

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.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2796–2806.

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.

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.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021a. Balancing out bias: Achieving fairness through training reweighting. *CoRR*, abs/2109.08253.

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

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021c. 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, Aili Shen, Yitong Li, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. fairlib: A unified framework for assessing and improving classification fairness. *arXiv preprint*.

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

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.

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

Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: One-Sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186.

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*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

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.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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

Harikrishna Narasimhan. 2018. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654.

Flávio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689.

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.

Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. 2020. Fr-train: A mutual information-based approach to fair and robust training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8147–8157.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. Fairbatch: Batch selection for model fairness. In *Proceedings of the 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. 2021. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*.

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

Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021b. Fairness-aware class imbalanced learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2045–2051.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9275–9293.

Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. 2011. Class imbalance, redux. In *Proceedings of the 11th IEEE International Conference on Data Mining*, pages 754–763.

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.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data*, pages 570–575.

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*, 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. 2018a. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3675–3685.

Junzhe Zhang and Elias Bareinboim. 2018b. 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. 2020. Conditional learning of fair representations. In *Proceedings of the 8th International Conference on Learning Representations*.

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.

A Experimental Settings

A.1 Adv Setup

For Adv, we use 3 sub-discriminators as Han et al. (2021c), where each sub-discriminator consists of two MLP layers with a hidden size of 256, followed by a classifier layer to predict the protected attribute. Sub-discriminators are optimised for at most 100 epochs after each epoch of main model training, leading to extra training time.

A.2 Hyperparameter Settings for Twitter Sentiment Analysis

For all models except for Adv, the learning rate is $3e-3$, and the batch size is 2,048. For INLP, following Ravfogel et al. (2020), we use 300 linear SVM classifiers. For Adv, the learning rate is $1e-3$, and the batch size is 2,048, the number of discriminators is 3, λ_{adv} is 0.5, and λ_{diff} is $1e-3$. For FairBatch, α is set as 0.1. For both EO_{CLA} and EO_{GLB}, λ is set as 0.5. All hyperparameters are finetuned on the **Moji** dev set.

A.3 Hyperparameter Settings for Profession Classification

For all models except for Adv, the learning rate is $3e-3$, and the batch size is 2,048. For INLP, following Ravfogel et al. (2020), we use 300 linear SVM classifiers. For Adv, the learning rate is $1e-2$, and the batch size is 1,024, the number of discriminators is 3, λ_{adv} is $1e-2$, and λ_{diff} is $1e4$. For FairBatch, α is set as $5e-2$. For EO_{CLA}, λ is set as $1e-2$, and for EO_{GLB}, λ is set as $5e-3$. All hyperparameters are finetuned on the **Bios** dev set.

B Analysis

B.1 Ablation Study hyperparameter Settings

For all models, we have tuned the hyperparameter λ and selected model based on performance on the dev set. On the **Moji** dataset, for $\text{EO}^{\text{max}}_{\text{CLA}}$, $\lambda = 2$, for $\text{EO}^{\text{min}}_{\text{CLA}}$, $\lambda = 0.4$, for $\text{EO}^{\text{max}}_{\text{GLB}}$, $\lambda = 2$, for $\text{EO}^{\text{min}}_{\text{GLB}}$, $\lambda = 0.2$. On the **Bios** dataset, for $\text{EO}^{\text{max}}_{\text{CLA}}$, $\lambda = 0.05$, for $\text{EO}^{\text{min}}_{\text{CLA}}$, $\lambda = 0.005$, for $\text{EO}^{\text{max}}_{\text{GLB}}$, $\lambda = 0.005$, for $\text{EO}^{\text{min}}_{\text{GLB}}$, $\lambda = 1e-4$.