Improving Fairness without Demographic by Lagged Dynamic Grouping

Anonymous ACL submission

Abstract

Machine learning models are prone to social biases in datasets and thus could make discriminatory decisions against demographic minority groups. Most existing fairness-promoting methods usually assume access to the annotations of the demographic information. However, such information could be inaccessible due to the high data annotation cost and privacy restrictions. Recently, distributionally robust optimization (DRO) techniques have been applied to promote fairness without demographic labels. DRO-based methods optimize the individuals/groups with the worst prediction performance, with the intuition that these groups roughly correspond to the minority groups being biased against. However, in complex real-world settings with multiple strong bias attributes, the simple grouping schemes in the existing DRO-based methods can fail to identify the ground truth minority groups. In this paper, we propose FREEDRO, a demographic-free group DRO method featuring a more principled grouping scheme, call lagged dynamic grouping. Specifically, FREEDRO dynamically splits the training data based on the ground truth labels and the prediction of the model at an earlier iteration and then optimizes worst group performance. Extensive experiments on five real-world datasets show that our method can effectively alleviate the biases and even achieve comparable results with methods with full demographic annotations. The results also verify that our grouping scheme has a good correspondence with the ground truth demographic grouping.

1 Introduction

The fairness problem in machine learning (ML) has received increasing attention from both academia and industry (Holstein et al., 2019; Mehrabi et al., 2021). Machine learning models are found impacted by the biases in the human world and datasets, and thus make discriminative judgments against certain minority groups. For example, it is found that some abusive language detection systems tend to classify texts that contain mere mentioning of certain minority groups, *e.g.*, homosexual groups, as abusive content, even though the texts themselves are not abusive at all (Dixon et al., 2018). Furthermore, the biases present in these deep learning models can in turn deepen the biases in human society (Zhao et al., 2017).

Most existing fairness promoting methods nowadays require explicit labels of biased demographic features. In many real-world applications, however, such demographic labels are inaccessible, because it is often difficult to identify what demographic biases are present in a machine learning model, and even if the potential biases could be pinpointed, it is sometimes impossible to collect the demographic labels due to privacy constraints (Lahoti et al., 2020). For example, it is impossible for credit card companies to prevent racial discrimination using race labels, because they are not allowed to collect race information from their card applicants.

Motivated by this, many research efforts have been directed to promoting fairness without relying on demographic labels. Among these methods, a major line of works involves distributionally robust optimization (DRO) and group DRO (Hu et al., 2018; Oren et al., 2019; Sagawa et al., 2019, 2020; Liu et al., 2021; Lahoti et al., 2020), whose objective is to optimize the worst performance of individuals or groups. The intuition is that the minority groups being biased against usually corresponds well with the groups with the worst prediction performance. However, in many real-world settings, the data distributions are complex, often involving multiple strong bias features. As a result, the simple grouping schemes as in many DRO-based methods would fail to identify minority groups. A more carefully-designed grouping scheme is needed to maintain the effectiveness of DRO-based methods under these complicated settings.

In this paper, we propose FREEDRO , a demographic-free group-DRO-based debiasing method featuring a more principled grouping scheme, called lagged dynamic grouping, which dynamically partitions the training data based on true labels and the prediction results of the classifier at a previous iteration. The algorithm then optimizes the worst-off group performance with a soft-weighted group DRO objective. We have performed extensive experiments on five real-world datasets, which show that FREEDRO outperforms other demographic-free baselines and can even achieve comparable results with methods that use full demographic information. Further experiments confirm that lagged dynamic grouping has a good alignment with the true demographic grouping.

2 Related Work

2.1 Fairness in Machine Learning

The fairness problem in machine learning has received increasing attention (Chouldechova and Roth, 2018; Sun et al., 2019; Holstein et al., 2019; Mehrabi et al., 2021).

The definitions of fairness can be broadly divided into individual fairness, group fairness, causal-based fairness, and Rawlsian fairness (Chouldechova, 2017; Makhlouf et al., 2020; Lahoti et al., 2020). Among them, individual fairness asks models to achieve similar performance on similar individuals (Dwork et al., 2012; Joseph et al., 2016) while suffering from the difficulty of defining the sample distance measure (Chouldechova and Roth, 2018). Group fairness consists of a family of notions which pursues similar model performance on all demographic groups (Calders and Verwer, 2010; Hardt et al., 2016; Chouldechova, 2017). Despite its effectiveness, different group fairness notions are found incompatible with each other (Kleinberg et al., 2017; Mitchell et al., 2018). Causal-based fairness (Kusner et al., 2017; Wu et al., 2019; Zhang and Bareinboim, 2018a,b) notions define fairness based on causal graphs, which could be hard to acquire without expert knowledge (Makhlouf et al., 2020). In this work, we adopt the Rawlsian fairness (Hashimoto et al., 2018; Lahoti et al., 2020; Lees et al., 2019) defined as the worst-group performance, which originates from the distributive justice theory (Rawls and Kelly, 2001).

The literature on mitigating the fairness problem could be roughly divided into three lines: (1) pre-processing methods (Kamiran and Calders, 2011; Zemel et al., 2013; Feldman et al., 2015; du Pin Calmon et al., 2017; Park et al., 2018; Dixon et al., 2018; Grover et al., 2020; Zhang et al., 2020), which remove the biases by modifying or re-weighting the training data; (2) in-processing algorithms (Zafar et al., 2017; Agarwal et al., 2018; Kamishima et al., 2012; Baharlouei et al., 2019; Pérez-Suay et al., 2017), which add regularization techniques during training; (3) post-processing methods (Hardt et al., 2016; Chzhen et al., 2019; Dwork et al., 2018; Kim et al., 2018, 2019; Woodworth et al., 2017; Mishler and Kennedy, 2021; Lohia et al., 2019), which calibrate the model outputs to satisfy fairness constraints. However, most of these works need the access to the ground-truth demographic information, while the usage of such information may be expensive and privacy risky.

2.2 Fairness without Demographic

A line of works which try to tackle unfairness problem without demographics is the proxy-based methods (Chen et al., 2019; Gupta et al., 2018; Kallus et al., 2020; Diana et al., 2021; Romanov et al., 2019; Zhao et al., 2021), which use proxy variables to replace the demographic information (*e.g.*, zip code as race). However, these methods need prior knowledge of the biases and risks exaggerating the disparities in the datasets due to the estimation bias (Chen et al., 2019; Lahoti et al., 2020). Another branch for solving the problem is to use pseudo group information generated by clustering (Yan et al., 2020) or causal variational autoencoder (Grari et al., 2021), while these methods highly rely on the assumptions of the data distribution. There are also works focusing on achieving fairness with imperfect demographic information (Awasthi et al., 2020; Dai and Wang, 2021; Coston et al., 2019), and with third-party provided privacy-preserved form demographic (Veale and Binns, 2017; Hu et al., 2019; Jagielski et al., 2019).

Another line of works utilizes the distributional robust optimization (DRO) techniques to alleviate the unfairness problems. Without the explicit group information, Hashimoto et al. (2018); Duchi et al. (2020) develop a DRO framework (Namkoong and Duchi, 2017), which considers any subset exceeding a certain size as a potential demographic group. Lahoti et al. (2020) make use of the computational-identifiability to identify the minority groups and remove biases in an adversarial re-weighting man-

ner. Our work falls into the category of group DRO-based methods, which focus on optimizing the worst performances over pre-defined groups and has been shown effective in training unbiased models (Hu et al., 2018; Oren et al., 2019; Sagawa et al., 2019, 2020). Different from prior works, our work defines the groups in a dynamic lagged grouping manner with the predicted and the true labels and optimizes a soft-weighted group DRO objective in these groups.

There are also a series of works for removing unknown datasets biases, which is similar to our focus as the fairness problem can also be partly seen as a dataset bias problem (Zhang et al., 2020). Arjovsky et al. (2019) learn the bias-free representations by optimizing across different environments. Bao et al. (2021a,b) make use of the invariance among separate environments to partial out the impacts of biases. Utama et al. (2020); Sanh et al. (2021) assume the shallow models with limited capacity or under-trained are more prone to biases and force the core model to learn from mistakes of the shallow models. Liu et al. (2021) adopt a two-stage training setting where a shallow model is firstly trained and its misclassified samples are then up-weighted in the second phase training. By contrast, our work introduces a dynamic lagged grouping strategy where groups are iteratively generated based on model predictions and true labels.

3 Background and Problem Formulation

3.1 Problem Formulation

Consider a classification problem, where $\boldsymbol{X}$ represents the input features, and $Y \in \mathcal{Y}$ represents the class label. In addition, denote $Z \in \mathcal{Z}$ as an *unobserved* sensitive attribute, which is often spuriously correlated with Y. Denote $G \in \mathcal{G} = \mathcal{Y} \times \mathcal{Z}$ as the group partitioned by different combinations of Y and Z. Specifically, if Z is a one-dimensional variable, then

$$G = (Y, Z). \tag{1}$$

For example, if Y is a one-dimensional binary variable, say the hair color of a person (blond or black), and Z is a one-dimensional gender feature (we use binary gender for simplicity of illustration), then G can take on four values, (black, male), (black, female), (blond, male) and (blond, female).

On the other hand, if $\boldsymbol{Z}$ is a k-dimensional variable, then $\boldsymbol{G}$ is also a k-dimensional variable, with

$$G_i = (Y, Z_i). \tag{2}$$

In other words, $\boldsymbol{Z}$ provides k different ways to group the data. In the example above, if $\boldsymbol{Z}$ has an additional dimension, (*e.g.*, white or non-white), then $\boldsymbol{G}$ would also have two dimensions. G_1 takes on the four values defined by the output label and gender. G_2 also takes on four values defined by the output label and race. For simplicity of our illustration, the remainder of this section will focus on the case with one-dimensional $\boldsymbol{Z}$, unless stated otherwise.

As the result of the spuriously-correlated attribute $\boldsymbol{Z}$, a regular classifier is prone to produce a high error rate in some of these groups. In the example of predicting hair color based on images of a person, $\boldsymbol{X}$ represents the input image, and $Y \in$ {blond, black} represents the hair color. The gender is a spuriously correlated feature, *i.e.*, $Z \in$ {male, female}, because female people are more likely to have blond hair and male people black hair. As a result, among the four groups, a regular classifier tends to make many errors in the (male, blond) and (female, black) groups. Our goal is to derive a training paradigm that can overcome the bias against these groups.

3.2 Rawlsian Fairness and Group DRO

The fairness metric we would like to optimize for is the Rawlsian fairness, which stems from the distributive justice theory (Rawls and Kelly, 2001). Rawlsian fairness measures the utility of the worst off group. Formally, if we define the parameters of the classifier as θ, the Rawlsian fairness objective is defined as follows,

$$\max_{\boldsymbol{\theta}} \min_{g} \mathbb{E}[U(\boldsymbol{\theta})|G = g], \tag{3}$$

where $U(\boldsymbol{\theta})$ is short for $U(\boldsymbol{X}, Y, Z; \boldsymbol{\theta})$, representing some utility function of an individual. Rawlsian fairness has a strong connection to a wide range of machine learning algorithms (Hashimoto et al., 2018; Lahoti et al., 2020). For example, if we set the utility function of to the negative loss function of the classifier, $\ell(\boldsymbol{\theta})$, such as cross-entropy, then Eq. (3) becomes the objective of the group DRO method:

$$\min_{\boldsymbol{\theta}} \max_{g} \mathbb{E}[\ell(\boldsymbol{\theta})|G = g]. \tag{4}$$

If $\boldsymbol{Z}$ is multidimensional, the objective becomes,

$$\min_{\boldsymbol{\theta}} \max_{i} \max_{g_i} \mathbb{E}[\ell(\boldsymbol{\theta})|G_i = g_i], \tag{5}$$

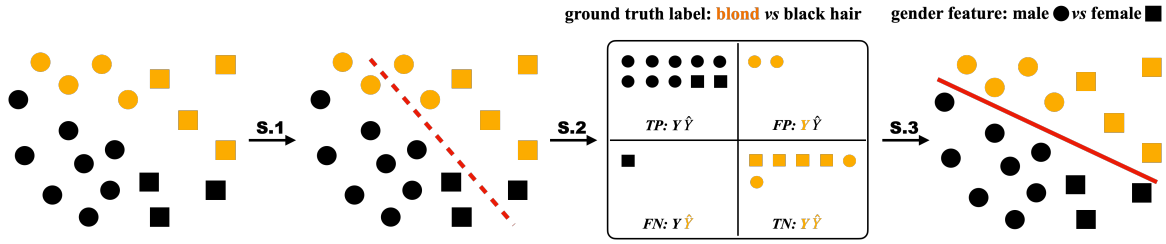

Figure 1: The pipeline of our proposed FREEDRO method with three main steps. **(S.1)** Use the lagged model the make predictions on the current data; **(S.2)** Partition the data based on the model predictions and the ground truth labels; **(S.3)** Optimize the Equation 9 across all partitions for rawlsian fairness promotion. Best viewed in color.

which is essentially minimizing the loss within the worst-off group.

If the sensitive attribute, Z, were observed, we could directly apply the group DRO method, which has been shown effective in mitigating bias (Hu et al., 2018; Oren et al., 2019; Sagawa et al., 2019, 2020). In our case, however, since Z is unobserved, we will explore ways to approximately recover the grouping information to bring group DRO back to the stage in the following.

4 Methodology

The key idea of our proposed FREEDRO is to introduce an alternative grouping, called lagged dynamic grouping, to partition the training data, and then optimize the performance of the worst group. We will first introduce our algorithm FREEDRO in the first subsection, and then explain why our grouping can approximately recover the group demographic grouping in the following subsections.

4.1 Lagged Dynamic Grouping

Denote $\boldsymbol{\theta}^{(t)}$ as the parameter of the classifier at iteration t. Our grouping, $\hat{G}(\boldsymbol{\theta}^{(t)})$, is defined as

$$\hat{G}(\boldsymbol{\theta}^{(t)}) = (Y, \hat{Y}(\boldsymbol{\theta}^{(t)})), \tag{6}$$

where $\hat{Y}(\boldsymbol{\theta}^{(t)})$ denotes the predicted label using the classifier at iteration t. Comparing Equations (1) and (6), we can see that the only difference between our proposed grouping and the oracle demographic grouping is that we replace the ground truth demographic information, Z, with the predicted output label, $\hat{Y}(\boldsymbol{\theta}^{(t)})$. As a concrete example, if Y is binary, then our proposed grouping partitions the data into four groups: true positive, true negative, false positive, and false negative, using the classifier at iteration t.

FREEDRO can then be formulated as follows. At iteration t, the algorithm involves two steps: 1) The algorithm uses the classifier at $\tau = t \bmod T$ iterations *earlier* to generate grouping $\hat{G}(\boldsymbol{\theta}^{(t-\tau)})$, where T is the delay interval length. 2) Then, FREEDRO tries to minimize the worst group performance. Formally, the objective at iteration t is

$$\min_{\boldsymbol{\theta}} \mathbb{E}[\ell(\theta) | \hat{G}(\boldsymbol{\theta}^{(t-\tau)}) = g^*]. \tag{7}$$

where

$$g^* = \operatorname*{argmax}_{g} \mathbb{E}[\ell(\theta) | \hat{G}(\boldsymbol{\theta}^{(t-\tau)}) = g]. \tag{8}$$

$\boldsymbol{\theta}^{(t+1)}$ is obtained by performing one (stochastic) gradient step of the objective above. This objective is very similar to Equation (4), except that the oracle demographic grouping, G, is replaced with our proposed grouping with delay τ, $\hat{G}(\boldsymbol{\theta}^{(t-\tau)})$.

To stabilize training, we further generalize Equation (7) to a soft-weighted version:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \big[& w_s \mathbb{E}[\ell(\theta) | \hat{G}(\boldsymbol{\theta}^{(t-\tau)}) = g^*] \\ & + w_u \mathbb{E}[\ell(\theta) | \hat{G}(\boldsymbol{\theta}^{(t-\tau)}) \neq g^*] \big], \end{aligned} \tag{9}$$

where w_s should be greater than w_u to emphasize the worst off group. When $w_u = 0$ and $w_s = 1$, Objective (9) reduces to Objective (7).

To sum up, our algorithm FREEDRO has three features.

- **Grouping with predicted and true labels:** The data are partitioned into different groups based on the predicted and true labels of the classifier.
- **Lagged grouping:** Classifier trained at an earlier step is used to produce the grouping.
- **Dynamic grouping:** At each iteration, a different grouping is generated.

Hence we name our algorithm lagged dynamic grouping. In the following, we will explain why each feature is essential.

4.2 Why Lagged Dynamic Grouping?

As mentioned, the overall goal of the proposed algorithm is to approximately recover the ground truth demographic grouping, G. In this subsection, we will explain why the three features of FREEDRO can contribute to this goal.

Why grouping based on predicted and true labels? If the classifier primarily uses the biased demographic feature to make predictions, then the grouping based on its predicted label, together with the true labels, can recover the ground truth demographic grouping. To see why this is the case, Figure 1 illustrates, using the hair color prediction example. Recall that the true grouping G in this case is {(blond, female), (black, female), (blond, male), (black, male)}; whereas our grouping is {(blond, predicted blond), (black, predicted blond), (blond, predicted black), (black, predicted black)}. In the figure, the shape ● denotes male, and the shape ■ denotes female), where the hair color label **blond** hair is represented by blond color while **black** hair is represented by black color. If the classifier primarily uses gender to predict hair color, the decision boundary should look like the dashed line, where the male samples (●) mostly lie on one side and female samples (■) mostly lie on the other side. As a result, (blond, predicted blond) mostly corresponds to (blond, female), and (black, predicted black) mostly corresponds to (black, male). Likewise, the remaining two groups also have good correspondence. Having established such correspondence, the next question is, how to find a classifier that primarily relies on biased demographic features, the answer to which lies in the second feature of FREEDRO .

Why use lagged grouping? Consistent with Utama et al. (2020); Sanh et al. (2021); Liu et al. (2021), we assume that the biased demographic features usually take on simple, easy-to-use forms, just like the gender feature in the previous example. Hence, neural models tend to focus on such features at an early training stage, before switching their attention to complicated features.

As discussed in Section 5.3, we conduct a case study and observe a similar behavior that models trained at earlier epochs pick up more bias. This observation motivates us to use a model trained at an earlier stage to produce the grouping so that the classifier would more heavily utilize the demographic info and the resulting grouping can better align with the ground truth demographic grouping.

Why use dynamic grouping? So far, our discussion has only focused on the case with only one biased demographic feature. In the presence of multiple biased demographic features, our algorithm will fail to recover the grouping, if our biased lagged classifier only relies on a subset of, but not all, the biased demographic features. However, this problem can be fixed by dynamic grouping. Assume, for example, there are two biased demographic features, gender, and race, and assume that the classifier initially only focuses on gender. As a result, lagged dynamic grouping would only fix the gender bias, and so race becomes the only unresolved bias. With the dynamic grouping, the classifier would redirect its attention to any unresolved biases, in this example race, and hence fix all the biases at the end.

5 Experiment

In this section, we first introduce our experiment settings, which include the datasets, baselines, model architectures, training details, and evaluation metrics. We then elaborate on the experiment setup of the motivating example we have shown in Figure 2. After that, we present our main results on five real-world datasets. Lastly, we provide ablation studies on the key parameters of the proposed methods.

5.1 Experiment Settings

Datasets We evaluate the proposed method using five real-world datasets with reported biases: `COMPAS` (Flores et al., 2016), `Sexist Tweets` (Waseem, 2016; Waseem and Hovy, 2016; Park et al., 2018), `Jigsaw Toxicity` (AI, 2019), `Civil Comments` (Koh et al., 2021), `Wiki Comments` (Dixon et al., 2018). These datasets are all used for binary classification. For `COMPAS`, we use the data split which is used in Lahoti et al. (2020). We use the original data split for `Civil Comments` (Koh et al., 2021) and `Wiki Comments` (Dixon et al., 2018). For `Sexist Tweets` and `Jigsaw Toxicity`, we randomly split the dataset as 8:1:1 for training, validation and testing. More details about these datasets can be found in Table 1.

Baselines We compare our method with the following baselines:

Dataset	Size	Task	Positive (%)	avg. Length	Bias
COMPAS	7,214	Recidivism Prediction	48.1	Categorical	Race, Gender
Sexist Tweets	12,096	Toxicity Detection	24.7	14.4	Gender
Jigsaw Toxicity	1,999,516	Toxicity Detection	8.0	50.7	Race, Gender, Religion, Sex orientation
Civil Comments	448000	Toxicity Detection	11.3	59.9	Race, Gender, Religion, Sex orientation
Wiki Comments	159,686	Toxicity Detection	9.6	66.0	Gender, Race, Sex orientation, Age, Nationality, Religion

Table 1: The details of the datasets that we use for evaluation. The "Positive (%)" column indicates the ratio of positive labels (*i.e.*, toxic for toxicity detection). The "Bias" column indicates the demographic features that are previously reported to be spuriously correlated with labels in the corresponding datasets.

• **ERM:** conventional empirical risk minimization.
• **ARL** (Lahoti et al., 2020): adversarially reweighting learning assumes no access to demographic information. An adversarial network is co-trained with the final model to generate instance weights to highlight hard samples.
• **JTT** (Liu et al., 2021): just train twice is a two-stage demographic-free training approach, which minimizes the loss over a reweighted dataset (second stage) with those training samples that are misclassified at the end of first-stage standard training up-weighted.
• **POE** and **REWEIGHTING** (Utama et al., 2020): A shallow model is firstly trained with limited steps in the training set. Product-of-expert trains the main model in an ensemble with the shallow model, where the softmax outputs of two models are combined. Reweighting adjusts the importance of a training instance by directly assigning a scalar weight calculated by the shallow model. Higher weights are assigned to those that the shallow model wrongly predicts. An annealing mechanism is applied for both methods.
• **IPW:** inverse probability weighting is an instance re-weighting technique. We specifically consider two variants of the IPW methods denoted as IPW-Z (Höfler et al., 2005) and IPW-ZY (Lahoti et al., 2020). The former one uses $1/P(z)$ as the constant weight while the latter one uses $1/P(z, y)$. These methods need to have explicit demographic information.
• **GDRO:** group distributional robust optimization (Hu et al., 2018; Sagawa et al., 2019) uses the ground-truth demographic information together with the label to partition the dataset. We optimize the worst-off loss over the considered groups.

It is worth emphasizing that all but IPW and GDRO are demographic-free methods. IPW and GDRO have an unfair advantage of accessing the demographic labels.

Evaluation Metrics We evaluate all the methods with average accuracy and worst-group accuracy in the testing set. One challenge for calculating worst-group fairness is that one sentence can contain multiple demographic values (*e.g.*, The boy and the girl are happy.) Following Koh et al. (2021); Liu et al. (2021), we partition the dataset into multiple overlapping groups, where each sentence contains a specific demographic value (*e.g.*, "male") and a specific label. For example, if we have two demographic values $\{$"$male''$, "$female''\}$ and two labels $\{0, 1\}$, then the dataset will be partitioned into four. With these partitioned groups, we report the worst-off accuracy among them.

The ground-truth demographic information is needed to calculate worst-group accuracy during validation and testing. For COMPAS, we use the provided gender and race annotations in the dataset. For Civil Comments, we directly use the provided eight demographics in the original testing set. For Jigsaw Toxicity, we use the provided demographic annotations and manually cluster them into eight demographics which are identical to Civil Comments following Koh et al. (2021). For Wiki Comments and Sexist Tweets, due to the lack of demographic annotations, we follow Zhang et al. (2020) and match each sentence with a pre-defined demographic word dict to extract the demographic information. We filter out those demographic values with the low occurrence and get two demographics for Sexist Tweets and sixteen for Wiki Comments. The specific details of the demographic identities can be found in Appendix A.1.

Model architectures and training details All approaches we evaluate are trained with the same model architecture and training setup. For the COMPAS dataset, we use a fully connected two-layer feed-forward network with $\{64, 32\}$ hidden dimension, and train the models using an Adagrad optimizer with 0.01 learning rates for 100 epochs. For the other four biased datasets, we use the BERT-base-uncased model (Devlin et al., 2019) implemented with Transformers (Wolf et al., 2020). An

Method	Need Demographic	COMPAS	Sexist Tweets	Jigsaw Toxicity	Civil Comments	Wiki Comments
ERM	No	47.4 (66.8)	78.0 (91.9)	52.2 (95.1)	55.1 (91.9)	57.1 (96.6)
ARL	No	46.9 (67.5)	82.0 (92.3)	52.9 (95.2)	57.1 (91.9)	**60.0** (96.3)
JTT	No	52.7 (58.9)	80.0 (92.1)	54.6 (95.1)	70.1 (90.9)	58.3 (96.0)
POE	No	46.5 (60.6)	76.0 (91.2)	52.3 (93.4)	66.0 (90.7)	57.1 (96.3)
REWEIGHTING	No	51.6 (57.1)	82.0 (92.1)	65.2 (93.3)	69.2 (90.1)	57.1 (96.6)
IPW-Z	Yes	46.9 (67.1)	84.0 (92.0)	58.5 (95.2)	56.6 (91.7)	**60.0** (96.1)
IPW-ZY	Yes	47.0 (67.3)	82.0 (92.1)	53.5 (95.1)	56.4 (91.8)	58.3 (96.5)
GDRO	Yes	**63.3** (65.9)	84.9 (92.8)	50.5 (95.2)	70.0 (90.2)	**60.0** (96.6)
FREEDRO (Ours)	No	52.3 (63.1)	**85.3** (92.1)	**69.6** (94.6)	**72.1** (90.6)	**60.0** (95.7)

Table 2: Performance of different methods. All the models are evaluated on the *Original* testing set. The results are reported in the *worst group accuracy* ↑ *(average accuracy* ↑*)* manner. % is neglected.

AdamW optimizer (Loshchilov and Hutter, 2017) with a linearly-decaying learning rate (with initial value 1e-5) and gradient clipping (ℓ_2-norm = 1) are applied. We train for 10 epochs on Sexist Tweets and 5 epochs for other datasets due to the limits on computational resources. The batch size is set as 16 and weight decay is 0.01.

Besides the above hyper-parameters shared across all methods, we tune the additional hyper-parameters of each algorithm based on the highest worst-group accuracy calculated on the validation set. Specifically, we list more details about hyper-parameter tuning for all approaches in the Appendix A.2.

5.2 Experiment Results

Table 2 shows the performance of FREEDRO with the other baselines. Below are our key observations.

First, our method consistently outperforms the other demographic-free baselines across all four text datasets. For example, FREEDRO improves at least 4.4% worst group accuracy on the Jigsaw Toxicity dataset compared with other demographic-free baselines. FREEDRO outperforms these baselines by 3.3% on the Sexist Tweets and 2.0% on Civil Comments.

Second, our method achieves a better trade-off between average accuracy and worst group accuracy compared with demographic-free baselines. For example, FREEDRO improves the worst-group accuracy by at least 3.3% and achieves comparable average accuracy with other demographic-free methods in Sexist Tweets. In Civil Comments, our method improves worst-group accuracy by at least 2.0% compared with JTT, POE and REWEIGHTING with at most 0.3% drop in average accuracy. Compared with Arl, our method brings 15.0% worst-group accuracy improvement in cost of only 1.3% loss on average accuracy. We also note there is still a gap between average and worst group accuracy, which suggests that biases are not completely eliminated. We ascribe this to the imperfect grouping without true demographics.

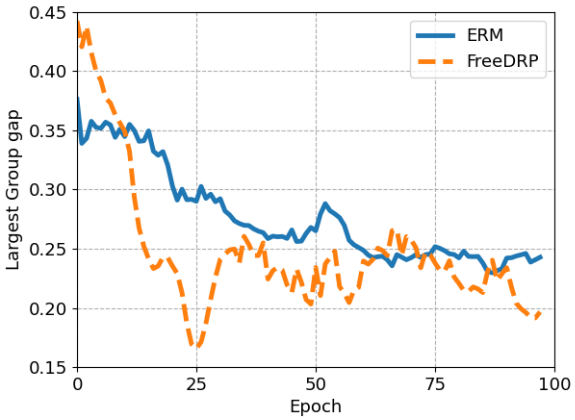

Figure 2: The training trajectory in the COMPAS dataset. The y-axis indicates the gap between the best and the worst group accuracy. Such disparity indicates the biases of models.

Third, our method achieves comparable performances to the methods with ground-truth demographics in the four text datasets. For example, FREEDRO achieves comparable worst group accuracy with GDRO on Sexist Tweets and Wiki Comments, and even outperforms GDRO in Wiki Comments and Jigsaw Toxicity. One possible cause is the training instability of GDRO as it needs to optimize the worst group performance over all demographics and labels, while FREEDRO only considers four groups.

5.3 Case Studies on COMPAS

We conduct a case study on the COMPAS dataset to verify some of our hypotheses that motivate the design of FREEDRO.

First, as discussed in Section 4.2, one of our motivating hypotheses is that a neural model trained at an earlier iteration would pick up more bias. To verify the hypothesis, in Figure 2, we report the gap between the best and the worst group accuracy as a function of iterations. The blue line and the orange line correspond to ERM and FREEDRO re-

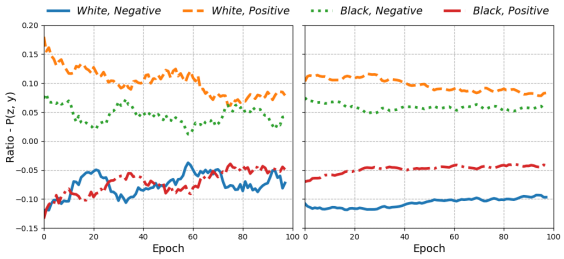

Figure 3: The ratio of the (race, label) groups minus the prior probability in the selected groups for training of GDRO (left) and FREEDRO (right). Greater than zero in y-axis means the group is up-weighted during training.

K	w_u	w_s	Worst Group	Average
0	1.0	6.0	65.4	91.1
1	1.0	6.0	67.3	90.5
2	**1.0**	**6.0**	**69.0**	**90.7**
2	0.1	1.0	60.4	89.5
2	0.2	1.0	69.0	90.7
2	0.5	1.0	62.0	91.8
2	1.0	10.0	66.6	89.7
2	**1.0**	**6.0**	**69.0**	**90.7**
2	1.0	3.0	65.9	91.2

Table 3: The effects of the lagged grouping frequency T and the soft weights w_u and w_s of Equation 9. *Worst Group* and *Average* indicate the correspondingly measured accuracy. The results are evaluated in the validation set. The selected model corresponding to FREEDRO in Table 2 is **bolded**.

spectively. This gap measures how much bias is captured in each model. The more the classifier relies on the biases in the dataset, the greater the gap between the best and the worst group performance. If the classifier does not rely on the bias at all, the gap should be 0. As can be observed, the gap decreases for both algorithms, which confirms our assumption that models trained at an earlier epoch rely more heavily on bias features. In addition, we observe that FREEDRO achieves a smaller gap compared to ERM in the training trajectory, demonstrating that our method can effectively mitigate the biases.

Second, we would also like to investigate how well the grouping of FREEDRO aligns with the true demographic grouping. To measure the group alignment, for every epoch, we collect the samples in the worst-performing group as selected by FREEDRO and calculate the ratio of the four ground-truth demographic groups (*e.g.*, {*(white, negative), (white, positive), (black, negative), (black, positive)*}) in it. If FREEDRO can identify the true demographic groups well, the ratio of the minority groups (in this case *(white, positive)* and *(black, negative)*) should be high.

Figure 3 plots the ratio curves for GDRO (left) and FREEDRO (right) as functions of training epochs. To make it easier to read, we subtract each ratio with the prior probability p(z, y) in the training data. In other words, if the demanded ratio is greater than zero, it means the corresponding group is up-weighted during training of FREEDRO. We show the result of GDRO as a reference because it uses the ground-truth demographic information to partition the training data and select the worst-performing group, and thus its ratio curves should exhibit the most ideal behavior. As can be observed in the figure, the minority groups *(white, positive)* and *(black, negative)* are successfully up-weighted by our method, which is consistent with the results of GDRO. These results demonstrate the effectiveness of our grouping strategy.

5.4 Ablation Studies

We perform a parameter sensitivity study on `Civil Comments` to investigate the effects of the key parameters of FREEDRO including the delay interval T and the soft weights w_u and w_s of Equation 9.

Effect of lagged grouping Table 3 shows the results on the delay interval T, *i.e.*, the number of lagged epochs at which the classifier is used to generate the grouping. When $T = 0$, the grouping is generated with the concurrent model. We see that with the increase of the T, the worst group accuracy improves, which verifies our hypothesis that lagged models can produce better grouping. We note that the average accuracy also drops when T increases, indicating that larger T can reduce the gap between the worst and the average accuracy.

Effect of relaxation coefficient Table 3 shows the results on the soft weights w_u and w_s as in Equation 9. We see that with larger w_s/w_u comes better average accuracy, while the best worst group occurs in a sweet spot at around $w_s/w_u = 6$.

6 Conclusion

In this paper, we introduce a novel DRO-based debiasing method without the use of demographic information, termed as FREEDRO. A lagged dynamic grouping strategy is used to iteratively partition the training data with the model predictions and true labels, and the worst-off performance over the splits is optimized with a soft-weighted group DRO. Extensive experiments are provided to demonstrate the superiority of our method over state-of-the-art demographic-free debiasing methods.

7 Broader Impact

Machine learning models are prone to the unfairness biases in datasets and thus risk making discriminatory decisions towards minority groups (Holstein et al., 2019; Mehrabi et al., 2021). Biased ML systems could even amplify the disparities and deepen the biases in human society (Zhao et al., 2017). Along with the proliferation of the use of ML techniques, it's critical to make sure that ML systems treat every demographic equally.

Despite the recent advances in mitigating the biases, most existing works need access to the ground-truth demographic annotations, which can be difficult due to the high cost and privacy risks (Holstein et al., 2019). To alleviate the problem, we propose a novel training method termed as FREEDRO which can promote model fairness without the use of demographics. The experiments demonstrate that the proposed method can effectively improve model fairness. We believe that our research could help build more fair and responsible ML systems and provide broad positive impacts on both research and industry.

Despite the effectiveness, we note that our method also has a few limitations and potential risks. First, our method, together with all other compared approaches, still needs a demographic-annotated development set to tune the hyper-parameters. Although it is cheaper to acquire such a set, it demands the users notice the biases in advance, which could be hard in practice. It remains an open problem to find more effective validation measures for fairness that do not use the demographics. Second, as demonstrated by the experiment results, the biases may not be completely eliminated. We leave it for future work to further dig into this problem. In practice, our method should be used with careful checks on potential ethical risks.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A reductions approach to fair classification. *ArXiv*, abs/1803.02453.

Jigsaw/Conversation AI. 2019. Jigsaw unintended bias in toxicity classification.

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *ArXiv*, abs/1907.02893.

Pranjal Awasthi, Matthäus Kleindessner, and Jamie H. Morgenstern. 2020. Equalized odds postprocessing under imperfect group information. In *AISTATS*.

Sina Baharlouei, Maher Nouiehed, and Meisam Razaviyayn. 2019. Rényi fair inference. *arXiv: Learning*.

Yujia Bao, Shiyu Chang, and R. Barzilay. 2021a. Predict then interpolate: A simple algorithm to learn stable classifiers. *ArXiv*, abs/2105.12628.

Yujia Bao, Shiyu Chang, and Regina Barzilay. 2021b. Learning stable classifiers by transferring unstable features. *arXiv preprint arXiv:2106.07847*.

Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21:277–292.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, G. Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

A. Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163.

A. Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *ArXiv*, abs/1810.08810.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, L. Oneto, and Massimiliano Pontil. 2019. Leveraging labeled and unlabeled data for consistent fair binary classification. In *NeurIPS*.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

Enyan Dai and Suhang Wang. 2021. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Multiaccurate proxies for downstream fairness. *ArXiv*, abs/2107.04423.

Lucas Dixon, John Li, Jeffrey Scott Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

Flávio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *NIPS*.

John C. Duchi, Tatsunori B. Hashimoto, and Hongseok Namkoong. 2020. Distributionally robust losses for latent covariate mixtures. *ArXiv*, abs/2007.13982.

C. Dwork, Moritz Hardt, Toniann Pitassi, O. Reingold, and R. Zemel. 2012. Fairness through awareness. *ArXiv*, abs/1104.3913.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark D. M. Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *FAT*.

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

Anthony W. Flores, Kristin A. Bechtel, and Christopher T. Lowenkamp. 2016. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks". *Federal Probation*, 80:38.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2021. Fairness without the sensitive attribute via causal variational autoencoder. *ArXiv*, abs/2109.04999.

Aditya Grover, Kristy Choi, Rui Shu, and Stefano Ermon. 2020. Fair generative modeling via weak supervision. In *ICML*.

Maya R. Gupta, Andrew Cotter, M. M. Fard, and Serena Wang. 2018. Proxy fairness. *ArXiv*, abs/1806.11212.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *NIPS*.

T. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *ICML*.

Michael Höfler, Hildegard Pfister, Roselind Lieb, and Hans-Ulrich Wittchen. 2005. The use of weights to account for non-response and drop-out. *Social Psychiatry and Psychiatric Epidemiology*, 40:291–299.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miroslav Dudík, and H. Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Hui Hu, Yijun Liu, Zhen Wang, and Chao Lan. 2019. A distributed fair machine learning framework with private demographic data protection. *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1102–1107.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. 2018. Does distributionally robust supervised learning give robust classifiers? In *ICML*.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. 2019. Differentially private fair learning. In *ICML*.

Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *ArXiv*, abs/1605.07139.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2020. Assessing algorithmic fairness with unobserved protected class using data combination. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *ECML/PKDD*.

Michael P. Kim, Amirata Ghorbani, and James Y. Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

Michael P. Kim, Omer Reingold, and Guy N. Rothblum. 2018. Fairness through computationally-bounded awareness. In *NeurIPS*.

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

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Wei hua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *NIPS*.

Preethi Lahoti, Alex Beutel, Jilin Chen, K. Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. 2020. Fairness without demographics through adversarially reweighted learning. *ArXiv*, abs/2006.13114.

Alyssa Lees, Ananth Balashankar, Chris Welty, and Lakshminarayanan Subramanian. 2019. Pareto-efficient fairness for skewed subgroup data.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *ICML*.

Pranay Kr. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2019. Bias mitigation post-processing for individual and group fairness. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on causal-based machine learning fairness notions. *ArXiv*, abs/2010.09553.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman, and A. G. Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54:1 – 35.

Alan Mishler and Edward H. Kennedy. 2021. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Shira Mitchell, Eric Potash, and Solon Barocas. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv: Applications*.

Hongseok Namkoong and John C. Duchi. 2017. Variance-based regularization with convex objectives. In *NIPS*.

Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. *ArXiv*, abs/1909.02060.

J. Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *EMNLP*.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. 2017. Fair kernel learning. In *ECML/PKDD*.

J. Rawls and Erin I. Kelly. 2001. Justice as fairness: A restatement.

Alexey Romanov, Maria De-Arteaga, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. *ArXiv*, abs/1904.05233.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv*, abs/1911.08731.

Shiori Sagawa, Aditi Raghunathan, P. W. Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. *ArXiv*, abs/2005.04345.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. *ArXiv*, abs/2012.01300.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding-Royer, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *ACL*.

Prasetya Ajie Utama, N. Moosavi, and Iryna Gurevych. 2020. Towards debiasing nlu models from unknown biases. *ArXiv*, abs/2009.12303.

Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *NLP+CSS@EMNLP*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *ArXiv*, abs/1702.06081.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. Pc-fairness: A unified framework for measuring causality-based fairness. In *NeurIPS*.

Shen Yan, Hsien-Te Kao, and Emilio Ferrara. 2020. Fair class balancing: Enhancing model fairness without observing sensitive attributes. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *AISTATS*.

R. Zemel, Ledell Yu Wu, Kevin Swersky, Toniann Pitassi, and C. Dwork. 2013. Learning fair representations. In *ICML*.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and T. Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *ACL*.

Junzhe Zhang and Elias Bareinboim. 2018a. Equality of opportunity in classification: A causal approach. In *NeurIPS*.

Junzhe Zhang and Elias Bareinboim. 2018b. Fairness in decision-making - the causal explanation formula. In *AAAI*.

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 *EMNLP*.

Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2021. You can still achieve fairness without sensitive attributes: Exploring biases in non-sensitive features. *ArXiv*, abs/2104.14537.

A Appendix

A.1 Demographic Identities of Datasets

We list the demographic identities used for evaluation in all five datasets below. We note that we have filtered out those identities with low frequency in `Wiki Comments` and `Jigsaw Toxicity`.

- `COMPAS`: *Black, White, Male, Female*
- `Sexist Tweets`: *Male, Female*
- `Jigsaw Toxicity`: *Male, Female, LGBTQ, Christian, Muslim, Other Religion, Black, White*
- `Civil Comments`: *Male, Female, LGBTQ, Christian, Muslim, Other Religion, Black, White*
- `Wiki Comments`: *Male, Female, LGBTQ, Heterosexual, American, European, Asian, Jewish, Black, White, Other Race, Old, Young, Christian, Catholic, Muslim*

A.2 Training Details for All Approaches

We list all the hyper-parameters we tuned with grid-search in the validation set for all methods for `Wiki Comments`, `Civil Comments` and `Jigsaw Toxicity`.

- For FREEDRO, we tune the parameter delay interval T in {0, 1, 2} and the w_u and w_s of Equation 9 in {(0.1, 1.0), (0.2, 1.0), (0.5, 1.0), (1.0, 3.0), (1.0, 6.0), (1.0, 10.0)}.
- For JTT, we tune its pretraining epoch number T in {1, 2} and tune its up-weights λ_{up} in {3.0, 6.0, 10.0}.
- For ARL, we tune its warm-up epoch number T in {0, 1}.
- For POE and REWEIGHTING, we tune their pretraining epoch number. T in {1, 2}.
- For IPW-Z and IPW-ZY, we tune if the average of the weights are normalized to 1 or not.
- For GDRO, we implement it with a soft-weighted version similar to Equation 9 except for using the ground-truth demographic and labels to partition the groups. We tune w_u and w_s in {(0.0, 1.0), (0.1, 1.0), (0.2, 1.0), (0.5, 1.0), (1.0, 3.0), (1.0, 6.0), (1.0, 10.0)}. It's worth noticing that when w_u, w_s = (0.0, 1.0), the GDRO conforms the Equation 4.

Method	Validation Acc.	Best Param.
ERM	53.0	-
ARL	52.6	$T = 1$
JTT	53.7	$T = 5, \lambda_{up} = 3.0$
POE	52.3	$T = 2$
REWEIGHTING	54.0	$T = 10$
IPW-Z	52.7	Normalize = $False$
IPW-ZY	52.8	Normalize = $False$
GDRO	62.6	$w_u = 0.1, w_s = 1.0$
FREEDRO	56.0	$T = 50, w_u = 0.2, w_s = 1.0$

Table 4: The validation results of all approaches in `COMPAS`. *Validation Acc.* column indicates the worst group accuracy evaluated on the validation set. *Best Para.* indicates the best parameter in the validation set.

Method	Validation Acc.	Best Param.
ERM	81.6	-
ARL	80.3	$T = 1$
JTT	77.6	$T = 1, \lambda_{up} = 10.0$
POE	80.3	$T = 5$
REWEIGHTING	81.6	$T = 1$
IPW-Z	80.3	Normalize = $False$
IPW-ZY	80.3	Normalize = $False$
GDRO	81.6	$w_u = 1.0, w_s = 10.0$
FREEDRO	81.6	$T = 2, w_u = 1.0, w_s = 6.0$

Table 5: The validation results of all approaches in `Sexist Tweets`. *Validation Acc.* column indicates the worst group accuracy evaluated on the validation set. *Best Para.* indicates the best parameter in the validation set.

Method	Validation Acc.	Best Param.
ERM	47.9	-
ARL	49.1	$T = 1$
JTT	49.7	$T = 1, \lambda_{up} = 6.0$
POE	54.5	$T = 2$
REWEIGHTING	64.6	$T = 1$
IPW-Z	52.2	Normalize = $True$
IPW-ZY	51.4	Normalize = $False$
GDRO	47.2	$w_u = 1.0, w_s = 3.0$
FREEDRO	63.4	$T = 2, w_u = 0.1, w_s = 1.0$

Table 6: The validation results of all approaches in `Jigsaw Toxicity`. *Validation Acc.* column indicates the worst group accuracy evaluated on the validation set. *Best Para.* indicates the best parameter in the validation set.

Method	Validation Acc.	Best Param.
ERM	54.9	-
ARL	54.4	$T = 1$
JTT	67.7	$T = 1, \lambda_{up} = 3.0$
POE	66.7	$T = 1$
REWEIGHTING	66.5	$T = 1$
IPW-Z	55.5	Normalize = $True$
IPW-ZY	55.5	Normalize = $True$
GDRO	64.8	$w_u = 1.0, w_s = 10.0$
FREEDRO	69.0	$T = 2, w_u = 1.0, w_s = 6.0$

Table 7: The validation results of all approaches in `Civil Comments`. *Validation Acc.* column indicates the worst group accuracy evaluated on the validation set. *Best Para.* indicates the best parameter in the validation set.

Method	Validation Acc.	Best Param.
ERM	41.7	-
ARL	50.0	$T = 0$
JTT	57.1	$T = 2, \lambda_{up} = 3.0$
POE	50.0	$T = 2$
REWEIGHTING	47.6	$T = 1$
IPW-Z	57.1	Normalize $= False$
IPW-ZY	50.0	Normalize $= False$
GDRO	41.7	$w_u = 1.0, w_s = 10.0$
FREEDRO	50.0	$T = 0, w_u = 0.2, w_s = 1.0$

Table 8: The validation results of all approaches in `Wiki Comments`. *Validation Acc.* column indicates the worst group accuracy evaluated on the validation set. *Best Para.* indicates the best parameter in the validation set.

For `Sexist Tweets`, we tune the hyper-parameters with more values for the following methods:

- For JTT, we tune its pretraining epoch number T in {1, 2, 5, 10} and tune its up-weights λ_{up} in {3.0, 6.0, 10.0, 20.0, 50.0}.
- For ARL, we tune its warm-up epoch number T in {0, 1, 2, 3}.
- For POE and REWEIGHTING, we tune their pretraining epoch number T in {1, 2, 5, 10}.

For `COMPAS`, we tune the hyper-parameters with more values for the following methods:

- For FREEDRO, we tune the parameter dynamic lagged grouping frequency T in {0, 1, 2, 5, 10} and the w_u and w_s in Equation 9 in {(0.1, 1.0), (0.2, 1.0), (0.5, 1.0), (1.0, 3.0), (1.0, 6.0), (1.0, 10.0)}.
- For JTT, we tune its pretraining epoch number T in {1, 2, 5, 10} and tune its up-weights λ_up in {3.0, 6.0, 10.0}.
- For ARL, we tune its warm-up epoch number T in {0, 1, 2, 5, 10}.
- For POE and REWEIGHTING, we tune their pretraining epoch number T in {1, 2, 5, 10}.

All experiments are run on a 16G Tesla V100 GPU. The validation results and the best parameters for all approaches are in Table 8, 6, 7, 5 and 4.