Fair Enough: Standardizing Evaluation and Model Selection for Fairness Research in NLP

Xudong Han[1,2] **Timothy Baldwin**[1,2] **Trevor Cohn**[1]
[1]The University of Melbourne
[2]MBZUAI
xudongh1@student.unimelb.edu.au, {tbaldwin,t.cohn}@unimelb.edu.au

Abstract

Modern NLP systems exhibit a range of biases, which a growing literature on model debiasing attempts to correct. However current progress is hampered by a plurality of definitions of bias, means of quantification, and oftentimes vague relation between debiasing algorithms and theoretical measures of bias. This paper seeks to clarify the current situation and plot a course for meaningful progress in fair learning, with two key contributions: (1) making clear inter-relations among the current gamut of methods, and their relation to fairness theory; and (2) addressing the practical problem of model selection, which involves a trade-off between fairness and accuracy and has led to systemic issues in fairness research. Putting them together, we make several recommendations to help shape future work.[1]

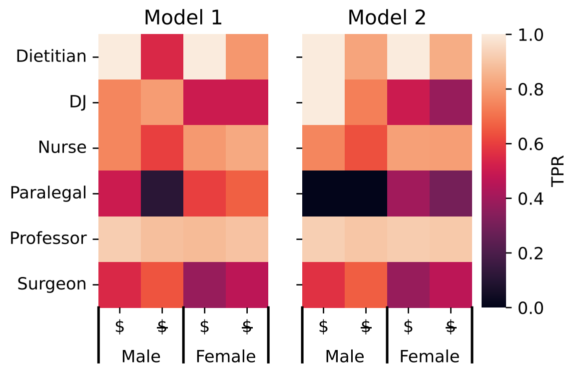


Figure 1: True positive rate (TPR) evaluation results over a biography classification dataset broken down by author demographic and selected profession classes. $ and ~~$~~ denote the economic status (wealthy vs. not, respectively). The pattern of results exhibits various biases, however it is difficult to distil this into a single figure of merit, and thus determine which is the better or fairer of the two models.

1 Introduction

In NLP and machine learning, there has been a surge of interest in fairness due to the fact that models often learn and amplify biases in the training dataset, leading to a range of harms (Badjatiya et al., 2019; Díaz et al., 2018). A central notion is group-wise fairness (Dwork et al., 2012; Chouldechova, 2017; Berk et al., 2021), which is typically measured as the model performance disparities across groups of data that are created by the combinations of protected attributes, such as race and gender. A broad range of bias evaluation metrics have been introduced in previous studies to capture different types of biases – such as *demographic parity* (Feldman et al., 2015) and *equal opportunity* (EO) (Hardt et al., 2016) – and different approaches have been adopted to both measure group disparities within each class, and aggregate over those disparities. Each of these choices implicitly encodes assumptions about the nature of fairness, but little work has been done to spell out what those assumptions are, or guide the selection of evaluation metric from first principles of what constitutes fairness.

As an illustration of this issue, Figure 1 depicts the true positive rate (TPR) values for two models.[2] Given that the EO fairness is satisfied if different groups achieve identical TPR, which model is fairer or "better" out of the two? The answer is far from clear, and in terms of evaluation practice, dictated by a series of choices which implicitly encode different assumptions about what fairness is.

In terms of research practice, these choices have led to a lack of consistency and direct empirical comparability between methods. Equally concerningly, given that fairness research involves an inherent trade-off between raw model performance and fairness, it has more subtly led to a lack of rigour in terms of how model selection has been carried out, meaning that methods are often deployed in

[1]Code available at https://github.com/HanXudong/Fair_Enough

[2]For further details see Appendix A.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 297–312
May 2-6, 2023 ©2023 Association for Computational Linguistics

suboptimal ways relative to a particular evaluation methodology.

In this paper, we seek to address these problems. We start by surveying current practices for fairness evaluation aggregation within an integrated framework, and discuss considerations and motivations for using different aggregation approaches. To ensure fairness metrics are fully comparable, we present a checklist for reporting fairness evaluation metrics, and also recommendations for aggregation method selections. We next survey model comparison methods, and demonstrate the issues stemming from using inconsistent model selection criteria. To ensure fair comparisons, we further introduce a metric for comparison without model selection, which measures the area under the trade-off curve of each method.

Overall this paper makes two key contributions: (1) we characterise current practices for fairness evaluation and their grounding in theory, proposing a best-practice checklist; and (2) we propose a new method which resolves several issues relating to model selection and comparison.

2 Related Work

In terms of bias metrics, there are mainly two lines of work in the literature on NLP fairness: bias in the geometry of text representations (intrinsic bias), and performance disparities across groups in downstream tasks (extrinsic bias), respectively. Based on the hypothesis that measuring and mitigating intrinsic bias will also reduce extrinsic bias, previous work has mainly focused on measuring and mitigating intrinsic bias, such as the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), Sentence Encoder Association Test (SEAT) (May et al., 2019), and Embedding Coherence Test (ECT) (Dev and Phillips, 2019). However, Goldfarb-Tarrant et al. (2021) recently showed that there is no reliable correlation between intrinsic and extrinsic biases, and suggest future work focusing on extrinsic bias measurement (which is the focus of this work).

As for bias mitigation, debiasing methods for intrinsic and extrinsic bias generally suffer from performance–fairness trade-offs controlled by particular hyperparameters such as the number of principal components used to define the intrinsic bias subspace (Bolukbasi et al., 2016), and the strength of addition objectives for performance parity across groups (Shen et al., 2022b). In measuring performance (perplexity and LM score for sentence embeddings, for example) and fairness simultaneously, the model comparison framework presented in this paper is generalizable for both intrinsic and extrinsic fairness.

3 Fairness Metrics

In this section, we discuss the considerations involved in fairness evaluation. We start with a survey of different methods for aggregating scores, and propose a two-step aggregation framework for fairness evaluation.

3.1 Formal Notation Preliminaries

We consider fairness evaluation in a classification scenario. Evaluation is based on a test dataset consisting of n instances $\mathcal{D} = \{(\boldsymbol{x}_i, \mathrm{y}_i, \mathrm{z}_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i$ is an input vector, $\mathrm{y}_i \in \{\mathrm{c}\}_{\mathrm{c}=1}^{\mathrm{C}}$ represents target class label, and $\mathrm{z}_i \in \{\mathrm{g}\}_{\mathrm{g}=1}^{\mathrm{G}}$ is the group label, such as gender.[3]

Given a model that has been trained to make predictions w.r.t. the target label $\hat{\mathrm{y}} = f(\boldsymbol{x})$, fairness evaluation metrics generally measure group-wise performance disparities for a particular metric $m(\mathrm{y}, \hat{\mathrm{y}})$. For example, positive predictive rate and true positive rate have been employed as the metric for demographic parity (Feldman et al., 2015) and equal opportunity (Hardt et al., 2016), respectively.

For each group, the results of a metric m are C-dimensional vectors, one dimension for each class. Given G protected groups, the full results are organized as a C $\times$ G matrix, denoted as $\boldsymbol{M}$. For the subset of instances $\mathcal{D}_{\mathrm{c,g}} = \{(\boldsymbol{x}_i, \mathrm{y}_i, \mathrm{z}_i) | \mathrm{y}_i = \mathrm{c}, \mathrm{z}_i = \mathrm{g}\}_{i=1}^{n}$, we denote the corresponding evaluation results as $M_{\mathrm{c,g}}$. Taking Figure 1 as an example, $\boldsymbol{M}$ refers to the heatmap plot, and $M_{\mathrm{c,g}}$ is the cell in the c-th row and g-th column.

Given $\boldsymbol{M}$, the question is how exactly to aggregate the result matrix as a single number that measures the degree of fairness. We split the aggregation into two steps: (1) group-wise aggregation, which aggregates evaluation results of all groups within a class ($[M_{\mathrm{c},1}, \ldots, M_{\mathrm{c,G}}]$) into a single number ($\beta_{\mathrm{c}}$); and (2) class-wise aggregation, which aggregates $[\beta_1, \ldots, \beta_{\mathrm{C}}]$ scores of all classes into a single number δ.[4]

[3] When considering multiple protected attributes, z can be intersectional identities as shown in Figure 1.

[4] Mathematically, it would be possible to do the class-wise aggregation first, and then the group-wise aggregation. However, aggregating class-wise performances within a particular group essentially measures the long-tail learning problem

Aggregation	Method	Description	Unit	Reference
Group-wise	mean gap	$\beta_c = \frac{1}{G}\sum_g \lvert M_{c,g} - \overline{M}_c\rvert$	G	(Shen et al., 2022b)
	variance	$\beta_c = \frac{1}{G-1}\sum_g \lvert M_{c,g} - \overline{M}_c\rvert^2$	G	(Lum et al., 2022)
	max gap	$\beta_c = \max_g \lvert M_{c,g} - \overline{M}_c\rvert$	G	(Yang et al., 2020)
	min score	$\beta_c = \min_g M_{c,g}$	S	(Lahoti et al., 2020)
	min ratio	$\beta_c = \min_g \frac{M_{c,g}}{\overline{M}_c}$	R	(Zafar et al., 2017)
	max difference	$\beta_c = \max_g M_{c,g} - \min_g M_{c,g}$	S	(Bird et al., 2020)
	max ratio	$\beta_c = \frac{\max_g M_{c,g}}{\min_g M_{c,g}}$	R	(Feldman et al., 2015)
	difference threshold (γ)	$\beta_c = \frac{1}{G}\sum_g \mathbb{1}_{[0,\gamma]}(\lvert M_{c,g} - \overline{M}_c\rvert)$	G	(Kearns et al., 2019)
	ratio threshold (γ)	$\beta_c = \frac{1}{G}\sum_g \mathbb{1}_{[0,\gamma]}(\lvert \frac{M_{c,g}}{\overline{M}_c} - 1\rvert)$	R	(Barocas et al., 2019)
Class-wise	binary	$\delta = \sum_c \beta_c \mathbb{1}_{\{1\}}(c)$	β	(Roh et al., 2021)
	quadratic mean	$\delta = \sqrt{\frac{1}{C}\sum_c \beta_c^2}$	β	(Romanov et al., 2019)
	mean	$\delta = \frac{1}{C}\sum_c \beta_c$	β	(Li et al., 2018)

Table 1: Summary of different aggregation approaches. Based on the basic unit, group-wise aggregations are additionally categorized into three types: **S**core ($M_{c,g}$), **G**ap ($\lvert M_{c,g} - \overline{M}_c\rvert$), and **R**atio ($\frac{M_{c,g}}{\overline{M}_c}$).

3.2 Existing Aggregation Approaches

Table 1 summarizes several aggregation approaches from previous work, which are categorized based on the level of aggregation.

3.2.1 Basic Unit

The **basic unit** refers to the inputs to an aggregation function.

Group-wise Broadly, there are three types of basic units for group-wise aggregation:

1. the original **score** ($M_{c,g}$), which maintains the actual performance level under aggregation and larger is better;
2. the **gap**, i.e., absolute difference, between the evaluation results of a group and the average performance ($\lvert M_{c,g} - \overline{M}_c\rvert$), where smaller is better and the minimum is 0; and
3. **ratio** of the evaluation results of a group to the average ($\frac{M_{c,g}}{\overline{M}_c}$), where closer to 1 is better.

Score describes the actual performance of each group, and is generally used to measure extrema of actual performances. For example, the *Rawlsian Max Min* criterion (Rawls, 2001) is satisfied if the utility of the worst-performing group is maximized. Related fairness notions are also known as per-group fairness (Hashimoto et al., 2018; Lahoti et al., 2020).

The other two units, **gap** and **ratio**, support the notion of group fairness, and evaluate whether or not ŷ is fair w.r.t. z. Taking EO (Hardt et al., 2016) as an example, it requires the true positive rate to be independent of z. Formally, for a particular class c, the EO criterion is satisfied iff

$$\mathrm{TPR}_{c,g} = \overline{\mathrm{TPR}}_c, \forall g \in \{g\}_{g=1}^{G}.$$

As such, it is straightforward to directly measure the absolute difference between $\mathrm{TPR}_{c,g}$ and $\overline{\mathrm{TPR}}_c$,

$$\mathrm{TPR}_{c,g} = \overline{\mathrm{TPR}}_c \Leftrightarrow \lvert \mathrm{TPR}_{c,g} - \overline{\mathrm{TPR}}_c\rvert = 0,$$

which is essentially the **gap** unit.

Alternatively, the **ratio** unit can be used to measure inequality as a percentage:

$$\mathrm{TPR}_{c,g} = \overline{\mathrm{TPR}}_c \Leftrightarrow \frac{\mathrm{TPR}_{c,g}}{\overline{\mathrm{TPR}}_c} = 1.$$

Ratio-based scores can also be interpreted via a "$q\%$-rule" (Zafar et al., 2017; Barocas et al., 2019), for example, the 80%-rule for *disparate impact* (Feldman et al., 2015), which requires that the ratio is no less than 0.8.

The $q\%$-rule can be captured more explicitly by a threshold (Kearns et al., 2018; Barocas et al., 2019), which is a relaxation of the equality based on a slack threshold $\epsilon \in \mathbb{R}^+$, $\lvert 1 - \frac{\mathrm{TPR}_{c,g}}{\overline{\mathrm{TPR}}_c}\rvert \leq \epsilon$. Similarly, the threshold can be applied to **gap**, resulting in $\lvert \mathrm{TPRc,g} - \overline{\mathrm{TPR}}_c\rvert \leq \epsilon$.

Class-wise The next step is class-wise aggregation, taking the group-wise aggregation for each class from above as inputs, $[\beta_1, \ldots, \beta_C]$.

3.3 Generalized Mean Aggregation

Before discussing each of these aggregation methods, we first introduce the basic concept of the

rather than fairness.

Power (p)	Formulation
$-\infty$	Minimum: $\min\{v_1, \ldots, v_n\}$
-1	Harmonic Mean: $\frac{n}{\sum_{i=1}^{n} v_i^{-1}}$
1	Arithmetic Mean: $\frac{1}{n}\sum_{i=1}^{n} v_i$
2	Quadratic Mean: $\sqrt{\frac{1}{n}\sum_{i=1}^{n} v_i^2}$
$+\infty$	Maximum: $\max\{v_1, \ldots, v_n\}$

Table 2: Commonly-used cases of generalized mean aggregation.

generalized mean as a framework for describing aggregation functions, and then make the link between the generalized mean and existing aggregation methods.

Formally, the generalized mean is defined as:

$$\mathcal{M}_p(v_1, \ldots, v_n) = \left(\frac{1}{n}\sum_{i=1}^{n} v_i^p\right)^{\frac{1}{p}},$$

where $v_i \in \mathbb{R}^+$ are positive real numbers to be aggregated, and p is the exponent parameter. A desired property of the generalized mean is its inequality, which states that,

$$\mathcal{M}_p(v_1, \ldots, v_n) > \mathcal{M}_{p'}(v_1, \ldots, v_n), \forall p > p'.$$

Essentially, a larger value of p encourage the aggregation to focus more on the larger-valued elements, which can be illustrated with the specific cases shown in Table 2.

By setting $p = \pm\infty$, generalized mean returns extremum values, including (in Table 1): (1) the maximum value of gap (Yang et al., 2020), difference (Bird et al., 2020), and ratio (Feldman et al., 2015); and (2) the minimum value of score (Lahoti et al., 2020) and ratio (Zafar et al., 2017).

For other p values, the generalized mean reflects the relative dispersion of its inputs. For example, group-wise mean gap aggregation (Shen et al., 2022b) and class-wise mean aggregation (Li et al., 2018) are both equivalent to $p = 1$. Class-wise quadratic mean aggregation (Romanov et al., 2019) is essentially $p = 2$, which focuses more on those classes with higher bias. Similarly, group-wise variance aggregation (Lum et al., 2022) is proportional to the $p = 2$ setting, implying that groups with larger gaps will influence results more.

The additional advantage of using generalized mean aggregation is that comparison across arbitrary p values can be easily stated. For example, group-wise aggregation is the $p = -5$ generalized mean with respect to the **score** units in a toxicity classification competition,[5] meaning that evaluation focuses more on groups with lower performance.

Other Aggregation Methods: Although generalized mean aggregation is a powerful tool for describing and interpreting the aggregation process, there are other ways that need further discussion. Previous work has also considered assigning different weights under aggregation, for instance, Kearns et al. (2018) assign larger weights to groups with larger populations. Such aggregations can be implemented as the weighted generalized mean: $\mathcal{M}_{p,w}(v) = \left(\frac{1}{n}\sum_{i=1}^{n} w_i v_i^p\right)^{\frac{1}{p}}$, where w is the weight vector, and $\sum_{i=1}^{n} w_i = 1$.

An example of the weighted generalized mean for class-wise aggregation is binary aggregation that only considers the positive class in a binary classification setting (Hardt et al., 2016; Zafar et al., 2017; Kearns et al., 2018; Zhao et al., 2019; Lahoti et al., 2020; Han et al., 2021; Lum et al., 2022). The positive class is often treated as the "advantaged" outcome, so the analysis focuses solely on the positive class. Moreover, the one-versus-all trick is not necessary for the binary setting, and natural derivations of the confusion matrix can be used to refer to a particular class, e.g., TPR for the positive class and TNR for the negative class.

3.4 Recommendations

We are now in a position to be able to provide recommendations for fairness evaluation.

Following the work of Dodge et al. (2019), we provide a checklist for fairness evaluation metric aggregation:

- ☐ Statistics of the dataset $\mathcal{D}$, e.g., the probability table of the joint distribution of y and z, and the size of each partition.
- ☐ The evaluation metric m (e.g., TPR for *EO* fairness).
- ☐ The basic unit of group-wise aggregation, including **score**, **gap**, and **ratio**, or other possible measures.
- ☐ The aggregation function for **group-wise** aggregation, and the corresponding motivation.

[5] Jigsaw Unintended Bias in Toxicity Classification: https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/

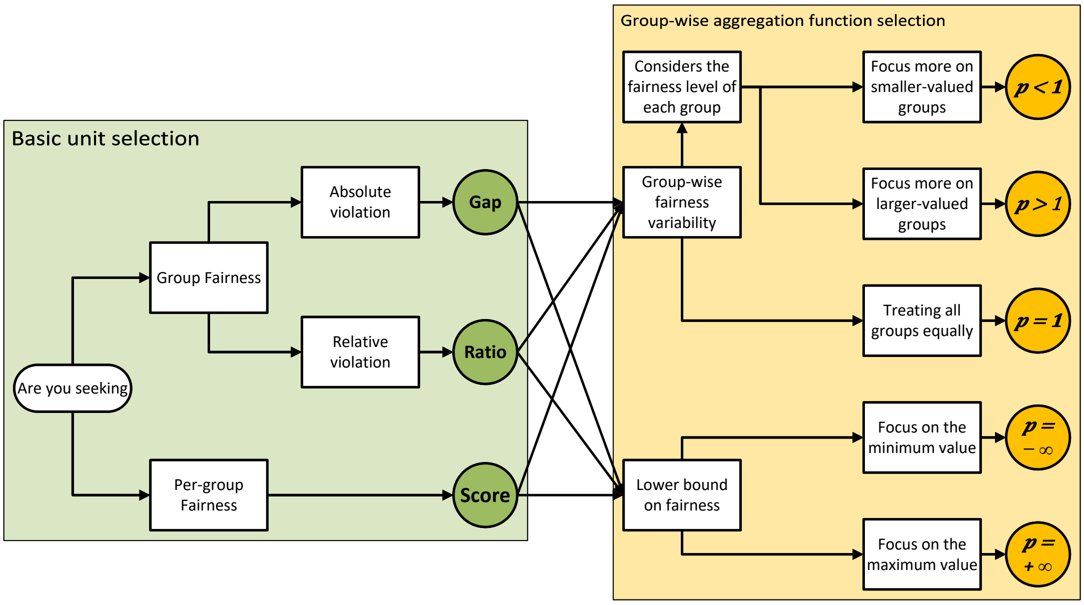


Figure 2: Decision path for exponent parameter selection for generalized mean aggregation.

☐ The aggregation function for **class-wise** aggregation, and the corresponding motivation.

Although the particular choice of evaluation dataset $\mathcal{D}$ and evaluation metric m are critical to the overall evaluation, they are not the main focus of this paper. Rather, we provide guidance based on the selection of basic unit, and methods for group- and class-wise aggregation, as detailed in Figure 2.

Basic Unit Selection: The circles in Figure 2 annotated as **Score**, **Ratio**, and **Gap** are the decision points for basic unit selection.

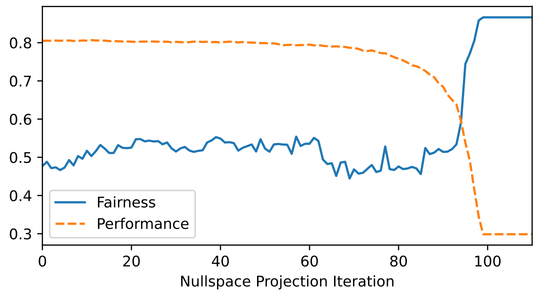
If per-group fairness is the primary criterion (e.g., Rawlsian Max-Min fairness (Rawls, 2001)), using **score** is the best practice, which maintains the original values under aggregation. On the other hand, if inter-group fairness is critical, **gap** and **ratio** are more appropriate choices. **Gap** reflects disparities in the same scale as the per-group scores, and is easy to visualize (e.g. as differences in height between clustered bars). However, if one wished to measure disparities in relative terms, e.g., the $q\%$-rule (Feldman et al., 2015), **ratio** is a better choice than **gap**.

Group-wise Aggregation Function Selection: The selection of group-wise aggregation functions is shown as the exponent parameters of the *generalized mean* aggregation.

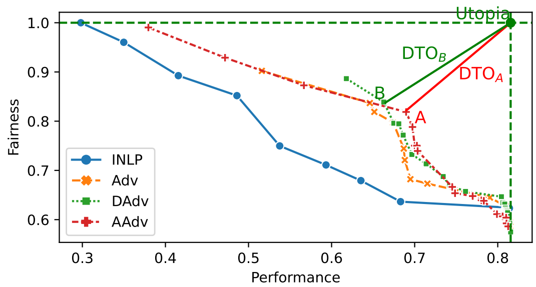
Measuring extrema is similar to the notion of per-group fairness, and encourages improvements in the worst-performing groups. For basic units where smaller is fairer, e.g., **gap**, aggregation generally focuses on the maximum (Yang et al., 2020), i.e., $p = +\infty$. For units like **score** ($p = -\infty$), on the other hand, the minimum value should be measured, as a lower bound.

Besides extrema, it is also reasonable to measure fairness variability across groups. A typical choice is taking the arithmetic mean (i.e., $p = 1$) across all groups, which implicitly assigns equal importance to each individual group. Similar to the signs in extremum aggregations, the value of p in variability aggregations should be selected based on the type of basic unit, to focus more on worse-performing groups. Taking the **gap** unit as an example, the quadratic mean ($p = 2$) is influenced more by larger gaps than the arithmetic mean. Moreover, quadratic mean aggregation based on **gap** is essentially the standard deviation of **scores**, and can be used to reconstruct variance aggregation (Lum et al., 2022).

Class-wise Aggregation Function Selection: Although our focus is on the fairness evaluation metric, class-wise aggregation is almost identical to aggregation methods for general utility metrics. Binary aggregation for fairness is the same as utility metrics, while mean aggregation (Li et al., 2018; Wang et al., 2019) for fairness evaluation is equivalent to "macro"-averaging in general evaluation.



(a) Tuning INLP trade-off hyperparameter



(b) Performance–fairness trade-offs

Figure 3: Figure 3a shows an example of performance and fairness with respect to different iterations of the nullspace projection in INLP. Figure 3b presents the Pareto frontiers in the performance–fairness trade-offs of four debiasing methods in recent work. Figure 3b also provides an illustration of DTO. The green dashed vertical and horizontal lines denote the best performance and fairness, respectively, and their intersection point is the *Utopia* point. The length of a line, e.g., the red line from A to the Utopia point, is the DTO for the corresponding candidate model.

4 Model Comparison

This section focuses on comparison of debiasing methods when considering utility and fairness simultaneously. We first introduce the performance–fairness trade-off curve (PFC) for debiasing methods, and then discuss the limitations of existing comparison frameworks. Finally, we propose a new metric, namely the area under the curve (AUC) w.r.t. PFC, which integrates existing approaches and reflects the overall goodness of a method.

4.1 Performance and Fairness Metrics

As discussed in Section 3, there are many options to measure performance and fairness. This paper is generalizable to all different metrics, but for illustration purposes, we follow Ravfogel et al. (2020); Subramanian et al. (2021) and Han et al. (2022c) in measuring the overall accuracy and equal opportunity fairness.

Specifically, equal opportunity fairness measures TPR disparities across groups, such as the situation depicted in Figure 1. We use the TPR **gap** across subgroups to capture absolute disparities. For group-wise aggregation, we treat all groups equally in computing the unweighted sum of *gap* scores ($\propto p = 1$). In the last step, class-wise aggregation, we focus more on less fair classes by using root mean square aggregation ($p = 2$).

4.2 Performance–Fairness Trade-off

It has been observed in previous work that a performance–fairness trade-off exists in bias mitigation (Li et al., 2018; Wang et al., 2019; Ravfogel et al., 2020; Han et al., 2022b; Shen et al., 2022b).

Typically, debiasing methods involve a trade-off hyperparameter to control the extent to which the model sacrifices performance for fairness. Examples of such trade-off hyperparameters include: (1) interpolation between the target and vanilla data distribution for pre-processing approaches (Wang et al., 2019; Han et al., 2022a); (2) the strength of additional loss terms for loss manipulation methods (Zhao et al., 2019; Lahoti et al., 2020; Han et al., 2021; Shen et al., 2022a); (3) the target level of fairness in constrained optimization (Kearns et al., 2018; Subramanian et al., 2021); and (4) the number of debiasing iterations for post-hoc bias mitigation methods (Ravfogel et al., 2020).

Taking INLP (Ravfogel et al., 2020) as an example, which debiases by iteratively projecting the text embeddings to the nullspace of the protected attributes, Figure 3a shows performance and fairness with respect to the number of nullspace projection iterations.[6] It is clear that more iterations lead to better fairness at the cost of performance.

Instead of looking at performance/fairness for different trade-off hyperparameter values, it is more meaningful to focus on the Pareto frontiers in trade-off plots (Figure 3b), where each point corresponds to a particular value of the trade-off hyperparameter in Figure 3a. The frontiers represent the best fairness that can be achieved at different performance levels, and vice versa.

One limitation of a trade-off plot is that it is hard to make quantitative conclusions based on the plot itself, and we cannot conclude that one method is better than another if there exists any intersection of their trade-off curves. As shown in Figure 3b, in addition to INLP, we also include the trade-

[6] Without loss of generality, we assume that for both fairness and performance, larger is better.

off curves for three recent adversarial debiasing variants: ADV (Li et al., 2018), DADV (Han et al., 2021), and A-ADV (Han et al., 2022b). Although A-ADV is better than the other methods under most conditions, there exist intersections between their trade-off curves. As such, we can only state that A-ADV is better than other methods within particular ranges, which is insufficient for making a precise comparison, especially when comparing multiple debiasing methods (as demonstrated in Figure 3b).

4.3 Model Selection

In order to conduct quantitative comparisons across different debiasing methods, current practice is to select a particular point on the frontier for each method, and then compare both the performance and fairness of the selected points.

One problem associated with model selection is that typically, no single method simultaneously achieves the best performance and fairness. For example, as shown in Figure 3b, if points **A** and **B** were the selected models for A-ADV and DADV, respectively, **A** would represent better performance and **B** better fairness. As such, although we have actual numbers for quantitative comparison, it is still hard to conclude which method is best.

Distance to the Optimal: To address this problem, we propose to measure the **D**istance **T**o the **O**ptimal point ("DTO") to quantify the performance–fairness trade-off (Salukvadze, 1971; Marler and Arora, 2004; Han et al., 2022a). A model is said to outperform others if it achieves a smaller DTO, i.e. the distance to the optimal (Utopia) point (the point at which performance and fairness are the maximum possible values) is minimized. Figure 3b illustrates the calculation of DTO for **A** and **B**, where the optimal point is the top-right corner[7] and DTO is measured by the normalized Euclidean distance (the length of the green and red lines) to the optimal point.

A notable advantage of DTO is that a Pareto improvement implies a smaller value of DTO. Therefore, DTO can be seen as relaxation of Pareto improvements, and the smallest DTO must be achieved by a point on the Pareto frontier. A key limitation of DTO is that it quantifies the trade-off of a single model rather than the full frontier, presupposing some means of model selection. This has been somewhat arbitrary in prior work, which is the problem we now seek to address.

[7]The location of the utopia point and the scale of metrics are discussed in Appendix C.

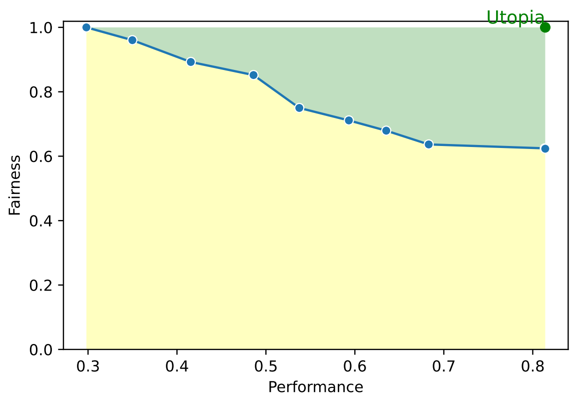


Figure 4: AUC of the performance–fairness trade-off curve. Taking the trade-off curve of INLP as an example, the yellow shaded area refers to the AUC-PFC, and the green shaded area refers to the integration of DTO in polar coordinates.

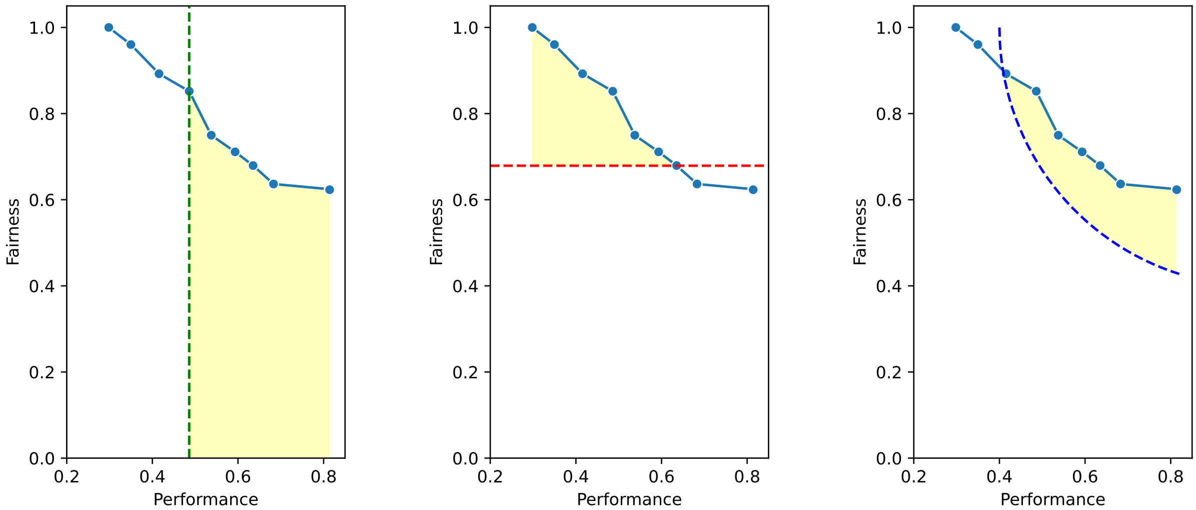
Selection Criteria: Similar to the aggregation of fairness metrics, model selection should be done in a domain-specific manner. Previous work has used different criteria for model selection, including: (1) minimum loss (Hashimoto et al., 2018; Li et al., 2018); (2) maximum utility (Lahoti et al., 2020), e.g., based on accuracy or F-measure; (3) manual selection based on visual inspection of the trade-off curve (Elazar and Goldberg, 2018; Ravfogel et al., 2020); (4) constrained selection (Han et al., 2021; Subramanian et al., 2021), by selecting the best fairness constrained to a particular level of performance, and vice versa; and (5) minimising DTO (Han et al., 2022b; Shen et al., 2022b).

Selection based on minimum loss and maximum utility is identical to classic model selection, and does not consider fairness explicitly. The other three types of criteria are based on trade-offs, differentiated by the method for aggregating fairness and performance.

Such inconsistency in model selection makes it very hard to rigorously compare methods. The question we want to address is: *how can we quantitatively compare methods without model selection?*

4.4 AUC-PFC

Recall that DTO is a metric for measuring the goodness of the trade-off of a particular model, and model selection is a process for selecting a particular frontier model from the Pareto curve. To address the problem associated with model selec-



(a) Accuracy is better than 0.49.
(b) Fairness is better than 0.68.
(c) DTO is better than 0.60.

Figure 5: Yellow shaded area denote the partial AUC-PFC score computed in the region where a particular condition applied.

tion, we propose to integrate DTO over the whole performance–fairness curve (PFC). Specifically, we integrate DTO in a polar coordinate system, where the reference point (pole) is the optimal point. Given that the DTO of a point on the trade-off curve is also the distance from the pole in the polar coordinate system, the trade-off curve can be treated as a function that maps angular coordinates to DTO. For example, as shown in Figure 4, the green area denotes the region enclosed by the performance–fairness trade-off curve of INLP and the utopia point with fairness = 1.00 and performance = 0.82.[8]

Alternatively, we can interpret the proposed metric from the performance–fairness perspective, in calculating the area under the Pareto curve, and subtracting this from the area under the optimal Pareto curve defined by the optimal point.

The magnitude of AUC-PFC differs from a single metric; for example, a 0.0001 improvement in the AUC-PFC score is equivalent to a 1 percentage point (pp) boost in both performance and fairness (0.01×0.01).

Partial AUC-PFC In practice, worse performance or fairness can be unacceptable, for example, one may want to prioritize fairness in particular applications. To address this problem, we present the Partial AUC-PFC score to focus on a specific region of the PFC Curve, where the AUC-PFC score is computed w.r.t. specific acceptable levels of performance and fairness.

Figure 5a shows an example of partial AUC-PFC, where the region can be considered if corresponding accuracy is better than 0.49. Similarly, Figures 5b and 5c show partial AUC-PFC scores with respect to particular fairness and DTO constraints.

With the partial AUC-PFC metric, one can explicitly compare different methods with a single number, and w.r.t. particular values of performance and fairness.

4.5 Case Studies

Experimental Details: We conduct experiments over the BIOS dataset (De-Arteaga et al., 2019) which was augmented with economic status by Subramanian et al. (2021), resulting in 28 professions as the target label and 4 intersectional demographic groups.[9] We use public implementations for all models in our experiments, primarily in *fairlib* (Han et al., 2022c).

Results: In Table 3, we investigate 7 different selection criteria and report the DTO score over the test set. Specifically, we conduct model selection over the development set based on: (1) minimum DTO; (2) maximum performance (**P**); (3) maximum performance within a fairness threshold of 5% improvement (**P@F+5%**); (4) maximum

[8] In the interests of consistent comparison, the Utopia point is typically $(1, 1)$, as in Table 3. In practice, this does not affect the calculation of AUC-PFC, as we discuss in Appendix C.

[9] Performance and fairness metrics have been introduced in Section 4.1.

Method	Selection Criteria							AUC↑
	DTO	P	P@F+5%	P@F+10%	F	F@P−5%	F@P−10%	
INLP (Ravfogel et al., 2020)	41.9	**41.9**	52.6	52.6	70.2	41.9	41.9	39.8
ADV (Li et al., 2018)	39.0	44.6	43.3	41.8	49.4	41.2	**41.2**	43.6
DADV (Han et al., 2021)	37.9	44.7	41.0	40.5	**39.9**	**40.4**	41.9	**44.5**
A-ADV (Han et al., 2022b)	**36.9**	45.4	**39.5**	**39.0**	62.1	43.8	42.8	44.0

Table 3: DTO scores of selected models over the BIOS dataset (smaller is better), based on the distances from mean performance and fairness to (1,1) over the test set. Models are selected based on the criterion listed for each column over the development set. The final column is the AUC, which does not involve model selection. **Bold** = the best score per column. See Appendix B for the full results.

performance within a fairness threshold of 10% improvement (**P@F+10%**); (5) maximum fairness (**F**); (6) maximum fairness within a performance trade-off threshold of 5% (**F@P−5%**); and (7) maximum fairness within a performance trade-off threshold of 10% (**F@P−10%**). Criteria (3), (4), (6), and (7) are constrained selections as discussed in Section 4.3, where we select the model with the highest fairness/performance within 5/10% of performance trade-off/fairness improvement relative to the STANDARD model. Taking F@P−10% as an example, the model with the highest fairness is selected within 10% performance trade-off over the vanilla model performance (i.e., with performance greater than 72% (82% − 10%)). Similarly, P@F+5% selects the model with highest performance subject to at least 5% fairness improvement over the vanilla model ($63\% = 58\% + 5\%$).

It can be seen that each of the four methods is the best for at least one selection criteria, as a stark illustration of our claim about model selection criteria biasing any possible conclusions about which method is best. For example, INLP and ADV are the best methods with respect to selection criteria **P** and **F@P−10%**, respectively.

On the contrary, our proposed AUC-PFC score (in the final column of Table 3) is unaffected by model selection, and reflects the overall trade-off of a method. Consistent with the trend in Figure 3b, AUC scores in Table 3 are smaller for worse-performing methods, e.g., INLP, and larger for better-performing methods such as A-ADV and DADV. Moreover, as the trade-off curves for A-ADV and DADV overlap one another (see Figure 3b), it is hard to pick a winner visually, let alone make quantitative comparisons. By using the AUC-PFC metric, we can conclude that overall, DADV is slightly better than A-ADV over this dataset.

Discussion The current DTO calculation assumes that users have no preference for performance over fairness or vice versa, where in practice it is possible that the choice of the fairness metric could be influenced by task-specific goals or the relative importance of fairness. Such problems have been widely studied in the literature on multi-objective learning, and a typical line of work is weighted generalized mean, which incorporates additional weight parameters in the generalized mean framework to reflect the importance or preference of each objective.

5 Conclusion

We have discussed the current practice in evaluation, model selection, and method comparison in the fairness literature, and shown how current practice in experimental fairness lacks rigour and consistency. We made recommendations for selecting a fairness evaluation metric, and introduced a new metric for measuring the overall performance–fairness trade-off of a method.

Acknowledgements

We thank Lea Frermann, Aili Shen, and Shivashankar Subramanian for their discussions and inputs. We thank the anonymous reviewers for their helpful feedback and suggestions. This work was funded by the Australian Research Council, Discovery grant DP200102519.

Limitations

This paper focuses on the notion of group fairness, under the assumption that each individual belongs to a particular demographic group. One limitation of methods in this space is that the demographic attributes must be observed (for the development and test data, at least) in order to evaluate fairness.

We only investigate the proposed evaluation aggregation framework in a classification setting. However, our framework is naturally generalizable to other tasks with discrete outcomes, such as generation and sequential tagging. Moreover, in terms of continuous labels, such as regression, one can skip the class-wise aggregation.

Ethical Considerations

This work focuses on current practice in fairness evaluation and method comparison. Our proposed "checklist" recommendations are specific to the fairness literature and complement existing frameworks, to encourage future research to think carefully about harms and what type of fairness is appropriate.

Demographics are assumed to be available only for evaluation purposes and are not used for model training or inference. We only use attributes that the user has self-identified in our experiments. All data and models in this study are publicly available and used under strict ethical guidelines.

References

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

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. http://www.fairmlbook.org.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

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

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

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.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR.

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.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

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.

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.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

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

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022a. Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11335–11350, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022b. Towards equal opportunity fairness through adversarial learning. *arXiv preprint arXiv:2203.06317*.

Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. 2022c. FairLib: A unified framework for assessing and improving fairness. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 60–71, Abu Dhabi, UAE. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

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

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

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*, volume 33, pages 728–740.

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 (Volume 2: Short Papers)*, pages 25–30.

Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-biasing "bias" measurement. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Seoul, Republic of Korea. ACM.

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

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

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

John Rawls. 2001. *Justice as Fairness: A Restatement*. Harvard University Press.

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*.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. 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 4187–4195.

M Ye Salukvadze. 1971. Concerning optimization of vector functionals. I. Programming of optimal trajectories. *Avtomat. i Telemekh*, 8:5–15.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022a. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022b. Optimising equal opportunity fairness in model training. In *Proceedings of* *the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4073–4084, Seattle, United States. Association for Computational Linguistics.

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

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 International Conference on Computer Vision*, pages 5310–5319.

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

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR.

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

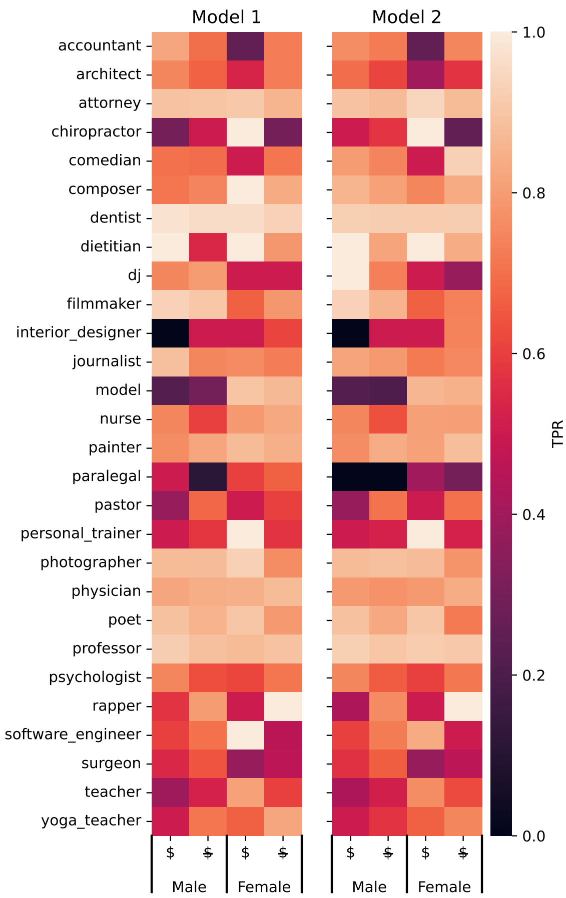


Figure 6: True positive rate (TPR) evaluation results over a biography classification dataset broken down by author demographics and profession classes. $ and ~~$~~ denote the economic status (high vs. low, respectively).

Profession	Total	Male		Female	
		$	~~$~~	$	~~$~~
professor	21715	46	9	37	7
physician	7581	42	8	41	8
attorney	6011	51	10	33	6
photographer	4398	53	11	30	6
journalist	3676	41	9	41	9
nurse	3510	8	1	76	15
psychologist	3280	31	6	52	11
teacher	2946	35	6	49	10
dentist	2682	52	11	30	6
surgeon	2465	73	12	13	2
architect	1891	64	12	21	3
painter	1408	47	9	36	8
model	1362	15	2	70	13
poet	1295	46	7	39	8
software engineer	1289	70	14	14	2
filmmaker	1225	56	10	29	6
composer	1045	70	14	14	2
accountant	1012	55	9	29	6
dietitian	730	5	1	82	12
comedian	499	69	9	19	3
chiropractor	474	62	14	21	3
pastor	453	59	15	23	4
paralegal	330	12	3	70	15
yoga teacher	305	13	3	71	12
interior designer	267	16	4	67	12
personal trainer	264	41	10	42	7
DJ	244	71	16	11	2
rapper	221	75	15	9	1
Total	72578	9	45	7	39

Table 4: Training set distribution of the BIOS dataset. For each profession, the table shows the number of individuals and the breakdown across demographics as a percentage. $ and ~~$~~ denote the economic status (high vs. low, respectively).

A Full Disaggregated Results

Figure 6 depicts the full TPR scores for two real-world models over a profession classification dataset, stratified across 4 protected attributes (male vs. female and high vs. low economic status) from Figure 1. Specifically, the two models are both trained naively without debiasing. They share the same hyperparameter settings and random seed, except that Model 1 and Model 2 are the 9th and 5th epochs, respectively. For professions such as Professor, there is little discernible difference either between the two models or across different combinations of protected attributes. For DJ, on the other hand, Model 1 appears to be reasonably fair w.r.t. economic status but biased for binary gender, whereas Model 2 is biased across both protected attributes but appears to have the higher overall TPR. Finally, with Paralegal, Model 2 appears to be fairer w.r.t. both economic status and binary gender but perform substantially worse than the more biased Model 1 in terms of the individual TPR scores for every combination of protected attributes. So it is hard to tell which model is fairer or "better" out of the two, without aggregation.

A.1 Dataset: BIOS

All experiments are based on a biography classification dataset (De-Arteaga et al., 2019; Ravfogel et al., 2020), where biographies were scraped from the web, and annotated for the protected attribute of binary gender and target label of 28 profession classes.

Besides the binary gender attribute, we additionally consider economic status as a second protected attribute. Subramanian et al. (2021) semi-automatically labeled economic status based on the individual's home country (wealthy vs. rest of world), as geotagged from the first sentence of the biography. For bias evaluation and mitigation, we consider the intersectional groups, i.e., the Cartesian product of the two protected attributes, leading to 4 intersectional classes: female–wealthy, female–rest, male–wealthy, and male–rest.

Since the data is not directly available, in order to construct the dataset, we use the scraping scripts of Ravfogel et al. (2020), leading to a dataset with 396k biographies.[10] Following Ravfogel et al. (2020), we randomly split the dataset into train (65%), dev (10%), and test (25%).

The augmentation for economic attributes follows previous work (Subramanian et al., 2021), which results in approximately 30% of instances that are labelled with both protected attributes.

Table 4 shows the target label distribution and protected attribute distribution.

A.2 Experimental Details

This work focuses on evaluation and model comparison in the fairness literature. Instead of training models from scratch, we use existing checkpoints from previous work (Han et al., 2022c), which are publicly available online.[11] Please refer to the original work (Han et al., 2022c) for experimental details.

A.3 Subset Confusion Matrices

Figure 7 presents the confusion matrices of all 4 subgroups. For each confusion matrix, the i-th row and j-th column entry indicates the number of samples which have the true label of the i-th class and predicted label of the j-th class. Since the distributions of classes within each group can be highly imbalanced, without further normalization and aggregation, it is difficult to draw any conclusion by just observing the number of samples in each cell.

A.4 Fairness Reproducibility

So far, we have listed critical factors underlying the choice of fairness metric, and provided recommendations for metric selection. However, we acknowledge that, in actual applications, the selection should be made in a domain-specific manner in close consultation with stakeholders or policymakers. In practice, countless types of fairness evaluation metrics could be derived from different combinations of aggregation methods.

Instead of reporting all possible fairness metrics, we suggest providing a set of confusion matrices for classification tasks, as it can form the basis of calculating a large number of metrics, including PPR, TPR, TNR, accuracy, and F-measure. The other key advantage of reporting confusion matrices is that the number of reported values is generally much smaller than the model or dataset size. Given a C-class classification dataset with G distinct protected groups, the combined size of the confusion matrices is $\mathrm{G} \times \mathrm{C}^2$ (one confusion matrix per group). Taking the BIOS dataset as an example, the sizes of the confusion matrices, test dataset, and model parameters (for a BERT-base classifier (Devlin et al., 2019)) are approximately 3×10^3, 4×10^4, and 1×10^8, respectively.

B Full Results of Case Studies

Table 5 shows the experimental results for both the test and development sets.

C AUC-PFC Extension

C.1 Weighted DTO

On the one hand, as suggested by Marler and Arora (2004), if fairness and performance have different scales, the Euclidean distance is not a suitable mathematical representation of closeness, resulting in worse approximation of Pareto optimality and efficiency. Therefore, the scales of performance and fairness should be normalized.

C.2 Selection of Utopia Point

Typically, most debiasing methods will share the same maximum performance, which is the performance of the vanilla model (corresponding to a hyperparameter setting where the debiasing method does nothing.) Accordingly this is a sensible choice for the performance of the Utopia point, as we have proposed for model selection. In terms of the calculation of areas of integration, moving the Utopia point to $(1, 1)$ has little effect, simply adding a constant triangular region which is identical for all methods, and thus irrelevant for model comparison. As such, it makes no difference whether we use 1 or the maximum-achieved model performance when comparing models based on AUC-PFC.

Distance to Arbitrary Ideal Point: Compared to the default value of DTO, moving the utopia point to the right (e.g., the $(1, 1)$ point) prioritizes methods with higher performance.

[10] There are slight discrepancies in the dataset composition due to data attrition: the original dataset (De-Arteaga et al., 2019) had 399k instances, while 393k were collected by Ravfogel et al. (2020).

[11] Bios_both at `https://github.com/HanXudong/fairlib/tree/main/analysis/results`

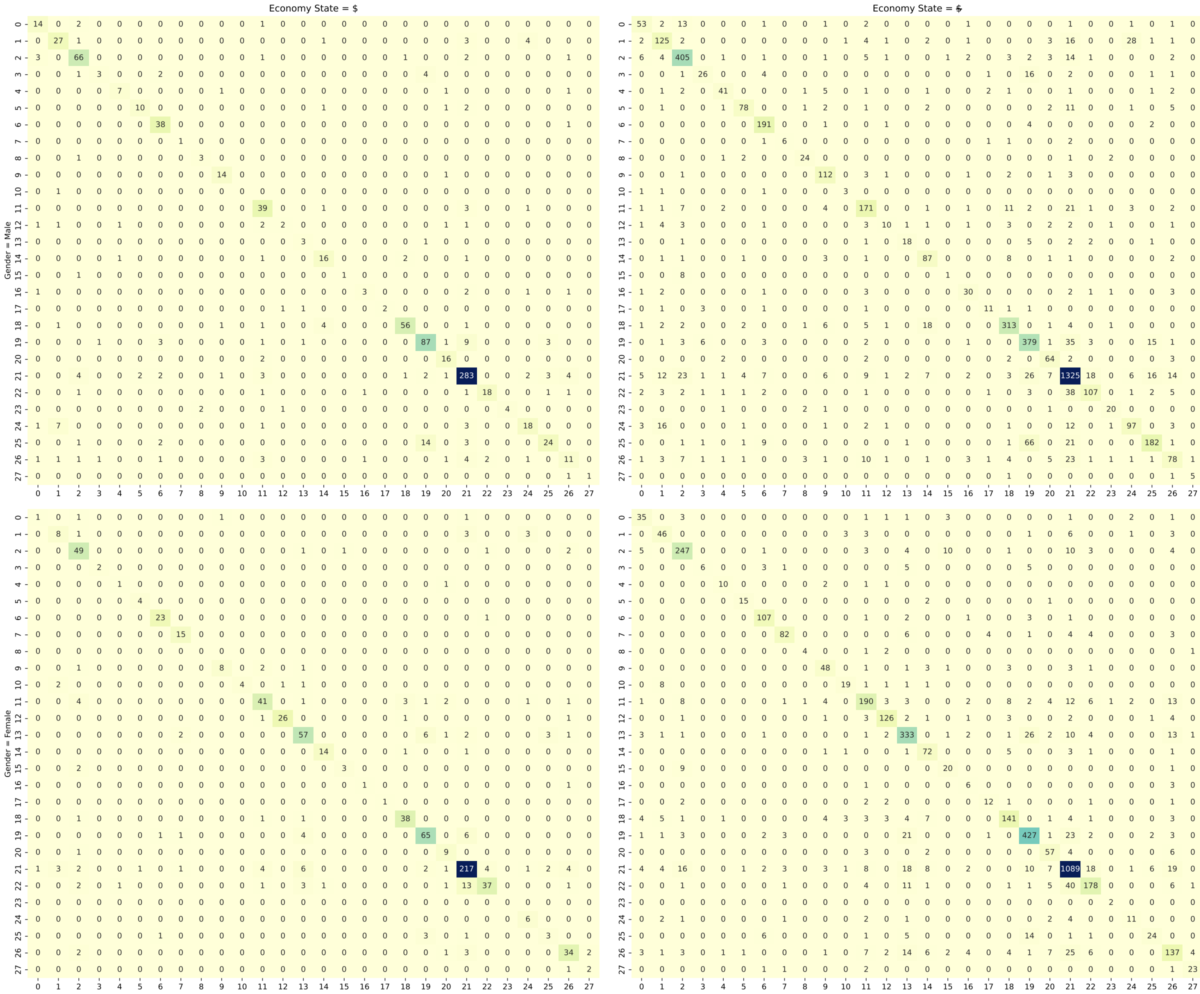


Figure 7: Confusion matrices of subgroups. Following Figure 6, confusion matrices are based on the predictions of Model 1, and class labels 0 to 27 are in the same order as the 28 professions in Figure 6.

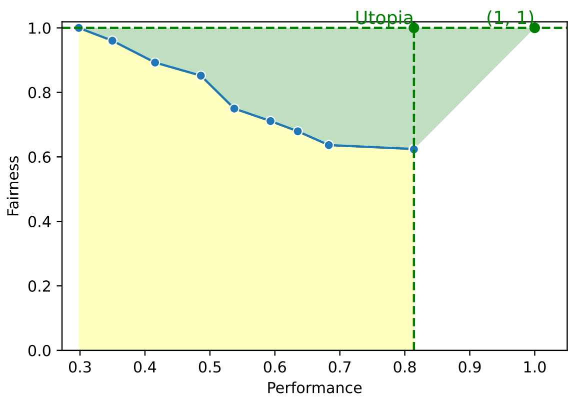


Figure 8: Integration of DTO with respect to an arbitrary ideal point, which is (1, 1) in this example.

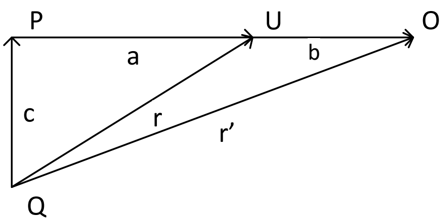


Figure 9: The influence of using different optimal points. Uppercase (Q, P, U, and O) and lowercase (c, a, b, r, and r') characters represent points and Euclidean distances between points.

As shown in Figure 9, without loss of generality, let

- $Q = (0, 0)$ denote the candidate point;
- $U = (c, a)$ denote the Utopia point, where c is the fairness distance from Q to the maximum fairness (which is 1), and a is the performance distance from Q to the maximum performance (which is 0.82 is Figure 9); and
- $O = (c, a + b)$ denote the arbitrary model,

Selection	Method	Test Set Performance	Test Set Fairness	Test Set DTO	Development Set Performance	Development Set Fairness	Development Set DTO
DTO	INLP	81.4 ± 0.0	62.4 ± 0.0	41.9	80.3 ± 0.0	54.4 ± 0.0	49.7
	ADV	64.6 ± 4.5	83.7 ± 1.1	39.0	63.9 ± 4.4	79.9 ± 3.0	41.3
	DADV	68.1 ± 5.5	79.5 ± 6.8	37.9	67.6 ± 5.5	75.5 ± 6.7	40.7
	A-ADV	69.7 ± 4.9	78.8 ± 7.7	36.9	69.3 ± 4.7	75.2 ± 5.4	39.5
P	INLP	81.4 ± 0.0	62.4 ± 0.0	41.9	80.3 ± 0.0	54.4 ± 0.0	49.7
	ADV	81.5 ± 0.2	59.5 ± 1.7	44.6	80.7 ± 0.2	55.2 ± 1.3	48.8
	DADV	81.4 ± 0.3	59.3 ± 1.7	44.7	80.6 ± 0.3	54.9 ± 1.7	49.1
	A-ADV	81.3 ± 0.3	58.6 ± 2.0	45.4	80.6 ± 0.3	54.4 ± 1.6	49.6
P@F+5%	INLP	53.8 ± 0.0	75.0 ± 0.0	52.6	53.8 ± 0.0	74.4 ± 0.0	52.8
	ADV	71.5 ± 5.7	67.3 ± 5.4	43.3	71.3 ± 5.8	66.2 ± 4.2	44.4
	DADV	73.4 ± 3.8	68.8 ± 4.9	41.0	73.0 ± 3.7	66.0 ± 3.7	43.5
	A-ADV	70.4 ± 5.7	73.9 ± 8.9	39.5	70.0 ± 5.8	67.4 ± 6.6	44.3
P@F+10%	INLP	53.8 ± 0.0	75.0 ± 0.0	52.6	53.8 ± 0.0	74.4 ± 0.0	52.8
	ADV	68.8 ± 4.8	72.1 ± 5.6	41.8	68.2 ± 4.8	70.3 ± 6.5	43.6
	DADV	69.7 ± 2.9	73.2 ± 6.8	40.5	69.2 ± 2.7	70.6 ± 4.0	42.6
	A-ADV	70.0 ± 3.3	75.0 ± 3.9	39.0	69.4 ± 3.4	70.8 ± 2.1	42.2
F	INLP	29.8 ± 0.0	100.0 ± 0.0	70.2	29.9 ± 0.0	86.6 ± 0.0	71.4
	ADV	51.6 ± 16.5	90.2 ± 9.3	49.4	51.2 ± 16.2	81.0 ± 6.0	52.4
	DADV	61.8 ± 3.7	88.6 ± 3.7	39.9	61.2 ± 3.5	82.2 ± 2.5	42.7
	A-ADV	37.9 ± 9.1	99.0 ± 1.2	62.1	37.6 ± 8.8	86.5 ± 0.2	63.8
F@P−5%	INLP	81.4 ± 0.0	62.4 ± 0.0	41.9	80.3 ± 0.0	54.4 ± 0.0	49.7
	ADV	79.1 ± 1.1	64.5 ± 1.3	41.2	78.5 ± 0.9	58.7 ± 2.4	46.5
	DADV	80.4 ± 0.5	64.7 ± 1.4	40.4	79.9 ± 0.5	57.4 ± 1.1	47.1
	A-ADV	79.9 ± 2.2	61.1 ± 2.8	43.8	79.1 ± 2.2	58.1 ± 2.8	46.8
F@P−10%	INLP	81.4 ± 0.0	62.4 ± 0.0	41.9	80.3 ± 0.0	54.4 ± 0.0	49.7
	ADV	79.1 ± 1.1	64.5 ± 1.3	41.2	78.5 ± 0.9	58.7 ± 2.4	46.5
	DADV	74.2 ± 3.2	67.0 ± 2.4	41.9	73.7 ± 3.3	64.2 ± 1.1	44.4
	A-ADV	74.9 ± 5.1	65.4 ± 6.9	42.8	74.3 ± 5.3	64.2 ± 5.4	44.1

Table 5: Evaluation results ± standard deviation (%) of selected models over the BIOS dataset. DTO scores are the distance from mean performance and fairness to (1,1) over the test set.

where $b > 0$, e.g., $b = (1 - 0.82)$ for the running example.

Before discussing the influence of the optimum point selection, recall that the magnitude of vector sum, $|\vec{v}| = |\vec{v}_1 + \vec{v}_2|$ is:

$$|\vec{v}| = \sqrt{|\vec{v}_1|^2 + |\vec{v}_2|^2 + 2|\vec{v}_1||\vec{v}_2|\cos\alpha},$$

where α is the angle between $\vec{v}_1$ and $\vec{v}_2$.

Let QU denote the vector from candidate model Q to the Utopia point U, the DTO based on the Utopia point is the $r = \sqrt{a^2 + c^2}$.

When calculating DTO based on the arbitrary optimum point O, $r' = |QU + UO|$, which can be shown as:

$$r' = \sqrt{r^2 + b^2 + 2rb\cos\alpha'},$$

where α' is the angle between QU and UO, and is equivalent to $\angle PUQ$. Furthermore, as discussed in Section 4.4, given a trade-off curve, the DTO is a function of $\angle PUQ$, i.e., the green shaded area is $\int_0^{\pi/2} \text{DTO}(\angle PUQ) d\angle PUQ$.

Lemma C.1. *Let Q_1 and Q_2 be two models with the same DTO score ($r_1 = r_2$), r'_1 and r'_2 be the DTO to the new Utopia point O. If the performance of Q_1 is worse than Q_2, then $r'_1 > r'_2$.*

Proof. Assuming that $r'_1 > r'_2$,

$$\begin{aligned} |Q_1U + UO| &> |Q_2U + UO| \\ |Q_1U + UO|^2 &> |Q_2U + UO|^2 \\ 2r_1 b\cos\angle PUQ_1 &> 2r_2 b\cos\angle PUQ_2 \end{aligned} \quad (1)$$

Since $\angle PUQ \leq \pi/2, \forall Q$, and $r_1 = r_2$,

$$a_1 = r_1\cos\angle PUQ_1 > a_2 = r_2\cos\angle PUQ_2,$$

where a_1 and a_2 are the performance distances from Q_1 and Q_2 to the maximum performance, respectively.

□