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 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.

Figure 1: True positive rate (TPR) evaluation results over a biography classification dataset broken down by author demographic and selected profession classes. $\$\$\$ 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.

For further details see Appendix A.
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 \( D = \{(x_i, y_i, z_i)\}_{i=1}^n \), where \( x_i \) is an input vector, \( y_i \in \{c\}_{c=1}^C \) represents target class label, and \( z_i \in \{g\}_{g=1}^G \) is the group label, such as gender.\(^3\)

Given a model that has been trained to make predictions w.r.t. the target label \( \hat{y} = f(x) \), fairness evaluation metrics generally measure group-wise performance disparities for a particular metric \( m(y, \hat{y}) \). 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 \( M \). For the subset of instances \( D_{c,g} = \{(x_i, y_i, z_i)\}_{i=1}^n \), we denote the corresponding evaluation results as \( M_{c,g} \). Taking Figure 1 as an example, \( M \) refers to the heatmap plot, and \( M_{c,g} \) is the cell in the \( c \)-th row and \( g \)-th column.

Given \( M \), 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 \( \langle M_{c,1}, \ldots, M_{c,G} \rangle \) into a single number \( \langle \beta_c \rangle \); and (2) class-wise aggregation, which aggregates \( \langle \beta_1, \ldots, \beta_C \rangle \) scores of all classes into a single number \( \delta \).\(^4\)

\(^3\)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.
<table>
<thead>
<tr>
<th>Aggregation</th>
<th>Method</th>
<th>Description</th>
<th>Unit</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group-wise</td>
<td>mean gap</td>
<td>$\beta_c = \frac{1}{G} \sum_{g}</td>
<td>M_{c,g} - \bar{M}_c</td>
<td>\ $</td>
</tr>
<tr>
<td></td>
<td>variance</td>
<td>$\beta_c = \frac{1}{G} \sum_{g}</td>
<td>M_{c,g} - \bar{M}_c</td>
<td>^2$</td>
</tr>
<tr>
<td></td>
<td>max gap</td>
<td>$\beta_c = \max_{g}</td>
<td>M_{c,g} - \bar{M}_c</td>
<td>\ $</td>
</tr>
<tr>
<td></td>
<td>min score</td>
<td>$\beta_c = \min_{g} M_{c,g} \ $</td>
<td>S</td>
<td>(Lahoti et al., 2020)</td>
</tr>
<tr>
<td></td>
<td>min ratio</td>
<td>$\beta_c = \min_{g} \frac{M_{c,g}}{\bar{M}_c} \ $</td>
<td>R</td>
<td>(Zafar et al., 2017)</td>
</tr>
<tr>
<td></td>
<td>max difference</td>
<td>$\beta_c = \max_{g} M_{c,g} - \min_{g} M_{c,g} \ $</td>
<td>S</td>
<td>(Bird et al., 2020)</td>
</tr>
<tr>
<td></td>
<td>max ratio</td>
<td>$\beta_c = \max_{g} \frac{M_{c,g}}{\max_{g} M_{c,g}} \ $</td>
<td>R</td>
<td>(Feldman et al., 2015)</td>
</tr>
<tr>
<td></td>
<td>difference threshold ($\gamma$)</td>
<td>$\beta_c = \frac{1}{G} \sum_{g} \gamma</td>
<td>M_{c,g} - \bar{M}_c</td>
<td>\ $</td>
</tr>
<tr>
<td></td>
<td>ratio threshold ($\gamma$)</td>
<td>$\beta_c = \frac{1}{G} \sum_{g} \gamma</td>
<td>M_{c,g} - \bar{M}_c</td>
<td>- 1 \ $</td>
</tr>
<tr>
<td>Class-wise</td>
<td>binary</td>
<td>$\delta = \sum_{c} \beta_c [1(c)] \ $</td>
<td>$\beta$</td>
<td>(Roh et al., 2021)</td>
</tr>
<tr>
<td></td>
<td>quadratic mean</td>
<td>$\delta = \sqrt{\frac{1}{C} \sum_{c} \beta_c^2} \ $</td>
<td>$\beta$</td>
<td>(Romanov et al., 2019)</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>$\delta = \frac{1}{C} \sum_{c} \beta_c \ $</td>
<td>$\beta$</td>
<td>(Li et al., 2018)</td>
</tr>
</tbody>
</table>

Table 1: Summary of different aggregation approaches. Based on the basic unit, group-wise aggregations are additionally categorized into three types: Score ($M_{c,g}$), Gap ($|M_{c,g} - \bar{M}_c|$), and Ratio ($\frac{M_{c,g}}{\bar{M}_c}$).

### 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 ($|M_{c,g} - \bar{M}_c|$), where smaller is better and the minimum is 0; and
3. the **ratio** of the evaluation results of a group to the average ($\frac{M_{c,g}}{\bar{M}_c}$), 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 $\gamma$ is fair w.r.t. $z$. Taking EO (Hardt et al., 2016) rather than fairness.

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

$$TPR_{c,g} = TPR_c, \forall g \in \{g\}^G,$$

As such, it is straightforward to directly measure the absolute difference between $TPR_{c,g}$ and $TPR_c$,

$$TPR_{c,g} = TPR_c \iff |TPR_{c,g} - TPR_c| = 0,$$

which is essentially the **gap** unit.

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

$$TPR_{c,g} = TPR_c \iff \frac{TPR_{c,g}}{TPR_c} = 1.$$

**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 $\epsilon \in \mathbb{R}^+$, $|1 - \frac{TPR_{c,g}}{TPR_c}| \leq \epsilon$. Similarly, the threshold can be applied to **gap**, resulting in $|TPR_{c,g} - TPR_c| \leq \epsilon$.

**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
The basic unit of group-wise aggregation, in-
cluding score, gap, and ratio, or other possi-
ble measures.

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

<table>
<thead>
<tr>
<th>Power ($p$)</th>
<th>Formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-\infty$</td>
<td>Minimum: $\min{v_1, \ldots, v_n}$</td>
</tr>
<tr>
<td>$-1$</td>
<td>Harmonic Mean: $\frac{n}{\sum_{i=1}^{n} v_i^{-1}}$</td>
</tr>
<tr>
<td>1</td>
<td>Arithmetic Mean: $\frac{1}{n} \sum_{i=1}^{n} v_i$</td>
</tr>
<tr>
<td>2</td>
<td>Quadratic Mean: $\left(\frac{1}{n} \sum_{i=1}^{n} v_i^2\right)^{1/2}$</td>
</tr>
<tr>
<td>$+\infty$</td>
<td>Maximum: $\max{v_1, \ldots, v_n}$</td>
</tr>
</tbody>
</table>

The additional advantage of using generalized
mean aggregation is that comparison across arbi-
trary $p$ values can be easily stated. For example,
group-wise aggregation is the $p = -5$ general-
ized mean with respect to the score units in a toxicity
classification competition, meaning that evalua-
tion focuses more on groups with lower perfor-
man.

Other Aggregation Methods: Although gener-
alized mean aggregation is a powerful tool for
describing and interpreting the aggregation pro-
cess, there are other ways that need further dis-
cussion. Previous work has also considered as-
signing different weights under aggregation, for in-
stance, Kearns et al. (2018) assign larger weights to
groups with larger populations. Such aggregations
can be implemented as the weighted generalized
mean: $M_{p,w}(v) = \left(\frac{1}{\sum_{i=1}^{n} w_i v_i^p}\right)^{1/p}$, 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 “advan-
taged” 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 $D$, 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, in-
cluding score, gap, and ratio, or other possi-
ble measures.
- The aggregation function for group-wise
aggregation, and the corresponding motivation.

Jigsaw Unintended Bias in Toxicity Classification: http
n/
The aggregation function for class-wise aggregation, and the corresponding motivation.

Although the particular choice of evaluation dataset \( D \) 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 \textbf{Score}, \textbf{Ratio}, and \textbf{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 \textbf{score} is the best practice, which maintains the original values under aggregation. On the other hand, if inter-group fairness is critical, \textbf{gap} and \textbf{ratio} are more appropriate choices. \textbf{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), \textbf{ratio} is a better choice than \textbf{gap}.

**Group-wise Aggregation Function Selection:** The selection of group-wise aggregation functions is shown as the exponent parameters of the \textit{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., \textbf{gap}, aggregation generally focuses on the maximum (Yang et al., 2020), i.e., \( p = +\infty \). For units like \textbf{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 \textbf{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 \textbf{gap} is essentially the standard deviation of \textbf{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.
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.\(^6\)

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

\(^6\)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-
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 Distance To the Optimal 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 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.

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

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

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.

Performance and fairness metrics have been introduced in Section 4.1.
Table 3: DTO scores of selected models over the BtBS 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DTO</th>
<th>P</th>
<th>P@F+5%</th>
<th>P@F+10%</th>
<th>F</th>
<th>F@P−5%</th>
<th>F@P−10%</th>
<th>AUC↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>INLP (Ravfogel et al., 2020)</td>
<td>41.9</td>
<td>41.9</td>
<td>52.6</td>
<td>52.6</td>
<td>70.2</td>
<td>41.9</td>
<td>41.9</td>
<td>39.8</td>
</tr>
<tr>
<td>Adv (Li et al., 2018)</td>
<td>39.0</td>
<td>44.6</td>
<td>43.3</td>
<td>41.8</td>
<td>49.4</td>
<td>41.2</td>
<td>41.2</td>
<td>43.6</td>
</tr>
<tr>
<td>DADV (Han et al., 2021)</td>
<td>37.9</td>
<td>44.7</td>
<td>41.0</td>
<td>40.5</td>
<td>39.9</td>
<td>40.4</td>
<td>41.9</td>
<td>44.5</td>
</tr>
<tr>
<td>AADV (Han et al., 2022b)</td>
<td><strong>36.9</strong></td>
<td>45.4</td>
<td><strong>39.5</strong></td>
<td><strong>39.0</strong></td>
<td>62.1</td>
<td>43.8</td>
<td>42.8</td>
<td>44.0</td>
</tr>
</tbody>
</table>

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**


Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. *Intrinsic bias metrics do not correlate...*


Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022b. Optimising equal opportunity fairness in model training. In Proceedings of


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 Carte-

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).
sian 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. 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. 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 \( G \times C^2 \) (one confusion matrix per group). Taking the B10S 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 \( 3 \times 10^3 \), \( 4 \times 10^4 \), and \( 1 \times 10^8 \), 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.

---

10There 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).

11Bios 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.

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.
<table>
<thead>
<tr>
<th>Selection</th>
<th>Method</th>
<th>Test Set</th>
<th>Development Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTO</td>
<td>INLP</td>
<td>81.4 ± 0.0</td>
<td>80.3 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>ADV</td>
<td>64.6 ± 4.5</td>
<td>63.9 ± 4.4</td>
</tr>
<tr>
<td></td>
<td>DADV</td>
<td>68.1 ± 5.5</td>
<td>67.6 ± 5.5</td>
</tr>
<tr>
<td></td>
<td>A-ADV</td>
<td>69.7 ± 4.9</td>
<td>69.3 ± 4.7</td>
</tr>
<tr>
<td>P</td>
<td>INLP</td>
<td>81.4 ± 0.0</td>
<td>80.3 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>ADV</td>
<td>81.5 ± 0.2</td>
<td>80.7 ± 0.2</td>
</tr>
<tr>
<td></td>
<td>DADV</td>
<td>81.4 ± 0.3</td>
<td>80.6 ± 0.3</td>
</tr>
<tr>
<td></td>
<td>A-ADV</td>
<td>81.3 ± 0.3</td>
<td>80.6 ± 0.3</td>
</tr>
<tr>
<td>P+F-5%</td>
<td>INLP</td>
<td>53.8 ± 0.0</td>
<td>53.8 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>ADV</td>
<td>71.5 ± 5.7</td>
<td>71.3 ± 5.8</td>
</tr>
<tr>
<td></td>
<td>DADV</td>
<td>73.4 ± 3.8</td>
<td>73.0 ± 3.7</td>
</tr>
<tr>
<td></td>
<td>A-ADV</td>
<td>70.4 ± 5.7</td>
<td>70.0 ± 5.8</td>
</tr>
<tr>
<td>P+F+10%</td>
<td>INLP</td>
<td>53.8 ± 0.0</td>
<td>53.8 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>ADV</td>
<td>68.8 ± 4.8</td>
<td>68.2 ± 4.8</td>
</tr>
<tr>
<td></td>
<td>DADV</td>
<td>69.7 ± 2.9</td>
<td>69.2 ± 2.7</td>
</tr>
<tr>
<td></td>
<td>A-ADV</td>
<td>70.0 ± 3.3</td>
<td>69.4 ± 3.4</td>
</tr>
<tr>
<td>F</td>
<td>INLP</td>
<td>29.8 ± 0.0</td>
<td>29.9 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>ADV</td>
<td>51.6 ± 16.5</td>
<td>51.2 ± 16.2</td>
</tr>
<tr>
<td></td>
<td>DADV</td>
<td>61.8 ± 3.7</td>
<td>61.2 ± 3.5</td>
</tr>
<tr>
<td></td>
<td>A-ADV</td>
<td>37.9 ± 9.1</td>
<td>37.6 ± 8.8</td>
</tr>
<tr>
<td>F+P-5%</td>
<td>INLP</td>
<td>81.4 ± 0.0</td>
<td>80.3 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>ADV</td>
<td>79.1 ± 1.1</td>
<td>78.5 ± 0.9</td>
</tr>
<tr>
<td></td>
<td>DADV</td>
<td>80.4 ± 0.5</td>
<td>79.9 ± 0.5</td>
</tr>
<tr>
<td></td>
<td>A-ADV</td>
<td>79.9 ± 2.2</td>
<td>79.1 ± 2.2</td>
</tr>
<tr>
<td>F+P-10%</td>
<td>INLP</td>
<td>81.4 ± 0.0</td>
<td>80.3 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>ADV</td>
<td>79.1 ± 1.1</td>
<td>78.5 ± 0.9</td>
</tr>
<tr>
<td></td>
<td>DADV</td>
<td>74.2 ± 3.2</td>
<td>73.7 ± 3.3</td>
</tr>
<tr>
<td></td>
<td>A-ADV</td>
<td>74.9 ± 5.1</td>
<td>74.3 ± 5.3</td>
</tr>
</tbody>
</table>

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 $\alpha$ 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 $\alpha'$ 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 $\int_0^{\alpha/\pi} DTO(\angle PUQ)d\angle PUQ$.

**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 $r'_1 > r'_2$,

$$|Q_1U + UO| > |Q_2U + UO|$$

$$|Q_1U + UO|^2 > |Q_2U + UO|^2$$

$$2r_1b\cos \angle PUQ_1 > 2r_2b\cos \angle PUQ_2$$

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.

$\blacksquare$