Finding Pareto trade-offs in fair and accurate detection of toxic speech

Soumyajit Gupta, Venelin Kovatchev, Anubrata Das, Maria De-Arteaga, and Matthew Lease

Abstract

Introduction. Optimizing NLP models for fairness poses many challenges. Lack of differentiable fairness measures prevents gradient-based loss training or requires surrogate losses that diverge from the true metric of interest. In addition, competing objectives (e.g., accuracy vs. fairness) often require making trade-offs based on stakeholder preferences, but stakeholders may not know their preferences before seeing system performance under different trade-off settings.

Method. We formulate the GAP loss, a differentiable version of a fairness measure, Accuracy Parity, to provide balanced accuracy across binary demographic groups.

Analysis. We show how model-agnostic, *HyperNetwork* optimization can efficiently train arbitrary NLP model architectures to learn *Pareto-optimal* trade-offs between competing metrics like predictive performance vs. group fairness.

Results. Focusing on the task of toxic language detection, we show the generality and efficacy of our proposed GAP loss function across two datasets, three neural architectures, and three fairness loss functions.

Conclusion. Our GAP loss for the task of TL detection demonstrates promising results - improved fairness and computational efficiency. Our work can be extended to other tasks, datasets, and neural models in any practical situation where ensuring equal accuracy across different demographic groups is a desired objective.

Introduction

Toxic language in social media is often associated with various risks and harms: cyber bullying, discrimination, mental health, and even hate crimes. Given the massive volume of user-generated content online, manual review of all posts by human moderators simply does not scale. Consequently, natural language processing (NLP) methods have been developed to fully or partially automate toxicity detection (Schmidt & Wiegand, 2017). Prior work has achieved high Accuracy and F1 scores on toxicity detection (e.g., (Zampieri et al., 2020)) across various model architectures: e.g., convolutional (CNN) (Gambäck & Sikdar, 2017), sequential (BiLSTM) (Graves et al., 2005), and transformer (BERT) (Devlin et al., 2018). However, studies have also found that model accuracy can vary greatly across sensitive demographic attributes, such as race or gender (Das et al., 2021; Park et al., 2018; Sap et al., 2019). Subjective annotation in such tasks arise from personal biases and experiences of annotators. Traditional approaches relying on majority voting to resolve disagreements leads to oversimplification of the task. For example, a BERT-based classifier obtains 90.4% vs. 84.5% accuracy for White vs. African American author on Davidson's dataset (Davidson et al., 2017) when just optimized for overall accuracy, independent of author groups. Thus, in subjective domains, the minority viewpoint plays an important role (Sang & Stanton, 2022), where context and interpretations around data collections (Rahman et al., 2022), sources (Chaudhry & Lease, 2022) and targets, can heavily influence judgments.

While recent years have seen rapid progress in fairness research, it is often measured in a post hoc manner, and optimization is often indirect (*e.g.*, by improving training data through pre- or post-processing) (Sap et al., 2019). A particular challenge is that most existing measures are non-differentiable and thus cannot be optimized directly via gradient descent. While one can optimize differentiable surrogate loss functions instead, this risks *metric divergence* between the optimization criteria used in training *vs.* the actual metrics of interest (Metzler & Croft, 2007; Morgan et al., 2004; Swezey et al., 2021; Yue et al., 2007).

As (Friedler et al., 2021) and others have noted, different worldviews lead to conflicting definitions of fairness that are mutually incompatible and specific fairness measures must be selected (suitable to the given task, context, and stakeholders at hand). In this work, we adopt a popular fairness objective, Accuracy Parity (Zhao et al., 2020) to optimize a model to provide balanced accuracy across demographic groups (Berk et al., 2021; Das et al., 2021; Heidari et al., 2019; Mitchell et al., 2021). Because no differentiable version of this measure exists, we formulate a novel, differentiable version, *Group Accuracy Parity* (GAP) that can be directly used to optimize descent-based models. We provide both a theoretical derivation and an empirical justification for GAP.

However, optimizing GAP alone may reduce Overall Accuracy (OA) since seeking to better fit minority group may lead to worse fit of majority group that tends to drive OA. Ultimately, we face a trade-off between competing objectives, whether we balance between competing accuracy goals (*e.g.*, precision *vs.* recall), fairness goals, or any combination thereof. Multi-Objective Optimization (MOO) provides a principled framework and rigorous toolbox for approaching such competing trade-offs, instead of treating them as single objective regularization problems (Little, 2023; Sorensen et al., 2024; Soto et al., 2022; Suau et al., 2024). We believe such MOO work remains underexplored in NLP today, and to the best of our knowledge, ours is the first NLP work on MOO for fair toxic language detection.

Because competing objectives typically lack global optima, optimization requires choosing among a set of equally valid, *Pareto-optimal* trade-offs between objectives. Naturally, selection of a suitable trade-off depends on stakeholder needs, and they typically wish to see system performance under real trade-off conditions before having to commit to any particular trade-off. We demonstrate how the full Pareto manifold – for *any* underlying model architecture – can be efficiently induced, provided optimization can be performed via gradient descent (with

differentiable loss objectives). This is accomplished via recent advances in *Pareto front learning* (PFL) (Gupta et al., 2022; X. Lin et al., 2020; Navon et al., 2021) for *HyperNetworks* (Ha et al., 2017), which train one neural model to generate effective weights for a second, target model.

In summary, we pursue two distinct and complementary approaches for fair toxic language detection via model optimization. First, recognizing the repeated call for balancing accuracy across demographic groups, yet finding no differentiable metric doing so, we present the first differentiable version, GAP, enabling optimization for the first time via standard gradient descent. Our results show a clear benefit of optimizing directly for the target metric of interest rather than surrogate loss functions that diverge from it. Second, to demonstrate generality of PFL optimization over competing objectives, we induce the full Pareto front of optimal trade-offs between OA *vs.* three different fairness measures: GAP and two prior measures. To show generality of both techniques – single-objective GAP and multiobjective PFL – we show optimization over three distinct neural architectures (CNN, BiLSTM, and BERT) on two datasets: Davidson (Davidson et al., 2017) and Wilds (Koh et al., 2021).

Our results show that GAP better balances accuracy across demographic groups (authors and targets of potentially toxic tweets) than existing differentiable measures. With multi-objective PFL, we show that we can successfully induce the full manifold of Pareto-optimal trade-offs across all differentiable objectives and neural architectures considered. GAP also achieves the best empirical trade-offs for OA vs. balanced accuracy in comparison to the two other fairness metrics considered. Finally, we note that GAP and PFL are broadly applicable and can be adapted for a wide range of NLP tasks, beyond the task of toxicity detection. For reproducibility and adoption, we provide our GAP source code.

Related work

Toxic language detection and fairness

Many datasets now exist to train and test automated systems for TL detection (Poletto et al., 2021; Vidgen & Derczynski, 2020). Many NLP models have been proposed and continue to increase overall accuracy of detection (Fortuna & Nunes, 2018; MacAvaney et al., 2019; Schmidt & Wiegand, 2017). However, recent studies highlight the racial bias induced in such classification tasks. Davidson *et al.* (Davidson et al., 2017) introduced a dataset with a corpus of tweets collected from social media and human annotations on the toxicity of the tweet. Sap *et al.* (Sap et al., 2019) and Davidson *et al.* (Davidson et al., 2019) analyse the correlation between race and gold-label of toxicity in the (Davidson et al., 2017) dataset and find a strong association between AAE markers and toxicity annotation, where both works noisily infer author dialect via Blodgett *et al.* (Blodgett et al., 2017)'s model as a proxy for race. The Wilds (Koh et al., 2021) dataset contains targets of TL with different demographic information and human annotated majority voted labels. It provides predefined training/test splits for effectively measuring distribution shifts in TL models.

To address the problem of bias in automatic TL detection, some work has been done on improving the training and testing data (Park et al., 2018; Röttger et al., 2021; Sap et al., 2019), with the expectation that fairer data will lead to fairer learned models. Some of the most similar to our work, by Xia *et al.* (Xia et al., 2020), Burack *et al.* (Ball-Burack et al., 2021), and Shen *et al.* (Shen et al., 2022), seeks to reduce the bias towards AAE-authors in the algorithm rather than data.

Fairness measures

The amplification of systemic unfairness through AI applications has been pronounced across different critical application areas such as hiring, finance, legal applications, content moderation *etc.* (Angwin et al., 2016; Balashankar & Lees, 2022). It is of societal and ethical importance to examine if an AI is discriminative and develop methods to make the AI fair on grounds such as gender, ethnicity, or other forms of identity attributes (Ekstrand et al., 2022). To connect fairness concepts with statistical measures in machine learning, Mitchell *et al.* (Mitchell et al., 2021)

synthesizes fairness measures based on the confusion matrix. Friedler *et al.* (Friedler et al., 2019) further categorize fairness measures into largely three categories: 1) measures based on base rates, such as Disparate Impact (Feldman et al., 2015), 2) measures based on group-conditioned accuracy, and 3) measures based on group-conditioned calibration.

Pareto optimization of trade-offs

Multi-Objective Optimization (MOO) is increasingly pursued in fair classification (Caton & Haas, 2020). The complexity of real-world problem often leads to competing objectives such as accuracy vs. fairness. Pareto frameworks are powerful tools to balance between such competing objectives. Several works (Balashankar et al., 2019; Martinez et al., 2020) seek to balance accuracy vs. fairness. Valdivia et al. (Valdivia et al., 2020) presents a group-fairness based trade-off model for decision tree classifiers via a genetic algorithm. Wei et al. (Wei & Niethammer, 2020) uses Chebyshev scalarization to provide a neural architecture for fairness vs. accuracy Pareto front computation in classification. Lin et al. (X. Lin et al., 2019) claims Pareto optimality on the basis of KKT conditions. In this work, we adopt Gupta et al. (Gupta et al., 2022)'s SUHNPF framework, given its error tolerance bounds and strong empirical performance. We apply it as a HyperNetwork (Ha et al., 2017) to optimize a variety of neural network models for TL detection. While we only optimize the Pareto trade-off between a single accuracy measure vs. a single fairness measure, the framework itself is more general and directly supports optimizing arbitrary numbers of competing objectives (and constraints).

Group accuracy parity (GAP)

In this work, we focus on *accuracy parity* (AP) (Zhao et al., 2020), *i.e.,* balancing accuracy across groups (subpopulations based on some demographic criteria), sometimes known as *equal accuracy* (Mitchell et al., 2021), *equality of accuracy* (Heidari et al., 2019), *overall accuracy equality* (Berk et al., 2021), *accuracy equity* (Dieterich et al., 2016), or *accuracy difference* (Das et al., 2021). We do not claim any primacy of this particular notion of fairness, but show that if one is interested in it, it can be directly optimized via our Group Accuracy Parity (GAP) measure without *metric divergence* (Metzler & Croft, 2007; Morgan et al., 2004; Swezey et al., 2021; Yue et al., 2007) between loss function *vs.* evaluation metric.

Accuracy difference

While AP is an equality condition, we still need to quantify the deviation from equality in cases of unequal performance across groups. We therefore use Accuracy difference (AD) (Das et al., 2021), a continuous version of AP to measure this deviation. AD is shown in (Eq. 1), where ŷ,y,g are the predicted label, true label, and group attribute respectively.

$$AD = \underbrace{P[\hat{y} = y | g = 1]}_{\text{Acc Group 1 (g=1)}} - \underbrace{P[\hat{y} = y | g = 0]}_{\text{Acc Group 0 (g=0)}} \qquad (1)$$

AD being defined based on the confusion matrix, makes the formulation is probabilistic in nature, *i.e.,* ratio of numbers over the dataset, and not distribution over variable, AD becomes non-differentiable. Thus, AD can only be used in a post-hoc manner and cannot be directly used for gradient-based back propagation. Furthermore, Eq. 1 inherently assumes that the majority group accuracy ($g = 1$) will always be higher than the vulnerable group ($g = 0$), which might not always hold true, resulting in potential negative values of AD in the range [-1,1]. Naturally, as a post-hoc measure, AD is disconnected from the optimization objective of the model used during training. These limitations motivated us to define a differentiable, non-probabilistic form of AD we refer to as Group Accuracy Parity (GAP), which allows any descent-based model during training to optimize close to equal accuracy across sensitive attribute classes, and addresses the range issue of AD.

Formulation

Binary Cross Entropy (BCE), as formulated in Eq. 2 is typically used as a loss function for optimizing a classifier. Although not a strict one-to-one correspondence, it is observed that minimizing BCE leads to maximization of Accuracy.

$$BCE = -\frac{1}{N} \sum_N y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \tag{2}$$

Weighted Cross Entropy (WCE) is a variant of BCE that re-weights the error for the different classes proportional to their inverse frequency of labels in the data. The class re-weighting strategy is available in packages like SkLearn (Pedregosa et al., 2011) and is discussed in detail by Lin *et al.* (T.-Y. Lin et al., 2017). For balanced classification across sensitive attributes (*e.g.*, demographic information across author groups or gender information across targets in Hate Speech), we formulate our GAP loss function as follows: we first calculate the WCE for each sensitive attribute (*g*), then minimize the difference across them. The GAP loss function in Eq. 3 is minimized only when WCE errors match across the binary sensitive attribute.

$$GAP = \| \underbrace{WCE(g = 1)}_{\text{Acc Group 1 (g=1)}} - \underbrace{WCE(g = 0)}_{\text{Acc Group 0 (g=0)}} \|_2^2 \tag{3}$$

The GAP function has the following properties:

1. GAP maps to AD. GAP has a one-to-one correspondence to AD *i.e.*, minimizing GAP also minimizes AD.
2. GAP is differentiable. GAP is defined as the squared 2-norm difference between the Weighted Cross Entropy (WCE) across the two sensitive attribute. Since WCE is differentiable, so is the 2-norm difference. Hence GAP can optimize any descent-based model.
3. GAP is symmetric. GAP has a 2-norm formulation, ensuring the range of attainable values are within $GAP \in [0,1]$, avoiding the negativity issue faced in AD. Also being a 2-norm measure, the loss surface of GAP is smoother than other comparable measures like CLA (Shen et al., 2022), which uses 1-norm (Boyd et al., 2004).

For a step-by-step derivation from WCE to GAP, readers are referred to Appendix A, showing the strict correspondence between the loss measures. In this paper we implement GAP (Eq. 3) to correspond to AD (Eq. 1). As such, GAP can be optimized over binary labels and binary groups.

Optimizing competing objectives

Typically, toxicity detection systems are trained with the single objective of maximizing OA (Founta et al., 2018; Park et al., 2018; Röttger et al., 2021) or a custom defined objective (Xia et al., 2020). In contrast, we frame toxicity detection as a Multi-Objective Optimization (MOO) problem. It is important to highlight the distinction between an M(Multi)OO vs. S(Single)OO formulation and their interpretation. Consider the two objectives as f1: Cross-Entropy and f2: Fairness. Traditional fair classifiers operate by adding a penalty term corresponding to Fairness to the main objective Entropy with a hyper-parameter λ in Eq. 4.

$$\min \quad f_1 + \lambda f_2$$
$$\min \quad \text{Cross-Entropy loss} + \lambda \, \text{Fairness loss} \tag{4}$$

The reader is specifically requested to note that such optimization process does not have any control over the range of λ, and it can vary generally between (0,∞). During the optimization process, we tune λ till we get a desired performance in SOO setting. Furthermore, there is no explicit requirement of the scale of f_1 and f_2 to be the same. Thus, there is no simple correlation between the amount of Fairness we want *vs.* the value of λ.

An unconstrained MOO problem with two competing loss objectives is defined in Eq. 5. Note that this is a joint min-min problem instead of a single min problem. The objectives here need to be at the same scale $w.r.t.$ each other. If the expectation is to achieve a liner trade-off between them, the linear scalarized form of the MOO problem with trade-off $\alpha \in [0,1]$, minimizes both objectives simultaneously in Eq. 6. Solving this reformulated MOO problem would achieve balance between Entropy and Fairness, with α holding strict mathematical interpretation of linear trade-off. Decreasing Entropy causes Fairness to increase, while decreasing Fairness causes Entropy to increase.

$$\min \min \quad f_1 , f_2 \tag{5}$$

$$\min \quad \alpha f_1 + (1 - \alpha) f_2$$

$$\min \quad \alpha \, \text{Cross-Entropy loss} + (1 - \alpha) \, \text{Fairness loss} \tag{6}$$

Note that there are multiple mathematically optimal solutions to Eq. 6. Every optimal solution corresponding to each value of α in Eq. 6 is a member of the Pareto optimal solution set $i.e.$, the Pareto front contains the set of optimal model parameters given the dataset and the model. To solve this MOO problem, we adopt the SUHNPF Pareto framework (Gupta et al., 2022) as a HyperNetwork (Ha et al., 2017) to learn optimal TL detection neural model parameters over trade-offs. Hypernetworks train one neural model to generate effective weights for a second, target model.

SUHNPF efficiently learns the entire Pareto manifold of feasible trade-off values during training. This empowers users to then choose any solution point they prefer on the manifold, $a \ posteriori$, and extract the classifier weights configuration as per their desired trade-off α, without retraining the model for that α. Training the same model for K different α's, with R being the time for a single run, would result in total runtime of K × R $i.e.$, linear on the number of runs. Using the Hypernetwork to learn the manifold is computationally much more efficient $i.e.$, taking a constant time c × R, 1 < c ≪ K over feasible α's, rather than for each value of α. Refer to Appendix D for values on runs.

Experimental details

In this section, we describe our datasets, neural models, baseline losses and other evaluation details.

Datasets

We consider two datasets: Davidson et al. (Davidson et al., 2017) for author demographics and the Civil Comments (Borkan et al., 2019) portion of Wilds (Koh et al., 2021) for target demographics (Table 1). In each case, we frame the task as a binary classification problem (Toxic vs. non-Toxic, or "safe") with binary sensitive attributes (Majority vs. Minority, the under-represented, sensitive attribute). Note that "Majority" and "Minority" in our work simply refers to the statistical representation of the group in the data and does not carry any social or cultural meaning.

Dataset	Group	Toxic	Safe	Total
Davidson	Minority	8,725	302	9,027 (36%)
	Majority	11,895	3,861	15,756 (64%)
Wilds	Minority	5,973	33,762	39,735 (44%)
	Majority	6,832	42,950	49,782 (56%)

Table 1. Statistics of the two datasets used in this work. For Davidson et al. (Davidson et al., 2017), we consider the author demographics AAE vs. SAE as group attribute for minority vs. majority group. For Wilds (Koh et al., 2021), we consider the binary group target gender as male vs. female for minority vs. majority group attributes.

Author Demographics Dataset We consider fair moderation of posts written by authors from different demographic groups in (Davidson et al., 2017). Prior studies (Arango et al., 2019; Sap et al., 2019) have empirically demonstrated the existence of bias towards author demographics in toxic language classification. The sensitive attribute in this dataset is race, as identified by the dialect of the tweets. Following prior work, we apply Blodgett et al. (Blodgett et al., 2017)'s model to automatically-detect dialect labels for each of the tweet as African American English (AAE) or Standard American English (SAE), representing Minority and Majority groups, respectively. We acknowledge both that dialect is only a weak surrogate representation of demographic race, and that automatic detection of dialect will naturally incur noise. However, in this, we follow established practices from prior work. Our fairness methods are agnostic to the sensitive attribute labelled in the data, and our results are only intended to attest to the capabilities of our proposed methods, rather than provide findings regarding protection of any specific vulnerable population. Davidson et al. (Davidson et al., 2017)'s data includes 24,783 Twitter posts labelled as Hate, Offensive, or Normal. Following prior work (Park et al., 2018), we set the class label to 1 (Toxic) if the post contains hate speech or offensive language, and 0 otherwise. We note that tweets from Minority authors are annotated as toxic in 96% of the cases, compared to 75% for the tweets by Majority authors. While these statistics suggest an important risk of annotation bias in this dataset, dataset debiasing lies beyond the scope of our work. Our focus in this work is restricted to balancing accuracy across the groups, given the dataset as it is annotated.

Target Identity Dataset To assess fair protection of different groups targeted in posts, we use the Civil Comments (Borkan et al., 2019) portion of Wilds (Koh et al., 2021). This dataset has 448,000 training tweets labelled as Toxic or non-Toxic. Each tweet has explicit annotation for the demographics, gender, or religion of the target entity. We select tweets where more than 50% of annotators agreed on the gender of the target. In this work, we include only female (majority) and male (minority) genders in order construct a binary sensitive attribute for our experiments. In doing so, we fully-acknowledge both the non-binary nature of gender and individual freedom of self-identification. As noted above, our methods are agnostic as to the sensitive attribute labelled in the data, and our inclusion of only two genders merely reflects a convenient way to assess the capabilities of our proposed methods in regards to balancing accuracy across a binary sensitive attribute.

Neural models considered

To assess the generality of our methods across distinct neural architectures, we evaluate over three types of models: CNN (Gambäck & Sikdar, 2017), BiLSTM (Graves et al., 2005) and BERT (Devlin et al., 2018). For full experimental setup, please refer to Appendix C. For all three models, we freeze the feature representation layers and optimize the weights of the classification layer. In general, GAP loss optimization and the SUHNPF hypernetwork (Gupta et al., 2022) support such generalization across any models that can be trained via gradient descent.

Baseline loss functions

We compare against two baseline loss functions. The first fairness loss CLAss-wise equal opportunity (CLA) (Shen et al., 2022) seeks to balance False Negative Rate (FNR) across protected groups (Chouldechova, 2017), also known as equality of opportunity (Hardt et al., 2016). CLA minimizes the error in absolute differences between error *w.r.t.* a label ($BCE(y)$) and error *w.r.t.* a label given the sensitive attribute ($BCE(y, g)$), with hyperparameter $\lambda \in [0, \infty]$, which differs from minimizing AD. Due to the 1 norm nature of CLA, the optimization surface for the loss function is not smooth (Boyd et al., 2004).

$$CLA = BCE + \lambda \cdot \sum_{y \in C} \sum_{g \in G} |BCE(y, g) - BCE(y)| \qquad (7)$$

The second fairness loss (Xia et al., 2020) is an adversarial approach to demoting unfairness, which we denote as ADV. It seeks to provide false positive rate (FPR) balance (Chouldechova, 2017) across groups, otherwise known as *predictive equality*. Being adversarial in nature, this method, and others (Chen et al., 2024) does not have any correspondence to any evaluation measure. Thus, users should take caution of possible metric divergence while using such techniques, with tuner $\beta \in [0,1]$.

$$ADV = \beta \cdot BCE + (1 - \beta) \cdot (adversary(y, g) - 0.5) \tag{8}$$

However, while ADV is motivated by FPR balance, no equivalence between the loss function and the evaluation metric is shown, exemplifying *metric divergence* between loss function and evaluation goal. Their reported results also show only limited empirical correspondence between reducing the model loss and reducing FPR.

Experimental setup

We have two experimental setups with the weighted cross entropy (WCE) as f_1 and the Fairness criteria as f_2. First, we optimize the fairness measure directly as a SOO problem following Eq. 4 under a penalization setting, as proposed in CLA (Shen et al., 2022). Secondly, we use the MOO setting to find the best trade-offs between WCE and fairness measure following Eq. 6, with the SOO *vs*. MOO distinction described in Sec 4.

Evaluation measures

Our focus in this work is the tension between minimizing *accuracy difference* (AD) (Das et al., 2021) and maximizing overall accuracy (OA). We thus evaluate on four post-hoc measures: OA over the dataset (majority and minority groups together), accuracy of each group separately, and AD observed between groups. Although we do not directly optimize F1, since a differentiable version of F1 does not exist, we still report the values in Appendix E.

Results

Existing bias in CNN, BiLSTM, BERT

Table 2 presents results for three toxic language classifiers optimized to maximize OA (*i.e.*, WCE) on Davidson *et al.* (Davidson et al., 2017)'s dataset. The *Majority* class consistently shows 6-7% higher accuracy than the *Minority* class, across models and five random initialisations. Such imbalance serves as motivation for our work to optimize OA/AD across demographic groups. This inequality behavior in toxic language detection is consistent across all three neural models and both datasets. Due to space restrictions in the main body, we present the results only for the BERT-based classifier. However, our findings also apply to BiLSTM and CNN networks, whose results are available in Appendix F.

Models	Overall %	Majority %	Minority %	AD %
CNN	87.52 ± 0.3	89.12 ± 0.2	82.88 ± 0.3	6.24 ± 0.2
BiLSTM	87.60 ± 0.2	89.37 ± 0.2	82.46 ± 0.1	6.91 ± 0.3
BERT	88.84 ± 0.2	90.35 ± 0.2	84.47 ± 0.1	5.88 ± 0.1

Table 2. Baseline accuracy results on Davidson *et al.* (Davidson et al., 2017)'s dataset when maximizing overall accuracy (OA) only. Results show consistent bias of higher accuracy for the majority.

Single objective optimization (SOO)

Table 3 shows the results for the SOO experimental setup. The baseline BERT model optimized via Cross Entropy obtains 88.84% OA and 5.88% AD on Davidson *et al.* (Davidson et al., 2017) and 84.68% OA and 3.88% AD on Wilds (Koh et al., 2021). All three loss functions successfully reduce

the AD on both datasets. As expected, the improvement in fairness comes at the cost of lower OA. We evaluate the different optimization metrics by looking at both the change in AD and in OA.

ADV performs the worst of the three measures, most notably due to its relatively large drop in OA. Optimizing for GAP and CLA gives the same OA, where the two losses show no significant difference across 5 initializations. However, in terms of reducing AD, our GAP measure outperforms CLA by 0.9% on Davidson and 1.5% on Wilds. Looking at the results, we can conclude that GAP is the best performing measure in terms of reducing Accuracy Difference. The results are consistent across both datasets. These results show the value in optimizing a measure that correctly reflects the desired notion of fairness, as well as the benefit from directly optimizing the measure of interest, rather than surrogate or approximate loss functions, to avoid metric divergence.

Measure	Overall %	Majority %	Minority %	AD %
	Davidson			
Baseline	88.84 ± 0.2	90.35 ± 0.2	84.47 ± 0.1	5.88 ± 0.1
GAP (Ours)	87.32 ± 0.1	87.35 ± 0.1	87.26 ± 0.1	$\mathbf{0.09 \pm 0.0}$
CLA	87.57 ± 0.2	87.82 ± 0.1	86.87 ± 0.1	0.95 ± 0.0
ADV	86.27 ± 0.4	86.88 ± 0.2	84.52 ± 0.3	2.36 ± 0.1
	Wilds			
Baseline	84.68 ± 0.3	86.41 ± 0.2	82.49 ± 0.1	3.88 ± 0.2
GAP (Ours)	84.38 ± 0.1	84.51 ± 0.1	84.23 ± 0.0	$\mathbf{0.28 \pm 0.0}$
CLA	84.43 ± 0.1	85.23 ± 0.1	83.41 ± 0.0	1.82 ± 0.1
ADV	83.61 ± 0.2	84.17 ± 0.1	82.91 ± 0.1	1.26 ± 0.1

Table 3. Optimizing fairness in a SOO setup. We compare a BERT-based model trained using cross entropy (Baseline) with models trained using different fairness measures. Our proposed measure (GAP) obtains the best results in reducing AD while maintaining high overall accuracy.
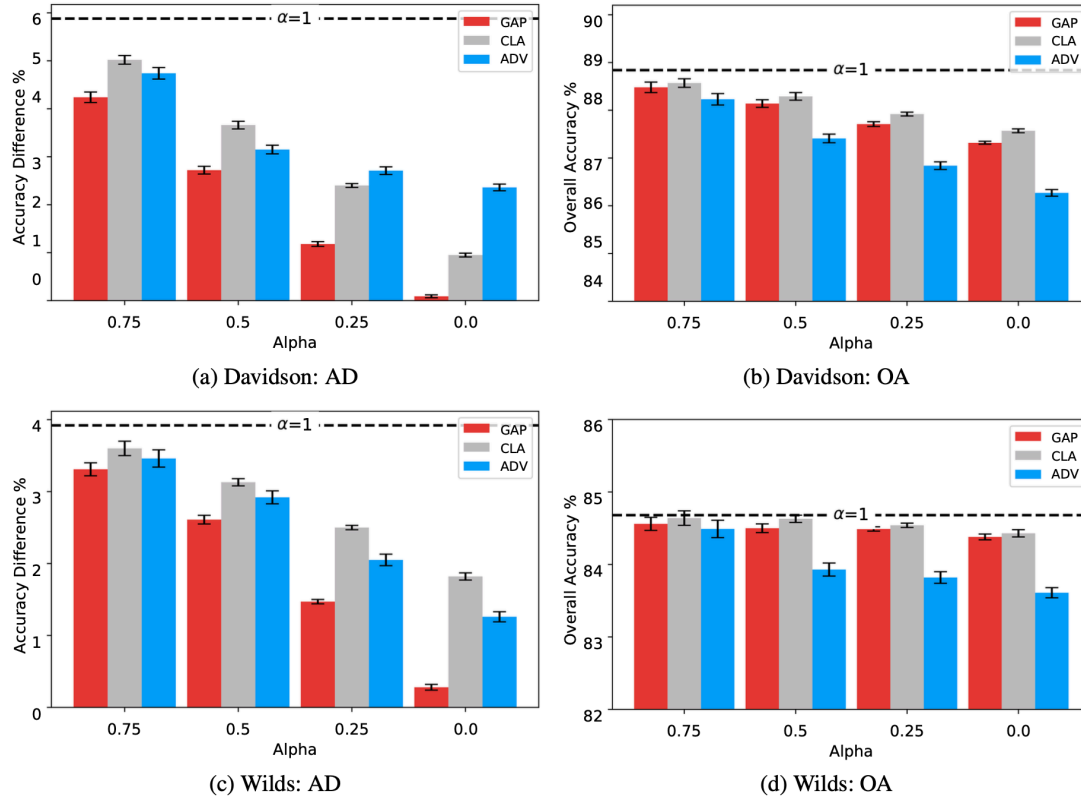
Figure 1. Trade-offs between accuracy difference (AD) and overall accuracy (OA), on the BERT based model with SUHNPF acting as hypernetwork for three methods — GAP (ours), CLA, and ADV – across the two datasets for $\alpha \in [0,1]$, with $\alpha = 0$ optimizing AD only and $\alpha = 1$ optimizing OA only. GAP achieves lower AD consistently across α settings and datasets, while a more modest drop in OA is observed across methods as AD is reduced.

Multi objective optimization (MOO)

In Section 6.2 we used GAP, CLA, or ADV to directly optimize fairness. However, the reduced AD comes at the cost of lower OA. In order to find the optimal trade-offs between fairness and accuracy, we use the SUHNPF framework in a MOO experimental setup. We use a BERT-based classifier and three different pairs of objective functions: WCE *vs.* GAP; WCE *vs.* CLA; and WCE *vs.* ADV, learning a linear MOO trade-off between the two competing objectives.

Fig. 1 shows the results of the MOO experiments. SUHNPF allows us to control how important is each objective (accuracy *vs.* fairness) by choosing the value of α. At α**=1**, we optimize only for Accuracy, and at α**=0**, only for fairness. We illustrate the different trade-offs at 4 points of the Pareto front (α = 0, 0.25, 0.5, and 0.75). We can observe that with decreasing α, both AD and OA decrease. For ADV we can see that the drop in AD is comparable to the drop in OA, which is not an efficient trade-off between accuracy *vs.* fairness. GAP and CLA maintain a relatively consistent OA, while GAP reduces AD far more than CLA, yielding the best trade-off for each α. See Appendix E for discussion on metric divergence and tabulated values in experiments. We can conclude that GAP is consistently the best metric, across SOO and MOO experimental setups and across different values of α for MOO.

Conclusion

Optimizing fairness: since fairness measures embody different underlying assumptions and statistical choices, selecting an appropriate fairness metric often depends on the task, use case, and stakeholder priorities. In this work, we focus on a popular fairness objective of balancing accuracy across different demographic groups, also known as minimizing Accuracy Difference. We

show that our *Group Accuracy Parity* (GAP) measure directly optimizes AD without *metric divergence* between loss function *vs.* evaluation metric. Results show GAP consistently achieves lower AD than prior work with modest loss in OA across datasets.

MOO and toxic language detection: rather than force the users to settle for any single accuracy or fairness measure, we further adopt SUHNPF, a multi-objective optimization (MOO) framing for joint pursuit of multiple objectives. We learn the full Pareto manifold over competing objectives so that users can view the full space of feasible trade-offs and choose any desired trade-off on the solution manifold, *a posteriori*. We empirically demonstrate that our measure GAP performs better than alternative differentiable fairness objectives in reducing AD. To the best of our knowledge this is the first use of MOO for fair toxic language detection.

Fairness and toxic language detection: we explore two different aspects of fairness in toxic language detection: 1) fair moderation of posts written by authors from different demographic groups; and 2) fair protection of different groups targeted by posts. We successfully improved the fairness of the models in both experimental setups, demonstrating the generality of the proposed approach.

Extending GAP to multiple classes and demographic groups: we formulate GAP following the strict definition of AD, which is for two classes and two demographic groups. Fairness literature has discussed heuristics and formulations for extending AD to multi-group and multi-class classification and balancing between multiple groups. As a future work, GAP can be extended based on those hypotheses.

Group identification: with author demographics in Davidson *et al.* (Davidson et al., 2017)'s dataset, we rely on automatic detection of author dialect, which is noisy. With target group demographics in Wilds (Koh et al., 2021), we assume oracle knowledge of target groups from annotation, which would have to be noisily detected in practice. In both cases, therefore, we make simplifying assumptions in this work. Optimizing trade-offs with awareness of noise in detection of demographic groups thus remains another direction for future work.

Dataset debiasing: recent studies highlight the risks of annotation bias, be it by annotator guidelines or the annotators themselves. Sap *et al.* (Sap et al., 2019) and Davidson *et al.* (Davidson et al., 2019) analyse the correlation between race and gold-label of toxicity in several datasets and find a strong association between African American English (AAE) markers and toxicity annotation. Because our work is restricted to balancing accuracy across the sensitive attribute, given the dataset as it is annotated, our results our limited by any such bias present in the data (Ludwig et al., 2024), Addressing such annotation bias thus remains another key direction for future work.

Generality and scope of this work: we implement GAP and SUHNPF for the task of TL detection and demonstrate promising results - improved fairness and computational efficiency. However, our work can be extended to other tasks, datasets, and neural models in any practical situation where ensuring equal accuracy across different demographic groups is a desired objective. Recently, Kovatchev and Lease (Kovatchev & Lease, 2024) demonstrated the significant impact of imbalanced data in popular NLP benchmarks. Our work can help address that challenge.

Acknowledgments

About the authors

Soumyajit Gupta is a graduate student in Computer Science, University of Texas at Austin, USA. He received his Ph.D. from UT Austin and his research interests are in Machine Learning and Interpretable Neural Design. He can be contacted at smjtgupta@utexas.edu

Venelin Kovatchev is an Associate Professor in Computer Science, University of Birmingham, UK. He did his Post Doctorate from UT Austin and his research interests are in Computational Linguistics and Natural Language Processing. He can be contacted at v.o.kovatchev@bham.ac.uk

Anubrata Das is a graduate student in School of Information, University of Texas at Austin, USA. He is under PhD candidacy at UT Austin and his research interests are in Human Computer Interactions and Natural Language Processing. He can be contacted at anubrata.das@utexas.edu

Maria De-Arteaga is an Assistant Professor in McCombs School of Business, University of Texas at Austin, USA. She is a Good Systems researcher. Her research interests are in Algorithmic Fairness and Human-AI complementarity. She can be contacted at dearteaga@mccombs.utexas.edu

Matthew Lease is a Professor in School of Information, University of Texas at Austin, USA. He is a Good Systems researcher and co-director of UT's CosmicAI Project. His research interests are in Artificial Intelligence and Human-Computer Interaction. He can be contacted at ml@utexas.edu

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 45–54.

Balashankar, A., & Lees, A. (2022). The need for transparent demographic group trade-offs in credit risk and income classification. *International Conference on Information*, 344–354.

Balashankar, A., Lees, A., Welty, C., & Subramanian, L. (2019). What is fair? exploring pareto-efficiency for fairness constrained classifiers. arXiv preprint arXiv:1910.14120.

Ball-Burack, A., Lee, M. S. A., Cobbe, J., & Singh, J. (2021). Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 116–128. https://doi.org/10.1145/3442188.3445875

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arxiv preprint. arXiv preprint arXiv:1810.01943.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 50(1), 3–44.

Blodgett, S. L., Wei, J., & O'Connor, B. (2017). A dataset and classifier for recognizing social media English. Proceedings of the 3rd Workshop on Noisy User-generated Text, 56–61.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. Companion proceedings of the 2019 world wide web conference, 491–500.

Boyd, S., Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge university press.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053.

Chaudhry, P., & Lease, M. (2022). You are what you tweet: Profiling users by past tweets to improve hate speech detection. International Conference on Information, 195–203.

Chen, T., Wang, D., Liang, X., Risius, M., Demartini, G., & Yin, H. (2024). Hate speech detection with generalizable target-aware fairness. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 365–375.

Chollet, F. (2015). Keras.

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

Das, S., Donini, M., Gelman, J., Haas, K., Hardt, M., Katzman, J., Kenthapadi, K., Larroy, P., Yilmaz, P., & Zafar, M. B. (2021). Fairness measures for machine learning in finance. The Journal of Financial Data Science, 3(4), 33–64.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc, 7(4).

Ekstrand, M. D., Dass, A., Burke, R., Diaz, F., et al. (2022). Fairness in information access systems. Foundations and Trends® in Information Retrieval, 16(1-2), 1–177.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 259–268.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1–30.

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. https://open.bu.edu/handle/2144/40119

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. Communications of the ACM, 64(4), 136– 143.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the conference on fairness, accountability, and transparency, 329–338.

Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. Proceedings of the first workshop on abusive language online, 85–90.

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional lstm networks for improved phoneme classification and recognition. International conference on artificial neural networks, 799–804.

Gupta, S., Singh, G., Bollapragada, R., & Lease, M. (2022). Learning a Neural Pareto Manifold Extractor with Constraints. Proceedings of the 38th International Conference on Uncertainty in Artificial Intelligence (UAI).

Ha, D., Dai, A. M., & Le, Q. V. (2017). Hypernetworks. 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. https://openreview.net/forum? id=rkpACe1lx

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2019). A moral framework for understanding fair ml through economic models of equality of opportunity. Proceedings of the conference on fairness, accountability, and transparency, 181–190.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. International Conference on Machine Learning, 5637–5664.

Kovatchev, V., & Lease, M. (2024, June). Benchmark transparency: Measuring the impact of data on evaluation. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers) (pp. 1536– 1551). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.86

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2980–2988.

Lin, X., Yang, Z., Zhang, Q., & Kwong, S. (2020). Controllable pareto multi-task learning. arXiv preprint arXiv:2010.06313.

Lin, X., Chen, H., Pei, C., Sun, F., Xiao, X., Sun, H., Zhang, Y., Ou, W., & Jiang, P. (2019). A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. Proceedings of the 13th ACM Conference on Recommender Systems, 20–28.

Little, C. (2023). To the fairness frontier and beyond: Identifying, quantifying, and optimizing the fairness-accuracy pareto frontier [Master's thesis, Rice University].

Ludwig, F., Dolos, K., Alves-Pinto, A., & Zesch, T. (2024). Unraveling the dynamics of semi-supervised hate speech detection: The impact of unlabeled data characteristics and pseudo-labeling strategies. Findings of the Association for Computational Linguistics: EACL 2024, 1974–1986.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8).

Martinez, N., Bertran, M., & Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. Proceedings of the 37th International Conference on Machine Learning.

Metzler, D., & Croft, W. B. (2007). Linear feature-based models for information retrieval. Information Retrieval, 10(3), 257–274.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application, 8, 141–163.

Morgan, W., Greiff, W., & Henderson, J. (2004). Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. Proceedings of HLT-NAACL 2004: Short Papers, 93–96.

Narayanan, A. (2018). 21 fairness definitions and their politics: A tutorial [https://shubhamjain0594.github.io/post/ tlds-arvind-fairness-definitions/]. Proceedings of the ACM FAccT Conference on Fairness, Accountability and Transparency.

Navon, A., Shamsian, A., Fetaya, E., & Chechik, G. (2021). Learning the pareto front with hypernetworks. International Conference on Learning Representations.

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2799–2804. https://doi.org/10.18653/ v1/D18-1302

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. Language Resources and Evaluation, 55(2), 477–523.

Rahman, M. M., Kutlu, M., & Lease, M. (2022). Understanding and predicting characteristics of test collections in information retrieval. International Conference on Information, 136–148.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. 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), 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

Sang, Y., & Stanton, J. (2022). The origin and value of disagreement among data labellers: A case study of individual differences in hate speech annotation. International Conference on Information, 425–444.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. Proceedings of the 57th annual meeting of the association for computational linguistics, 1668–1678.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 1–10. https://doi.org/ 10.18653/v1/W17-1101

Shen, A., Han, X., Cohn, T., Baldwin, T., & Frermann, L. (2022). Optimising equal opportunity fairness in model training. Proceedings of the 2022 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies, 4073–4084. https://doi.org/10.18653/v1/2022.naacl-main.299

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. (2024). A roadmap to pluralistic alignment. arXiv preprint arXiv:2402.05070.

Soto, C. P., Nunes, G. M., Gomes, J. G. R., & Nedjah, N. (2022). Application-specific word embeddings for hate and offensive language detection. Multimedia Tools and Applications, 81(19), 27111–27136.

Suau, X., Delobelle, P., Metcalf, K., Joulin, A., Apostoloff, N., Zappella, L., & Rodríguez, P. (2024). Whispering experts: Neural interventions for toxicity mitigation in language models. arXiv preprint arXiv:2407.12824.

Swezey, R., Grover, A., Charron, B., & Ermon, S. (2021). Pirank: Scalable learning to rank via differentiable sorting. Advances in Neural Information Processing Systems, 34, 21644–21654.

Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2020). How fair can we go in machine learning? assessing the boundaries of fairness in decision trees. arXiv preprint arXiv:2006.12399.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. Plos one, 15(12), e0243300.

Wei, S., & Niethammer, M. (2020). The fairness-accuracy pareto front. arXiv preprint arXiv:2008.10797.

Xia, M., Field, A., & Tsvetkov, Y. (2020). Demoting racial bias in hate speech detection. Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, 7–14. https://doi.org/10.18653/ v1/2020.socialnlp-1.2

Yue, Y., Finley, T., Radlinski, F., & Joachims, T. (2007). A support vector method for optimizing average precision. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 271–278.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). Proceedings of SemEval.

Zhao, H., Coston, A., Adel, T., & Gordon, G. J. (2020). Conditional learning of fair representations. 8th International Conference on Learning Representations, ICLR 2020.

Appendix

Relating BCE and GAP measures

We provide a step-by-step derivation from BCE to GAP measures and analyse how each of the measures are correlated, to highlight their interplay. Before delving into the measures, we setup the notation and classes to illustrate the relation.

Binary cross entropy

Binary cross entropy (BCE), as formulated in Eq. 2 is typically used as a loss function for optimizing a classifier. Although not a strict one-to-one correspondence, it is generally observed that minimizing the BCE loss leads to maximization of Accuracy. The BCE formulation does not consider imbalance across class frequency, hence might be biased towards the majority class label. It also does not consider the sensitive attributes.

Weighted cross entropy

One way to account for the imbalance across toxic and non-toxic labels (y) is weighted cross entropy (WCE), a variation of BCE that re-weights the error for the different classes proportional to their inverse frequency of labels (y). This re-weighting strategy is available in popular packages like SkLearn (Pedregosa et al., 2011) and is discussed in detail by (T.-Y. Lin et al., 2017). In Eq. 9 we replicate BCE (Eq. 2) terms twice which only introduces a duplication without formulation alteration. To ensure class balancing across toxic and non-toxic classes, we scale each of the duplicate terms *w.r.t.* to the sample count (toxic: P, non-toxic: Q) of the opposite class, while performing summation. When there's no class imbalance *i.e.*, P = Q, WCE reduces to 2·BCE, which has the same loss trajectory as BCE. This definition of WCE in Eq. 10 is differentiable, owing to its similar form to BCE and shares all the properties of BCE which allows it to be used as a loss for optimizing binary classifiers. WCE attempts to reduce the bias of the majority label due to the inverse sample count scaling, *i.e.*, majority and minority classes scaled by their opposite sample counts respectively.

$$BCE \propto -(\frac{1}{N}\sum_N y\log(\hat{y}) + (1-y)\log(1-\hat{y})) - (\frac{1}{N}\sum_N y\log(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{9}$$

$$WCE = \underbrace{-\frac{Q}{N}\sum_N y\log(\hat{y}) + (1-y)\log(1-\hat{y})}_{\text{BCE loss for toxic class } (y=1) \text{ with scaling}} \quad \underbrace{-\frac{P}{N}\sum_N y\log(\hat{y}) + (1-y)\log(1-\hat{y})}_{\text{BCE loss for non-toxic class } (y=0) \text{ with scaling}} \tag{10}$$

Remark 1. Rescaling the majority and minority labels (y) with their inverse frequency only ensures reduced bias towards the majority label. It does not optimize for equal accuracy across both the labels.

WCE w.r.t sensitive group attribute

While WCE accounts for the label imbalance in the dataset, it still does not consider the notion of fairness and the different sub-populations. The core idea behind WCE is that we can '*copy*' the loss function twice and then apply mathematical transformations to it, while maintaining the property of differentiability. We apply that same idea to derive our loss function for fairness. We calculate two separate instances of WCE: WCE (g = 1), calculated for data samples of group 1, and WCE (g = 0), calculated for data samples of group 0. GAP in essence is the 2-norm difference between the WCE's across each sensitive attribute *s*. The GAP loss function in Eq. 3 obtains a minimum only when both WCE errors match across the binary sensitive attribute. Note that unlike WCE, our measure GAP is defined as the difference between the two loss functions, rather than their weighted sum. Therefore, GAP reaches its minimum when the two sub-populations of sensitive group attribute (*s*) achieve the same accuracy.

Datasets

We consider two datasets: Davidson *et al.* (Davidson et al., 2017) for author demographics and the *Civil Comments* (Borkan et al., 2019) portion of *Wilds* (Koh et al., 2021) for target demographics. In each case, we frame the task as a binary classification problem (Toxic *vs.* non-Toxic, or "safe") with binary sensitive attributes (Majority *vs.* Minority, the under-represented, sensitive attribute). For Davidson, since an explicit train-test split does not exist, we randomly seed the dataset into train-test splits of 90% – 10%, following SkLearn's (Pedregosa et al., 2011) stratified sampling to ensure similar proportion of positive and negative tweets across the splits. For Wilds (Koh et al., 2021) we select tweets where more than 50% of annotators agreed on the gender of the target, and the toxicity label as well. Note that the annotation for male and female in the dataset is carried out separately, so it is possible that a tweet is targeted both towards male and female. We include such tweets in both portions as independent samples. Such pre-processing has been done across both train and test splits for evaluation purposes.

Setup

Experiments use a Nvidia 2060 RTX Super 8GB GPU, Intel Core i7-9700F 3.0GHz 8-core CPU and 16GB DDR4 memory. We use the Keras (Chollet, 2015) library on a Tensorflow 2.0 backend with Python 3.7 to train the networks in this paper. For optimization, we use AdaMax (Kingma & Ba, 2014) with parameters (*lr*=0.001) and 1000 steps per epoch. For each configuration, we did five independent runs to report mean and variance.

Runtime

The benefit of any Pareto HyperNetwork is to trace out the approximated front of feasible values during the training time, so that uses can extract neural weights corresponding to their desired trade-off values *a posteriori*. In our experiments, for the five trade-off values shown, one can achieve it in two ways.

1. Run the Bert model five times, each with different trade-off in the loss function
2. Run the Bert model one time, with the Pareto HyperNetwork supervising it.

The Bert model ran for 10 epochs with ~ 10 mins per epoch, for a total runtime ~ 100 mins. If we run the same configuration for five trade-offs, that would equate to ~ 500 mins of runtime. Thus, any additional trade-off measures the user desires would cost an extra ~ 100 mins each. The SUHNPF Pareto HyperNetwork on the other hand approximated a manifold of trade-off values supervising the Bert model, where the Bert model still takes ~ 100 mins with the supervising network taking additional ~ 60 mins for manifold approximation. Extracting the weights of the Bert model post-hoc takes an additional ~ 20 mins for each trade-off. Therefore, while both the prescribed approaches would roughly yield similar results from optimization of the Bert model, Approach 1 would take ~ 500 mins, while Approach 2 would take ~ 260 mins, resulting in a ~ 2× speedup via PFL.

Discussion on metric divergence

Table 4 reports the Accuracy Difference (AD) and Overall Accuracy (OA) values achieved for the different trade-off configurations of the Bert model, across three loss measures. This is a tabulated version of Fig. 1 (main text). Note that for trade-off $\alpha = 1$, only OA is maximized, hence none of the losses play any part, thus a common number across three columns, for each dataset. As the trade-off takes into account each of the loss measures, we empirically observe GAP to be performing best *w.r.t.* the other measures, since it is being optimized *w.r.t.* minimizing AD.

α	Accuracy Difference			Overall Accuracy			F1		
	GAP (Ours)	CLA	ADV	GAP (Ours)	CLA	ADV	GAP (Ours)	CLA	ADV
Davidson									
1.00	5.9 ± 0.1			88.9 ± 0.2			0.71 ± 0.02		
0.75	4.2 ± 0.1	5.0 ± 0.1	4.7 ± 0.1	88.5 ± 0.3	88.6 ± 0.2	88.2 ± 0.4	0.70 ± 0.01	0.69 ± 0.01	0.68 ± 0.00
0.50	2.7 ± 0.1	3.7 ± 0.1	3.2 ± 0.1	88.1 ± 0.5	88.3 ± 0.5	87.4 ± 0.6	0.69 ± 0.02	0.67 ± 0.01	0.65 ± 0.01
0.25	1.2 ± 0.1	2.4 ± 0.0	2.7 ± 0.1	87.7 ± 0.2	87.9 ± 0.4	86.8 ± 0.6	0.67 ± 0.01	0.65 ± 0.00	0.64 ± 0.01
0.00	0.1 ± 0.0	0.9 ± 0.0	2.4 ± 0.1	87.3 ± 0.1	87.6 ± 0.2	86.3 ± 0.4	0.66 ± 0.00	0.64 ± 0.02	0.61 ± 0.01
Wilds									
1.00	3.9 ± 0.2			84.7 ± 0.3			0.65 ± 0.02		
0.75	3.3 ± 0.1	3.6 ± 0.1	3.5 ± 0.1	84.6 ± 0.2	84.6 ± 0.1	84.5 ± 0.3	0.63 ± 0.02	0.62 ± 0.01	0.62 ± 0.02
0.50	2.6 ± 0.1	3.1 ± 0.1	2.9 ± 0.1	84.5 ± 0.4	84.6 ± 0.6	83.9 ± 0.4	0.62 ± 0.0	0.61 ± 0.01	0.60 ± 0.01
0.25	1.5 ± 0.0	2.5 ± 0.0	2.0 ± 0.1	84.5 ± 0.1	84.5 ± 0.2	83.8 ± 0.5	0.60 ± 0.01	0.60 ± 0.01	0.57 ± 0.01
0.00	0.3 ± 0.0	1.8 ± 0.1	1.3 ± 0.1	84.4 ± 0.1	84.4 ± 0.1	83.6 ± 0.2	0.58 ± 0.02	0.58 ± 0.01	0.55 ± 0.02

Table 4. Performance of GAP *vs.* CLA, ADV across two datasets in terms of Accuracy Difference (AD) and Overall Accuracy (OA). GAP achieves lower AD consistently across α settings and datasets, while a more modest drop in OA is observed across methods. $\alpha = 1$ minimizes WCE over labels only, hence same error across the three measures.

Performance of models on Wilds dataset

Table 5 shows the baseline results on the Wilds (Koh et al., 2021) dataset. The performance of the classifiers is similar *w.r.t.* Table 2, where due to focus on Overall Accuracy (OA), there is a gap between the group specific accuracies. This shows the existing bias across the three neural models, with the BERT based model performing relatively better than the rest.

Models	Overall %	Majority %	Minority %	AD %
CNN	83.90 ± 0.2	86.11 ± 0.1	81.27 ± 0.2	4.84 ± 0.2
BiLSTM	83.94 ± 0.1	85.98 ± 0.2	81.52 ± 0.2	4.46 ± 0.1
BERT	84.71 ± 0.3	86.53 ± 0.1	82.49 ± 0.2	4.04 ± 0.2

Table 5. Baseline accuracy results on Wilds (Koh et al., 2021) dataset when maximizing overall accuracy (OA) only. Results show consistent bias of higher accuracy for the Majority.