# Towards Equal Opportunity Fairness through Adversarial Learning

Anonymous ACL submission

## Abstract

Adversarial training is a common approach for bias mitigation in natural language processing. Although most work on debiasing is motivated by equal opportunity, it is not explicitly captured in standard adversarial training. In this paper, we propose an augmented discriminator for adversarial training, which takes the target class as input to create richer features and more explicitly model equal opportunity. Experimental results over two datasets show that our method substantially improves over standard adversarial debiasing methods, in terms of the performance–fairness trade-off.

## 1 Introduction

While natural language processing models have achieved great successes across a variety of classification tasks in recent years, naively-trained models often learn spurious correlations with confounds like user demographics and socio-economic factors (Badjatiya et al., 2019; Zhao et al., 2018; Li et al., 2018a).

Various fairness criteria have been proposed to quantify fairness under different conditions. *Equal opportunity*, for example, is satisfied if a binary classification model has an equal positive prediction rate for the advantaged class as for other disadvantaged classes, as measured by the difference in true positive rate (TPR GAP) between protected groups (Hardt et al., 2016). In addition to TPR, *equalized odds* also considers the FPR GAP, and as such is satisfied when model predictions are independent of the protected attribute, conditioned on the true label (Hardt et al., 2016). *Demographic parity* is another well-known fairness metric (Feldman et al., 2015), which is satisfied if protected groups have equal positive prediction rates (with no further conditioning).

A common way of mitigating bias relies on "unlearning" discriminators during the debiasing process. For example, in adversarial training, an encoder and discriminator are trained such that the encoder attempts to prevent the discriminator from identifying protected attributes (Zhang et al., 2018; Li et al., 2018a; Han et al., 2021c). In this, each training instance must be annotated with both the main task label and protected attribute.

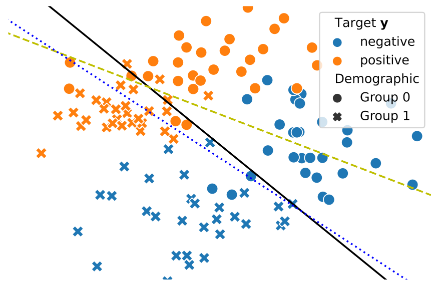
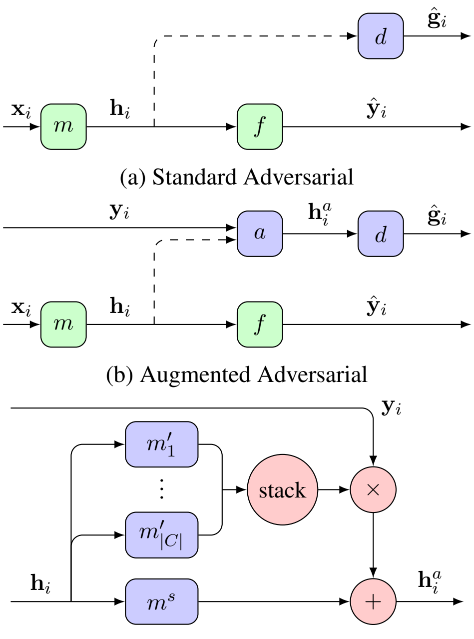



Figure 1: The black solid, yellow dashed, and blue dotted lines are the decision boundaries of linear discriminators for demographic trained over all instances, $\mathbf{y}$ = positive, and $\mathbf{y}$ = negative, resp.

Although the most popular fairness metric is equal opportunity, standard adversarial training does not consider the target label, which is fundamental to equal opportunity (acknowledging the correlation between target labels and protected attributes). Figure 1 shows a toy example where hidden representations are labelled with the associated target labels via colour, and protected labels via shape. Taking the target label information into account and training separate discriminators for each of the two protected attributes, it can be seen that the linear decision boundaries are quite distinct, and each is different from the decision boundary when the protected attribute is not taken into consideration.

In this paper, we propose a novel discriminator architecture that captures the individual protected attributes during adversarial training. Experiments show that our method consistently outperforms



(a) Standard Adversarial

(b) Augmented Adversarial

(c) Augmentation layer

Figure 2: Proposed model architectures. Dashed lines denote gradient reversal in adversarial learning. Green and blue rounded rectangles are the trainable neural network layers for target label classification and bias mitigation, resp. Red circles are operations.

standard adversarial learning.

## 2 Methods

Here we describe the methods employed in this paper. Formally, as shown in Figure 2a, given an input $\mathbf{x}_i$ annotated with main task label $\mathbf{y}_i$ and protected attribute label $\mathbf{g}_i$, a main task model consists of two connected parts: the encoder $\mathbf{h}_i = m(\mathbf{x}_i; \boldsymbol{\theta}^m)$ is trained to compute the hidden representation from an input $\mathbf{x}_i$, and the classifier makes prediction, $\hat{\mathbf{y}}_i = f(\mathbf{h}_i; \boldsymbol{\theta}^f)$. During training, a discriminator $d$, parameterized by $\boldsymbol{\phi}^d$, is trained to predict $\hat{\mathbf{g}}_i = d(\mathbf{h}_i; \boldsymbol{\phi}^d)$ from the final hidden-layer representation $\mathbf{h}_i$.

### 2.1 Adversarial Learning

Following the setup of Li et al. (2018a); Han et al. (2021c), the optimisation objective for standard adversarial training is:

$$\min_{\boldsymbol{\theta}^*} \max_{\boldsymbol{\phi}^*} \mathcal{X}(\mathbf{y}, \hat{\mathbf{y}}) - \lambda \mathcal{X}(\mathbf{g}, \hat{\mathbf{g}}) \qquad (1)$$

where $\boldsymbol{\theta}^* = \{\boldsymbol{\theta}^m, \boldsymbol{\theta}^f\}$, $\boldsymbol{\phi}^* = \{\boldsymbol{\phi}^d\}$, $\mathcal{X}$ is the cross entropy loss, and $\lambda$ is a trade-off hyperparameter. Solving this minimax optimization problem encourages the main task model hidden representation $\mathbf{h}$ to be informative to $f$ and to be uninformative to $d$.

### 2.2 Discriminator with Augmented Representation

As illustrated in Figure 2b, we propose **augmented discrimination**, a novel means of strengthening the adversarial component. Specifically, an extra augmentation layer $a$ is added between $m_{\mathbf{y}}$ and $d$, where $a$ takes the $\mathbf{y}$ into consideration to create richer features, i.e., $\hat{\mathbf{g}}_i = d(a(\mathbf{h}_i; \mathbf{y}_i; \boldsymbol{\phi}^a); \boldsymbol{\phi}^d)$.

**Augmentation Layer** Figure 2c shows the architecture of the proposed augmentation layer. Inspired by the domain-conditional model of Li et al. (2018b), the augmentation layer $a$ consists of one shared projector and $|C|$ specific projectors,$\{m^s, m'_1, m'_2, \ldots, m'_{|C|}\}$, where $|C|$ is the number of target classes.

Formally, let $m^s(\mathbf{h}; \boldsymbol{\phi}^s)$ be a function parameterized by $\boldsymbol{\phi}^s$ which projects a hidden representation $\mathbf{h}$ to $\mathbf{h}^s$ representing features w.r.t $\mathbf{g}$ that are *shared* across classes, and $m'_j(\mathbf{h}; \boldsymbol{\phi}^j)$ be a class-specific function to the $j$-th class which projects the same hidden representation $\mathbf{h}$ to $\mathbf{h}'^j$ capturing features that are *private* to the $j$-th class. In this paper, we employ the same architecture for shared and all private projectors. The resulting output of the augmentation layer is

$$\mathbf{h}_i^a = a(\mathbf{h}_i; \mathbf{y}_i; \boldsymbol{\phi}^a) = \mathbf{h}_i^s + \sum_{j=1}^{|C|} \mathbf{y}_{i,j} \mathbf{h}'^j,$$

where $\boldsymbol{\phi}^a = \{\boldsymbol{\phi}^s, \boldsymbol{\phi}^1, \ldots, \boldsymbol{\phi}^{|C|}\}$, and $\mathbf{y}_{i,:}$ is 1-hot. Moreover, let $\boldsymbol{\phi}^* = \{\boldsymbol{\phi}^d, \boldsymbol{\phi}^a\}$, the training objective is the same as Equation 1.

Intuitively, $d$ is able to make better predictions over $\mathbf{g}$ based on $\mathbf{h}^a$ than the vanilla $\mathbf{h}$ due to the enhanced representations provided by $a$. More formally, as the augmented discriminator models the conditional probability $\Pr(g|h, y)$, the unlearning of the augmented discriminator encourages conditional independence $\mathbf{h} \perp \mathbf{g}|\mathbf{y}$, which corresponds directly to the equal opportunity criterion.

## 3 Experiments

In order to compare our method with previous work, we follow the experimental setting of Han et al. (2021c). We provide full experimental details in Appendix B.[1]

[1] We will release source code and datasets upon acceptance.

### 3.1 Evaluation Metrics

Following Han et al. (2021c); Ravfogel et al. (2020), we use overall accuracy as the performance metric, and measure TPR GAP for equal opportunity fairness. For multiclass classification tasks, we report the quadratic mean (RMS) of TPR GAP over all classes. While in a binary classification setup, TPR and TNR are equivalent to the TPR of the positive and negative classes, respectively, so we employ the RMS TPR GAP in this case also. For GAP metrics, the smaller, the better, and a perfectly fair model will achieve 0 GAP.

More specifically, the calculation of RMS TPR GAP consists of aggregations at the group and class levels. At the group level, we measure the absolute TPR difference of each class between each group and the overall TPR $GAP^{TPR}_{G,y} = \sum_{g \in G} |TPR_{g,y} - TPR_y|$, and at the next level, we further perform the RMS aggregation at the class level to get the RMS TPR GAP as $GAP = \sqrt{\frac{1}{|Y|}\sum_{y \in Y}(GAP^{TPR}_{G,y})^2}$.

### 3.2 Dataset

Following Subramanian et al. (2021), we conduct experiments over two NLP classification tasks — sentiment analysis and biography classification — using the same dataset splits as prior work.

**Moji** This sentiment analysis dataset was collected by Blodgett et al. (2016), and contains tweets that are either African American English (AAE)-like or Standard American English (SAE)-like. Each tweet is annotated with a binary 'race' label (based on language use: either AAE or SAE) and a binary sentiment score determined by (redacted) emoji contained in it.

**Bios** The second task is biography classification (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 label economic status (wealthy vs. rest) based on the country the individual is based in, as geotagged from the first sentence of the biography. For bias evaluation and mitigation, we consider the intersectional groups, i.e., the Cartesian product of the two protected attributes, leading to 4 intersectional classes: female–wealthy, female–rest, male–wealthy, and male–rest.

### 3.3 Models

We first implement a naively trained model on each dataset, without explicit debiasing. On the Moji dataset, we use DeepMoji (Felbo et al., 2017) as the fixed encoders to get 2304d representations of input texts. For the Bios dataset, we use uncased BERT-base (Devlin et al., 2019), taking the 'AVG' representations extracted from the pretrained model, without further fine-tuning.

For adversarial method, both the Adv and augmented Adv, we jointly train the discriminator and classifier. Again, we follow Han et al. (2021c) in using a non-linear discriminator, which is implemented as a trainable 3-layer MLP.

One problem is that the natural distribution of the demographic labels is imbalanced, e.g. in Bios 87% nurses are female while 90% surgeons are male. In order to deal with this label imbalance, we reweight each instance inversely proportional to the frequency of its demographic label within its target class when training the discriminators (Han et al., 2021a).

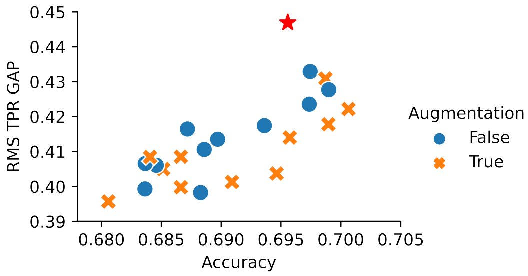
Another common problem is that a large number of instances are not annotated with protected attributes, e.g. only 28% instances in the Bios dataset are annotated with both gender and economic status labels. The standard adversarial method has required all training instances are annotated with protected attributes, and thus can only be trained over a full-labelled subset, decreasing the training set size significantly. To maintain the performance of the debiased model, we follow Han et al. (2021b) in decoupling the training of the model and the discriminator, making it possible to use all instances for model training at a cost of the performance-fairness trade-off.

### 3.4 Main results

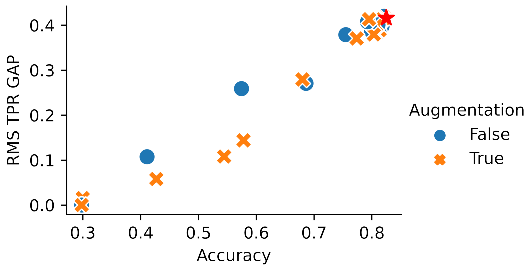
Now we compare the adversarial debiasing with our proposed augmented discriminator against the standard discriminator.

Recall that $\lambda$ is the most sensitive hyperparameter, which controls the performance–fairness trade-off. To explore trade-offs of our proposed method at different levels, we tune $\lambda$ log-uniformly to get a series of candidate models.

Figure 3 shows the results. Each point denotes a candidate model with a given $\lambda$, and we take the average over 5 runs with different random seeds.




(a) MOJI




(b) BIOS

Figure 3: Adversarial trade-offs. Red star denotes the naively-trained model without debiasing. Our proposed model (orange crosses) substantially outperforms standard adversarial training (blue circles). The bottom-right represents ideal model with the idea performance and fairness.

Over both datasets, our proposed method consistently achieves better performance–fairness trade-off. I.e., the adversarial method with augmented discriminator achieves smaller GAP (better fairness) at the same accuracy level, and achieves better accuracy at the same GAP level.

**Without Decoupling** As stated in Section 3.3, to use full datasets for the main task model training, we have been using decoupled adversarial training for both datasets at a cost of the trade-off. Due to the different training setting, such results are not directly comparable to previous work. To provide comparability with past work, we consider the full-labelled subset setting over the MOJI dataset without decoupling and use the best hyperparameters for adversarial training from Han et al. (2021c).

Consistent with the decoupled training in Figure 3, our method increase the trade-off of the adversarial training. Averaged over 5 runs with different random seeds, the standard adversarial training achieves 72.73% accuracy and 18.94% GAP, while our augmented method shows substantially better fairness (5.49% absolute improvement in GAP) and similar performance (73.01% Accuracy). We elaborate more on these results in Appendix C.

| Model | MOJI ↑ | BIOS ↑ |
|---|---|---|
| Random | 50.00 | 25.00 |
| DISCRIMINATOR | 88.25 | 89.87 |
| +LINEAR-AUGMENTED | 88.56 | 90.13 |
| +NONLINEAR-AUGMENTED | 88.68 | 90.53 |

Table 1: Demographic label prediction accuracy (%) for discriminators over the MOJI and BIOS datasets.

### 3.5 Analysis

We test our hypothesis that *the augmented discriminator can identify protected attributes better than the standard method.* Intuitively, adversarial debiasing relies on unlearning the discriminator, and thus the better the discriminators perform, the better the fairness.

On each dataset, we train the main task model until convergence, and then extract hidden representations, which are inputs to the adversary training.[2]

We compare three different discriminators: (1) DISCRIMINATOR, which is a vanilla discriminator that takes $\mathbf{h}$ as input; (2) DISCRIMINATOR with LINEAR-AUGMENTED inputs, i.e., all projectors within the augmentation layer are linear functions; and (3) DISCRIMINATOR with NONLINEAR-AUGMENTED inputs, which is used as our reported model.

Table 1 summarises the results over both datasets. By using augmented inputs based on the target labels, both LINEAR-AUGMENTED and NONLINEAR-AUGMENTED consistently outperforms DISCRIMINATOR on both datasets, confirming our hypothesis. Moreover, NONLINEAR-AUGMENTED DISCRIMINATOR learns nonlinear projections for each channel in the augmentation layer and achieves the best results.

## 4 Conclusion

We introduce an augmented discriminator for adversarial debiasing. We conducted experiments over a binary tweet sentiment analysis with binary author race attribute and a multiclass biography classification with the multiclass protected attribute. Results showed that our proposed method, considering the target label, can more accurately identify protected information and thus achieves better performance–fairness trade-off than the standard adversarial training.

[2] We focus on training the discriminators only, not joint training as done elsewhere.

## References


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

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

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

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

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

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

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018a. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018b. What's in a domain? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 474–479, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

| Profession | Total | male_rest | male_wealthy | female_rest | female_wealthy |
|---|---|---|---|---|---|
| professor | 21715 | 0.092 | 0.462 | 0.073 | 0.374 |
| physician | 7581 | 0.084 | 0.424 | 0.080 | 0.411 |
| attorney | 6011 | 0.099 | 0.512 | 0.062 | 0.327 |
| photographer | 4398 | 0.111 | 0.531 | 0.056 | 0.303 |
| journalist | 3676 | 0.093 | 0.407 | 0.086 | 0.414 |
| nurse | 3510 | 0.011 | 0.075 | 0.149 | 0.764 |
| psychologist | 3280 | 0.065 | 0.307 | 0.105 | 0.523 |
| teacher | 2946 | 0.061 | 0.351 | 0.095 | 0.492 |
| dentist | 2682 | 0.113 | 0.521 | 0.063 | 0.303 |
| surgeon | 2465 | 0.124 | 0.727 | 0.024 | 0.126 |
| architect | 1891 | 0.116 | 0.641 | 0.034 | 0.208 |
| painter | 1408 | 0.089 | 0.473 | 0.075 | 0.363 |
| model | 1362 | 0.025 | 0.149 | 0.130 | 0.696 |
| poet | 1295 | 0.073 | 0.459 | 0.082 | 0.385 |
| software_engineer | 1289 | 0.137 | 0.697 | 0.025 | 0.140 |
| filmmaker | 1225 | 0.096 | 0.556 | 0.059 | 0.289 |
| composer | 1045 | 0.142 | 0.704 | 0.017 | 0.137 |
| accountant | 1012 | 0.095 | 0.553 | 0.063 | 0.289 |
| dietitian | 730 | 0.012 | 0.051 | 0.121 | 0.816 |
| comedian | 499 | 0.090 | 0.693 | 0.030 | 0.186 |
| chiropractor | 474 | 0.143 | 0.618 | 0.032 | 0.207 |
| pastor | 453 | 0.146 | 0.594 | 0.035 | 0.225 |
| paralegal | 330 | 0.027 | 0.124 | 0.148 | 0.700 |
| yoga_teacher | 305 | 0.030 | 0.134 | 0.121 | 0.715 |
| interior_designer | 267 | 0.041 | 0.165 | 0.124 | 0.670 |
| personal_trainer | 264 | 0.098 | 0.413 | 0.068 | 0.420 |
| dj | 244 | 0.156 | 0.709 | 0.025 | 0.111 |
| rapper | 221 | 0.154 | 0.747 | 0.009 | 0.090 |
| **Total** | 72578 | 0.089 | 0.451 | 0.075 | 0.386 |

Table 2: Training set distribution of the BIOS dataset.

# A Dataset

## A.1 MOJI

We use the train, dev, and test splits from Han et al. (2021c) of 100k/8k/8k instances, respectively. This training dataset has been artificially balanced according to demographic and task labels, but artificially skewed in terms of race–sentiment combinations, as follows: AAE–happy = 40%, SAE–happy = 10%, AAE–sad = 10%, and SAE–sad = 40%.

## A.2 BIOS

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.[3] Following Ravfogel et al. (2020), we randomly split the dataset into train (65%), dev (10%), and test (25%).

Table 2 shows the target label distribution and protected attribute distribution.

# B Reproducibility

## B.1 Computing infrastructure

We conduct all our experiments on a Windows server with a 16-core CPU (AMD Ryzen Threadripper PRO 3955WX), two NVIDIA GeForce RTX 3090s with NVLink, and 256GB RAM.

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

## B.2 Computational budget

Over the MOJI dataset, we run experiments with 108 different hyperparameter combinations (each for 5 runs with different random seeds) in total, which takes around 300 GPU hours in total and 0.56 hrs for each run. Over the BIOS dataset, we run experiments with 162 different hyperparameter combinations for around 466 GPU hours and 0.58 hrs for each run.

## B.3 Model architecture and size

In this paper, we used pretrained models as fixed encoder, and the number of fixed parameters of DeepMoji (Felbo et al., 2017) for MOJI and uncased BERT-base (Devlin et al., 2019) for BIOS are approximately 22M and 110M, resp. The number of remaining trainable parameters of the main model is about 1M for both tasks.

As for the standard discriminator, we follow (Han et al., 2021b) and use the same architecture for both tasks, leading to a 3-layer MLP classifier with around 144k parameters. When comparing NONLINEAR-AUGMENTED DISCRIMINATOR with DISCRIMINATOR, we use the same number of hidden layers by replacing the hidden layer of the DISCRIMINATOR with the projectors in the augmentation layer. Taking the NONLINEAR-AUGMENTED DISCRIMINATOR as an example, we use 2 hidden layers with activation functions for each projector of the augmentation layer, and the DISCRIMINATOR is a single-layer MLP. Similarly, for the LINEAR-AUGMENTED DISCRIMINATOR, augmentation projectors and DISCRIMINATOR have 1 and 2 hidden layers, resp. The number of parameters of the non-augmentation layer correlated with the number of components, i.e. the number of classes for the main task. Thus there are 284k and 4M parameters for MOJI and BIOS, resp.

## B.4 Hyperparameters

For each dataset, all main task model models in this paper share the same hyperparameters as the standard model. Hyperparameters are tuned using grid-search, in order to maximize accuracy for the standard model. Table 3 summaries search space and best assignments of key hyperparameters.

To explore trade-offs of our proposed method at different levels, we tune $\lambda$ log-uniformly to get a series of candidate models. Specifically, the search space of $\lambda$ with repsect to MOJI and BIOS are *loguniform-float*$[10^{-4}, 10^{4}]$ and *loguniform-*

| | | Best assignment | |
|---|---|---|---|
| **Hyperparameter** | **Search space** | **MOJI** | **BIOS** |
| number of epochs | - | 100 | |
| patience | - | 10 | |
| embedding size | - | 2304 | 768 |
| hidden size | - | 300 | |
| number of hidden layers | *choice-integer*[1, 3] | 2 | |
| batch size | *loguniform-integer*[64, 2048] | 1024 | 512 |
| output dropout | *uniform-float*[0, 0.5] | 0.5 | 0.3 |
| optimizer | - | Adam (Kingma and Ba, 2015) | |
| learning rate | *loguniform-float*$[10^{-6}, 10^{-1}]$ | $3 \times 10^{-3}$ | $10^{-3}$ |
| **l**earning **r**ate **s**cheduler | - | reduce on plateau | |
| **LRS** patience | - | 2 epochs | |
| **LRS** reduction factor | - | 0.5 | |

Table 3: Search space and best assignments on the BIOS dataset

| **Model** | **Accuracy↑** | **GAP ↓** |
|---|---|---|
| STANDARD | $72.1 \pm 0.1$ | $40.8 \pm 0.3$ |
| ADV | $72.7 \pm 2.1$ | $18.9 \pm 2.5$ |
| DADV | $74.3 \pm 1.8$ | $14.6 \pm 3.0$ |
| Augmented ADV | $73.0 \pm 2.5$ | $13.4 \pm 1.9$ |

Table 4: Results over the sentiment analysis (MOJI) task. Evaluation results $\pm$ standard deviation (%) on the test set, averaged over 5 runs with different random seeds. "↑" and "↓" indicate that higher and lower performance, resp., is better for the given metric. STANDARD: naively trained mdoel without debiasing. ADV: the adversarial debiasing method presented by Li et al. (2018a). DADV: the recent STOA variation of adversarial debiaisng proposed by Han et al. (2021c).

*float*$[10^{-2}, 10^{2}]$, resp.

## C Ablation Study

Table 4 shows evaluation results over the MOJI dataset. Under the same training setting (i.e., without decoupling), our proposed approach consistently outputs the STANDARD and ADV. Our method achieves similar trade-off as the STOA method DADV, lower accuracy but better fairness. However, DADV relies on training multiple adversaries, leading to a much higher time complexity, and the training time of DADV with 3 subdiscriminators will be almost 3 times as long as ours.