Leveraging Prototypical Representations for Mitigating Social Bias without Demographic Information

Shadi Iskander **Kira Radinsky** **Yonatan Belinkov**
shadi.isk@campus.technion.ac.il
kirar@cs.technion.ac.il belinkov@technion.ac.il
Technion – Israel Institute of Technology

Abstract

Mitigating social biases typically requires identifying the social groups associated with each data sample. In this paper, we present DAFAIR, a novel approach to address social bias in language models. Unlike traditional methods that rely on explicit demographic labels, our approach does not require any such information. Instead, we leverage predefined prototypical demographic texts and incorporate a regularization term during the fine-tuning process to mitigate bias in the model's representations. Our empirical results across two tasks and two models demonstrate the effectiveness of our method compared to previous approaches that do not rely on labeled data. Moreover, with limited demographic-annotated data, our approach outperforms common debiasing approaches.[1]

1 Introduction and Background

The presence of social bias in training data presents a significant challenge in the development of language models for real-world applications. While these models possess remarkable capabilities, biases within the data can lead to unfair outcomes. Mitigating these biases is crucial, but it becomes particularly challenging when acquiring or accessing sensitive attribute labels is costly or unfeasible.

Studies showed that language models have the ability to capture demographic information about the writer, including race or gender, within their representations (Caliskan et al., 2017; Zhao et al., 2018). However, this capability can introduce unintended biases, leading to discriminatory outputs (De-Arteaga et al., 2019).

Common approaches for social bias mitigation require explicit annotation of biases for each sample in the data (Beutel et al., 2017; Zhang et al., 2018). Recent concept removal methods (Ravfogel et al., 2020, 2022a,b; Iskander et al., 2023) have shown promise in addressing social bias by removing sensitive attributes. These approaches rely on training classifiers for predicting the sensitive attribute, and training such classifiers typically requires a significant amount of annotated data.

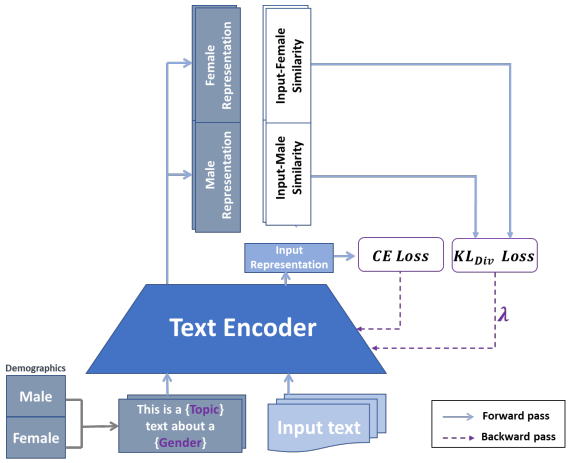


Figure 1: Our debiasing method consists of defining task-specific representations for each social attribute, measuring similarity in the representation space for each example, and utilizing the KL loss to encourage uniform probabilities across social groups.

A promising line of research has emerged that aims to mitigate bias without relying on explicit information about the biases present in the data. For instance, Just Train Twice (JTT) (Liu et al., 2021) employs a two-step training process. In the second step, a second model is trained on up-weighed training examples that were misclassified by the first model. Another method is BLIND (Orgad and Belinkov, 2023), which introduces a success detector and down-weighs examples for which the detector accurately predicts the outcome.

In this paper, we propose **DAFAIR**: **D**emographics-**A**gnostic **Fair**ness, a novel approach for mitigating social bias during the fine-tuning process of language models, without

[1]Our code is available at https://github.com/technion-cs-nlp/DAFair

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 379–390
June 16-21, 2024

relying on demographic information. Our approach aims to ensure equal similarity between the representation of a text and prototypical representations of different demographic groups. For instance, when classifying a biographical text of a person into their profession, our method aims to make the representation of the text equally similar to the representations of both males and females. More concretely, DAFAIR first defines prototypical texts, such as "This is a biography about a male" and "This is a biography about a female". It then adds a regularization term that makes the representation of a training example equally similar to the representations of each of the prototypical texts (Figure 1).

Furthermore, we extend our approach to scenarios where limited demographic-annotated data is available. In such cases, we obtain the prototypical representation by averaging the sample representations corresponding to each social attribute.

We evaluate the effectiveness of DAFAIR and its extension on two tasks: occupation prediction and sentiment analysis of twitter posts. In these tasks, we investigate the performance of our approach under the settings of limited demographic labels or no labels at all, reflecting real-world scenarios where labeled data is challenging to obtain. The experimental results with two base models demonstrate that our approach outperforms previous approaches that do not rely on demographic information, as well as common approaches with limited data.

2 Methodology

Assume a dataset $D = \{t_i, y_i, z_i\}_{i=1}^{n}$ of input texts $t_i \in \mathcal{T}$, main task labels $y_i \in \mathcal{Y}$, and sensitive attributes $z_i \in \mathcal{Z}$ that correspond to discrete demographic attributes, such as race. This sensitive attribute can either be unobserved during training or available in a small subset of the data. Our aim is to learn a model $F : \mathcal{T} \rightarrow \mathbb{R}^{|\mathcal{Y}|}$ that does not rely on the sensitive attribute z_i in its prediction.

2.1 Demographic-Agnostic Fairness Approach

Our method, depicted in Fig. 1, involves several key steps to mitigate social bias. First, we establish multiple representations for each group of sensitive attributes (Section 2.1.1). During fine-tuning, we measure similarity between the representation of an example and each attribute representation. These similarities are then transformed into a probability distribution. Subsequently, we use the Kullback-Leibler (KL) divergence loss (Kullback and Leibler, 1951) to compare the predicted probability distribution with a uniform distribution (Section 2.1.3). This loss term encourages the model to mitigate bias by penalizing deviations from a uniform distribution, promoting fair and unbiased predictions.

2.1.1 Social Attribute Representations

We employ two approaches to define representations for social attribute groups, depending on the availability of labels: no labels, or few labels.

Pre-defined Representations (No Labels). In the absence of labeled data, we leverage semantic similarity and define pairs of texts that capture the models' understanding of text describing different social attribute groups. For example, to represent gender in an occupation prediction task we can use the encoder's representations of "This biography is about a man" and "This biography is about a woman". To generate these pre-defined representations, we employ a generative model. We provided ChatGPT (OpenAI, 2022) with a description of the approach, DAFAIR, along with a description of each dataset and task, and instructed the model to produce 10 pairs of prototypical texts for each task. The prototypical texts (Tables 7 and 8) and the full prompt (Figure 4) are provided in the appendix.

Data-driven Representations (Few Labels). When a limited number of labels are available, we leverage the representations generated by the text encoder to derive data-driven representations for each labeled group. Specifically, we calculate the mean representation of each labeled group using the available labeled samples. We call this method **SEMI-DAFAIR**.

We will assume a binary case for simplicity and denote the pair of representations as $[X_A, X_B]$.[2]

2.1.2 Ensemble of Representations

Inspired by Stacey et al. (2020), we adopt an ensemble approach by leveraging multiple pairs of representations instead of using a single pair. We denote the ensemble of representations as $\{[X_A^j, X_B^j]\}_{j=1}^{K}$, where K represents the number of pairs.

In the case of pre-defined representations, we use multiple pre-defined pairs that capture different perspectives. For data-driven representations, we divide the labeled data into K partitions and calculate the mean representation for each partition, resulting in K pairs of representations.

[2] Our approach can be extended to handle multiple social attribute groups, denoted as $[X_A, X_B, X_C, ...]$.

By incorporating an ensemble of representations, we aim to capture a diverse range of information and perspectives related to biases.

2.1.3 Calculating KL Loss

During fine-tuning, we calculate the similarity between the representation of example X_i and each pair of attribute representations using dot product:

$$[sim_A^j, sim_B^j] = X_{\text{i}} \cdot [X_A^j, X_B^j] \quad (1)$$

Then we apply the softmax function $\sigma(a,b) = \frac{e^a}{e^a+e^b}$ to obtain the similarity distribution:

$$d_{sim}^j = \sigma(sim_A^j, sim_B^j) \quad (2)$$

To calculate the overall KL loss, we compute KL divergence between each of the similarity distributions d_{sim}^j and a uniform distribution d_{uni}:

$$L_{kl} = \sum_{j=1}^{K} D_{KL}(d_{sim}^j, d_{uni}) \quad (3)$$

Finally, we compute the total loss:

$$L_{total} = L_{ce} + \lambda L_{kl}, \quad (4)$$

where L_{ce} is the usual cross-entropy loss. The hyper-parameter λ adjusts the balance between task performance and fairness, providing flexibility to prioritize either aspect.

3 Experimental Setup

3.1 Tasks

We conduct experiments on two classification tasks: occupation prediction and sentiment analysis, focusing on social bias related to gender and race.

Occupation Prediction. We use the Bias in Bios Dataset (De-Arteaga et al., 2019). The task involves predicting the occupation of individuals based on their biographical information. The dataset consists of 394K biographies of 28 professions, with gender annotations.

Twitter Sentiment Analysis. We follow the setup of Elazar and Goldberg (2018), who leveraged a Twitter dataset originally gathered by Blodgett et al. (2016). Elazar and Goldberg (2018) used emojis in the tweets to derive sentiment labels for the classification task. Tweets are labeled with sociolects—African American English (AAE) or Standard American English (SAE)—based on the author's geo-location, serving as a proxy for their racial identity. We work with a subset of 100K samples, consistent with Orgad and Belinkov (2023).

3.2 Models

We use two pre-trained text encoders: BERT (Devlin et al., 2019) and DeBERTa-V3 (He et al., 2022). By considering two diverse tasks and different models, we can evaluate the effectiveness of our approach in mitigating social bias in various contexts and with different model architectures.

3.3 Metrics

Performance Evaluation. We evaluate the model's accuracy (**Acc**) on the downstream task to ensure that it has not been significantly affected.

Fairness Assessment. To evaluate extrinsic bias, we align with previous work (De-Arteaga et al., 2019; Ravfogel et al., 2020) and use the True Positive Rate Gap (**TPR-GAP**) as the main fairness metric to assess performance disparities across different protected attribute groups. Following the guidelines in Orgad and Belinkov (2022) for a comprehensive evaluation, we also incorporate statistical fairness metrics: **Independence**, **Separation** and **Sufficiency**. The metrics details and calculation procedures are provided in Appendix B.

3.4 Compared Methods

We compare our approach with several methods for bias mitigation and with a baseline (**Original**) without any debiasing procedure.

We compare with two existing methods that do not rely on demographic information:

JTT (Liu et al., 2021), which trains in a second phase on up-weighed hard examples.

BLIND (Orgad and Belinkov, 2023), which uses a success detector to down-weigh biased examples.

When only limited demographic labeled samples are available, we evaluate three methods:

INLP (Ravfogel et al., 2020) removes linear information from the neural representation by iteratively training a linear classifier to predict the demographic attribute from the representation, then projecting the representations to the null-space of the linear classifier.

RLACE (Ravfogel et al., 2022b) is similar to INLP with the goal of linear information removal from the neural representations. However, it uses a different approach of a linear minimax game.

IGBP (Iskander et al., 2023) overcome the drawbacks of INLP and RLACE which only remove linearly encoded information, and removes non-linear information from representations by gradient-based projections.

	Occupation Prediction		Sentiment Analysis	
Method	Accuracy ↑	TPR-GAP ↓	Accuracy ↑	TPR-GAP ↓
Original	83.43 ± 0.08	14.66 ± 0.51	79.18 ± 0.22	25.34 ± 0.82
JTT	81.34 ± 0.81	14.19 ± 1.08	78.08 ± 1.21	23.67 ± 1.87
BLIND	82.52 ± 0.23	13.76 ± 1.18	76.45 ± 0.67	22.68 ± 2.40
DAFAIR	82.32 ± 0.13	12.29 ± 0.32	77.20 ± 1.17	21.72 ± 0.82

Table 1: Evaluation results for occupation prediction and sentiment analysis tasks with BERT as the text encoder.

3.5 Settings

No Demographic Labels. In this setting, we explore scenarios where demographic labels are not available. We evaluate the performance of demographic-agnostic methods: JTT, BLIND and DAFAIR.

Limited Demographic Labels. Additionally, we investigate a scenario where we have limited access to demographic labels. In this setting, we apply information removal methods along with SEMI-DAFAIRwhile varying the size of the available demographic-labeled data to analyze their effectiveness.

We run each method using 5 random seeds and report the mean and standard deviation of the test results. More details on training setup and evaluation procedures are described in Appendix A.

3.6 DAFAIR Hyperparameters

Under the setting of no demographic labels, there is no validation set to optimize the selection of prototypical texts or the number of pairs. To avoid dependency on the choice of prototypical representations, we first generate $N > K$ pairs, and within each iteration, we randomly sample K pairs. For all experiments, we set $N = 10$, $K = 4$ to capture diverse associations of the training samples with demographic attributes, without relying on an extensive set of pairs. In Section 4.3, we analyze the impact of K on the model's performance and assess its implications on fairness and bias mitigation.

λ Tuning. To perform λ tuning without the need for a validation set with demographic annotations, we adopt Orgad and Belinkov (2023)'s strategy that prioritizes selecting the most radical parameter, while ensuring that the downstream task accuracy remains above 0.97 of the original accuracy. More details are described in Appendix A.

4 Results and Analysis

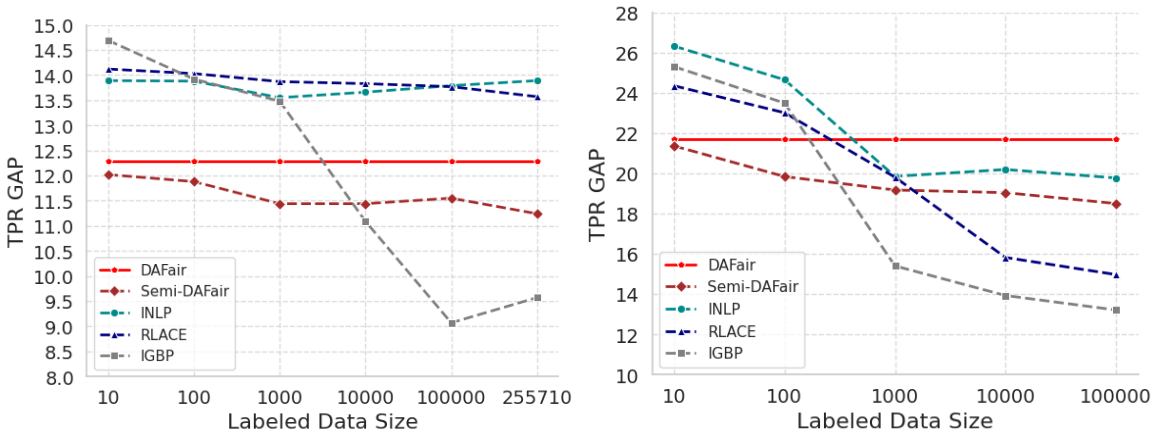
4.1 No Demographic Labels

Table 1 presents the evaluation results on the occupation prediction and twitter sentiemnt tasks using BERT as encoder. In both tasks, Our proposed method, DAFAIR, achieves a slightly lower accuracy compared to the finetuned model. However, it significantly reduces the TPR-GAP, outperforming BLIND and JTT in mitigating bias related to gender or race. This could be attributed to the fact that JTT and BLIND do not directly address social bias, but assign different weights to examples based on their difficulty. In contrast, DAFAIR uses a regularization term designed to lower the association of text representation with specific social groups, which might explain the superior reduction of social bias measures. Evaluation with other statistical fairness metrics reveals similar patterns to TPR-GAP (Appendix C). Results with the DeBERTA-V3 model exhibit same trend, as presented in Appendix C.

4.2 Limited Demographic Labels

Figure 2 presents TPR-GAP for both tasks under different levels of labeled data for social attributes, showcasing the performance of various debiasing methods (Section 3.4), including our proposed methods, DAFAIR and SEMI-DAFAIR.[3] While using no labels (horizontal solid lines), DAFAIR outperforms other methods even when they are provided a limited number of labels (up to 100 in twitter sentiment and 1000 in occupation prediction). DAFAIR further benefits from labels (SEMI-DAFAIR lines), even outperforming prior methods with limited labeled data. With an abundance of labeled data (1000 in sentiment and 100K in occupation prediction), other methods perform better.

[3]Accuracy figures are available in Appendix C.1.



(a) Occupation Prediction

(b) Sentiment Analysis

Figure 2: Effect of bias mitigation methods on TPR-GAP with varying labeled data sizes. In scenarios with limited demographic-annotated data, our approach outperforms common debiasing approaches.

The aim of information removal methods is to learn and neutralize decision boundaries between different social attributes using supervised learning. However, limited labeled examples may hinder classifiers from modeling social attribute subspace in high-dimensional spaces. In SEMI-DAFAIR, we aim to mitigate associations with specific social groups. Surprisingly, a small set of labeled data seems sufficient for this purpose, as more labeled data does not offer additional benefits.

4.3 Effect of Number of Prototypical Texts

To investigate the effect of the number of prototypical text pairs (K) on model performance, we conducted experiments with varying K values of (1, 2, 4, 8). The results presented in Table 2 reveal that all K values contributes to the reduction of the TPR-GAP without affecting accuracy. While larger values of K result in more substantial reductions, the incremental improvements become less significant for $K > 2$. These findings suggest that a small K may be sufficient for DAFAIR.

5 Conclusion

We introduced DAFAIR, a novel approach for mitigating social bias in language models without explicit demographic information. Our method leverages semantic similarity to manipulate the model's text representations during finetuning to promote fairness. Experimental results on two tasks and under different settings demonstrated the effectiveness of DAFAIR in reducing bias and improving fairness while maintaining competitive downstream task performance, even with limited or no labeled demographic data. With its focus on social bias, DAFAIR offers a flexible framework adaptable to address other forms of bias through the modification of prototypical texts.

In conclusion, our approach offers a practical and flexible solution for bias mitigation in real-world applications, contributing to the development of fairer language models.

	Occupation Prediction		Sentiment Analysis	
K	Accuracy ↑	TPR-GAP ↓	Accuracy ↑	TPR-GAP ↓
Original	83.43 ± 0.08	14.66 ± 0.51	79.18 ± 0.22	25.34 ± 0.82
1	82.85 ± 0.18	13.11 ± 0.62	77.52 ± 1.23	22.69 ± 1.81
2	82.55 ± 0.23	12.84 ± 0.59	77.16 ± 1.12	22.21 ± 1.08
4	82.32 ± 0.13	12.29 ± 0.32	77.20 ± 1.17	21.72 ± 0.82
8	82.29 ± 0.30	12.20 ± 0.25	77.47 ± 0.94	22.17 ± 1.17

Table 2: Effect of varying K on accuracy and TPR-GAP for occupation prediction and sentiment analysis tasks.

Limitations

While our approach shows promise in mitigating social bias in language models without relying on demographic labels, it is important to recognize its limitations. First, our method relies on predefined texts that represent different social attribute groups, which may not fully capture the complexity and diversity of these attributes. Language models are complex systems, and they may still exhibit bias or unintended associations despite our efforts.

Moreover, it is important to acknowledge that gender is non-binary, and the expirements we conducted were focused on addressing binary gender biases. Additionally, our analysis of racial biases is centered around the African-American race, using sociolect as a proxy which might be inaccurate. We believe there is a need for more comprehensive research to address biases related to African American rase and other racial and ethnic groups, in a more precise manner.

Ethics Statement

The development and implementation of our method for mitigating bias in language models require careful ethical considerations. By employing the KL loss regularization term with non-uniform probabilities, there is a possibility of inadvertently amplifying biases or introducing unintended consequences. Additionally, while our aim is to mitigate bias without relying on demographic labels, we acknowledge the need for evaluation and validation to minimize any unforeseen biases that may persist. To mitigate these risks, we strongly recommend the collection of a small validation set to assess the performance of the system and ensure its alignment with ethical considerations.

Acknowledgements

This research was supported by the Israel Science Foundation (grant 448/20), an Azrieli Foundation Early Career Faculty Fellowship, and an AI Alignment grant from Open Philanthropy.

References

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075.

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, Austin, Texas. Association for Computational Linguistics.

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

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*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. 2023. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977, Toronto, Canada. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy

Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

OpenAI. 2022. *OpenAI. OpenAI: Introducing ChatGPT, 2022. URL https://openai.com/blog/chatgpt.*

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167.

Hadas Orgad and Yonatan Belinkov. 2023. Debiasing NLP models without demographic information. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, ACL 2023, July 9-14, 2023*. 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, ACL 2020, Online, July 5-10, 2020*, pages 7237–7256. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022a. Linear adversarial concept erasure. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022b. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8281–8291, Online. Association for Computational Linguistics.

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

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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

A Setup

A.1 Training Setup

We conduct experiments using two models: BERT (Devlin et al., 2019) and DeBERTa-V3 (He et al., 2022). For BERT we used the `bert-base-uncased` and for DeBERTa we used the `microsoft/deberta-v3-base`, both from the Huggingface library (Wolf et al., 2020). We utilize the transformer models as a text encoder, where the input text is transformed into a contextualized representation. The [CLS] token of the encoder is then passed through a linear classifier for the downstream task. We used a 65/10/25 training-validation-test split ratio for all tasks. Training was done with a learning rate of $5e-5$ and a stochastic gradient descent optimizer for 1 epoch.

A.1.1 DaFair

To maintain suitable representations in the embedding space, we processed the pre-defined text (Sec 2.1.1) through the text encoder every 200 batches during fine-tuning. This fixed frequency was found to be effective for stable training and did not require further tuning.

λ Tuning To determine the appropriate value for the parameter λ, we adopted the approach outlined by Ganin et al. (2016). The parameter λ was initially set to 0 and gradually adjusted towards a predefined threshold value, denoted as $\lambda_{threshold}$, using a specific schedule. The schedule for updating λ is determined by the following formula:

$$\lambda = \left(\frac{2}{1+\exp(-\gamma \cdot p)} - 1\right) \cdot \lambda_{threshold}, \quad (5)$$

where p represents a measure of progress, and γ controls the rate of change. The parameter γ allows us to control the speed at which λ approaches $\lambda_{threshold}$.

In our experiments, we set γ to a fixed value of 5. This schedule allows for stable training in the early stages of training. In order to optimize the performance of DaFair and Semi-DaFair, we conducted a grid search to find the optimal value for the parameter $\lambda_{threshold}$. The search was performed over the following values $\lambda_{threshold} \in (0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100)$.

A.1.2 Compared Methods

For **BLIND**, we used their implementation and ran a grid search for $\gamma \in (1, 2, 4, 8, 16)$ and $T \in (2, 4)$.

For **JTT**, we provide our own implementation and tune for $\lambda_{up} \in (1, 2, 4, 6, 8, 10)$.

For the **INLP**, **RLACE**, and **IGBP** methods, we utilized the implementations provided by the respective authors. To ensure optimal performance, we conducted hyperparameter tuning specifically for the number of iterations. We applied these post-hoc methods on the model's representations extracted from a fine-tuned model (**Original**). In experiments with a limited number of labeled data for social attributes, we modify the concept removal methods by training the debiasing classifiers on the available labeled data.

A.2 Evaluation Setup

As discussed in Section 3.6, to perform parameter tuning without the need for a validation set with demographic annotations, we adopt Orgad and Belinkov (2023)'s strategy that prioritizes selecting the most radical hyperparameters, while ensuring that the downstream task accuracy remains above 0.97 of the original accuracy. This is done for all methods.

Intuitively, increasing the regularization term weight (in our case: $\lambda_{threshold}$) promotes fairness by encouraging the model to distribute its predictions more evenly among different social groups. However, it can also lead to a decrease in task accuracy if applied excessively. By setting a threshold of 0.97 for the accuracy, we strike a balance between bias mitigation and maintaining competitive performance.

B Fairness Metrics

TPR-GAP We calculate the True Positive Rate (TPR) by:

$$\mathrm{TPR}_{\mathrm{z,y}} = \mathrm{P}(\hat{\mathrm{Y}} = \mathrm{y} | \mathrm{Z} = \mathrm{z}, \mathrm{Y} = \mathrm{y}) \quad (6)$$

where $\hat{Y}$ represents the predicted label, Z denotes the protected attribute, and Y represents the true label.

The TPR Gap is computed as:

$$\mathrm{GAP}^{\mathrm{z,y}}_{\mathrm{TPR}} = \mathrm{TPR}_{\mathrm{z,y}} - \mathrm{TPR}_{\mathrm{z'},\mathrm{y}} \quad (7)$$

where z and z' correspond to different values of the protected attribute.

To assign a single bias measure across all values of y, we calculate the root mean square $\mathrm{GAP}^{\mathrm{z}}_{\mathrm{TPR}}$.

$$\text{GAP}^{\text{z}}_{\text{TPR}} = \sqrt{\frac{1}{|C|}\sum_{y=1}^{N}\left(\text{GAP}^{\text{z,y}}_{\text{TPR}}\right)^2} \quad (8)$$

where C represents the total number of label categories.

Statistical Measures. Another family of fairness metrics involves statistical measures based on probability distributions. We utilize three key metrics:

Independence This metric measures the statistical dependence between the model's prediction and protected attributes. It employs the Kullback–Leibler divergence between two distributions, namely $KL(P(y), P(\hat{y}|z = z))$, for $z \in Z$. The sum over z yields a single value describing the model's independence. It assesses how the model's behavior varies across different demographics.

Separation This metric assesses the statistical dependence between the model's prediction given the target label and the protected attributes. It utilizes $KL(P(\hat{y}|y = y), P(\hat{y}|y = y, z = z))$ for all $y \in Y$, $z \in Z$. This metric is similar to True Positive Rate (TPR) and False Positive Rate (FPR) gaps. It evaluates if the model behaves differently across classes and demographics.

Sufficiency This metric measures the statistical dependence between the target label given the model's prediction and the protected attributes. It employs $KL(P(y|\hat{y} = \hat{y}), P(y|\hat{y} = \hat{y}, z = z))$ for $\hat{y} \in Y$, $z \in Z$. The sum over $\hat{y}$ and z results in a single value. It intuitively assesses whether a model disproportionately promotes or penalizes specific demographic groups.

To measure these statistical fairness metrics, we used the AllenNLP fairness library. (https://github.com/allenai/allennlp).

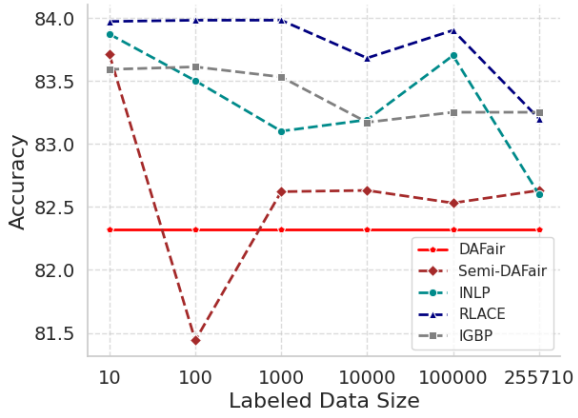
C Full Results

Tables 3 and 4 present comprehensive results for the occupation prediction and sentiment analysis tasks, respectively, employing BERT as the text encoder. Each method's performance is evaluated across multiple metrics, including Accuracy, TPR-GAP, Independence, Separation, and Sufficiency (Section B). Here we also see that our proposed method reduces Independence, Separation, and Sufficiency values, in both tasks.

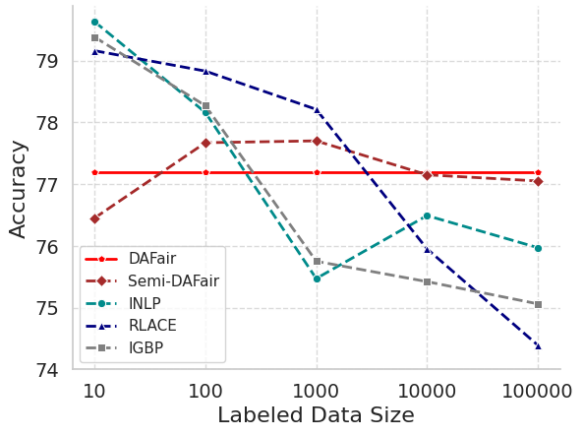
DeBERTA-V3 Results The evaluation results for the occupation prediction and sentiment analysis tasks using the DeBERTa-V3 model are presented in Tables 5, and 6. In the two tasks, Both JTT and BLIND methods demonstrate some success in reducing bias, although not substantial. However, DAFAIR , outperforms both JTT and BLIND in terms of mitigating bias related to social attributes. It achieves a lower TPR-GAP, Independence, Separation and Sufficiency in most cases, while maintaining a comparable level of accuracy. This indicates that our approach is more effective in reducing bias without sacrificing the overall performance of the model also with DeBERTa-V3 model.

C.1 Accuracy Figures

In Figure 3 we provide the accuracy of different methods across varying dataset sizes.



(a) Occupation Prediction



(b) Sentiment Analysis

Figure 3: Effect of bias mitigation methods on accuracy with varying labeled data sizes.

	Occupation Prediction				
Method	Accuracy ↑	TPR-GAP ↓	Indep ↓	Sep ↓	Suff ↓
Original	83.43 ± 0.08	14.66 ± 0.51	0.16 ± 0.002	2.59 ± 0.07	2.46 ± 0.08
JTT	81.34 ± 0.81	14.19 ± 1.08	0.16 ± 0.002	2.77 ± 0.01	2.22 ± 0.14
BLIND	82.52 ± 0.23	13.76 ± 1.18	0.16 ± 0.002	2.35 ± 0.10	2.21 ± 0.09
DAFAIR	82.32 ± 0.13	12.29 ± 0.32	0.14 ± 0.001	1.90 ± 0.09	2.20 ± 0.10

Table 3: Full results for the occupation prediction task with BERT as the text encoder.

	Sentiment Analysis				
Method	Accuracy ↑	TPR-GAP ↓	Indep ↓	Sep ↓	Suff ↓
Original	79.18 ± 0.22	25.34 ± 0.82	0.17 ± 0.003	0.11 ± 0.003	0.08 ± 0.004
JTT	78.08 ± 1.21	23.67 ± 1.87	0.18 ± 0.01	0.11 ± 0.008	0.08 ± 0.006
BLIND	76.45 ± 0.67	22.68 ± 2.40	0.14 ± 0.001	0.05 ± 0.001	0.06 ± 0.001
DAFAIR	77.20 ± 1.17	21.72 ± 0.82	0.15 ± 0.002	0.09 ± 0.002	0.08 ± 0.004

Table 4: Full results for the sentiment analysis task with BERT as the text encoder.

	Occupation Prediction				
Method	Accuracy ↑	TPR-GAP ↓	Indep ↓	Sep ↓	Suff ↓
Original	83.42 ± 0.26	14.60 ± 0.81	0.17 ± 0.003	2.67 ± 0.10	2.51 ± 0.09
JTT	81.98 ± 1.70	14.30 ± 0.55	0.18 ± 0.004	2.65 ± 0.04	2.38 ± 0.12
BLIND	82.41 ± 0.55	13.86 ± 1.49	0.16 ± 0.003	2.31 ± 0.12	2.29 ± 0.10
DAFAIR	82.15 ± 0.32	12.93 ± 0.39	0.14 ± 0.001	2.06 ± 0.09	2.25 ± 0.11

Table 5: Full results for the occupation prediction task with DeBERTa as the text encoder.

	Sentiment Analysis				
Method	Accuracy ↑	TPR-GAP ↓	Indep ↓	Sep ↓	Suff ↓
Original	78.36 ± 0.93	29.62 ± 1.58	0.19 ± 0.003	0.12 ± 0.004	0.09 ± 0.005
JTT	77.41 ± 1.76	28.06 ± 2.03	0.18 ± 0.013	0.13 ± 0.009	0.08 ± 0.005
BLIND	77.20 ± 1.22	27.32 ± 2.84	0.15 ± 0.001	0.07 ± 0.001	0.08 ± 0.001
DAFAIR	77.92 ± 1.21	26.24 ± 1.61	0.16 ± 0.002	0.08 ± 0.003	0.07 ± 0.003

Table 6: Full results for the sentiment analysis task with DeBERTa as the text encoder.

Male Prototypical Texts	Female Prototypical Texts
This is a biography about a male.	This is a biography about a female.
A man who excelled in his field.	A woman who excelled in her field.
He is known for his achievements in various industries.	She is known for her achievements in various industries.
A prominent male figure in history.	A prominent female figure in history.
His career and accomplishments are well-regarded.	Her career and accomplishments are well-regarded.
This biography focuses on the life of a distinguished man.	This biography focuses on the life of a distinguished woman.
An influential male individual.	An influential female individual.
He made significant contributions to his profession.	She made significant contributions to her profession.
This is a story about a man who shaped his industry.	This is a story about a woman who shaped her industry.
His impact on his field is noteworthy.	Her impact on her field is noteworthy.

Table 7: Pre-defined Representations for Male and Female Biographical Texts

AAE Prototypical Texts	SAE Prototypical Texts
This tweet reflects a [sentiment] from a white writer.	This tweet reflects a [sentiment] sentiment from a black writer.
A tweet expressing a [sentiment] moment by a white individual.	A tweet expressing a [sentiment] moment by a black individual.
A [sentiment] viewpoint shared by a writer using Standart American English.	A [sentiment] viewpoint shared by a writer using African American English.
This post, written in standard English, conveys [sentiment] from a white perspective.	This post, written in AAE, conveys [sentiment] from a black perspective.
A message filled with [sentiment] from a white communicator.	A message filled with [sentiment] from a black communicator.
A white person shares their [sentiment] thoughts in this tweet.	A black person shares their [sentiment] thoughts in this tweet.
This is an example of a tweet with [sentiment] in white sociolect.	This is an example of a tweet with [sentiment] sentiment in AAE.
A tweet written by a white speaker that conveys [sentiment].	A tweet written by a black speaker that conveys [sentiment].
This post by a white individual radiates [sentiment] and [sentiment].	This post by a black individual radiates [sentiment] and [sentiment].
A [sentiment] perspective presented by a writer using white sociolect.	A [sentiment] perspective presented by a writer using African American English.

Table 8: Pre-defined Representations for AAE and SAE Tweet Texts

Prototypical Text Pairs Generation Prompt

```
Demographics-Agnostic Fairness method is a novel approach for
mitigating social bias during the fine-tuning process of language
models, without relying on demographic information.
Our approach aims to ensure equal similarity between the
representation of a text and prototypical representations of
different demographic groups. For instance, when classifying a
biographical text of a person into their profession, our method
aims to make the representation of the text equally similar to
the representations of both males and females. More concretely,
our method first define prototypical texts, such as "This is a
biography about a male" and "This is a biography about a female".
It then adds a regularization term that makes the representation
of a training example equally similar to the representations of
each of the prototypical texts.

I want you to generate examples of these prototypical texts.

First Task: Bias-in-Bios: The task involves predicting the
occupation of individuals based on their biographical
information. I need 10 male prototypical texts and 10 female
prototypical texts

Second Task: The task entails classifying the sentiment expressed
in tweets. We utilize the DIAL dataset, a collection of Twitter
messages labeled with positive or negative sentiment. As a proxy
for the writer's racial identity, each tweet is associated with
the sociolect: African American English (AAE) or Mainstream US
English (MUSE; often called Standard American English, SAE) I
need 10 prototypical texts describing posts written by AAE people
and 10 prototypical texts written by MUSE people.
```

Figure 4: The prompt for generating prototypical text pairs.