Inference-Time Selective Debiasing to Enhance Fairness in Text Classification Models

Gleb Kuzmin[2,4] **Neemesh Yadav**[5] **Ivan Smirnov**[3,4]
Timothy Baldwin[1,6] **Artem Shelmanov**[1]
[1]MBZUAI [2]Weakly-Supervised NLP Group [3]RUDN University
[4]Laboratory for Analysis and Controllable Text Generation Technologies RAS
[5]IIIT Delhi [6]The University of Melbourne
kuzmin@airi.net neemesh20529@iiitd.ac.in ivs@isa.ru
{timothy.baldwin, artem.shelmanov}@mbzuai.ac.ae

Abstract

We propose selective debiasing – an inference-time safety mechanism designed to enhance the overall model quality in terms of prediction performance and fairness, especially in scenarios where retraining the model is impractical. The method draws inspiration from selective classification, where at inference time, predictions with low quality, as indicated by their uncertainty scores, are discarded. In our approach, we identify the potentially biased model predictions and, instead of discarding them, we remove bias from these predictions using LEACE – a post-processing debiasing method. To select problematic predictions, we propose a bias quantification approach based on KL divergence, which achieves better results than standard uncertainty quantification methods. Experiments on text classification datasets with encoder-based classification models demonstrate that selective debiasing helps to reduce the performance gap between post-processing methods and debiasing techniques from the at-training and pre-processing categories.[1]

1 Introduction

Fairness is an important safety characteristic of a machine learning (ML) model, representing the model's ability to classify instances without discrimination based on various sensitive attributes, such as race, gender, and age (Blodgett et al., 2020). For the past few years, numerous works have investigated and promoted fairness, and a variety of fairness definitions have been proposed (Blodgett et al., 2020; Han et al., 2022b). One prominent type of fairness is group fairness, also known as the equal opportunity criterion, which reflects the inequality of opportunities across different groups (Han et al., 2022a). The inequality in the model predictions usually comes from inadequate or biased training data, and to address this problem and achieve better fairness, researchers have proposed various debiasing techniques (Li et al., 2018; Han et al., 2021, 2022a; Belrose et al., 2023; Kuzmin et al., 2023). The majority of these techniques assume that one has access to the complete training data and the ability to retrain the model from scratch using some special loss function or reweighting the training instances. However, there are many situations when this assumption does not hold. There is a need for inference-time safety mechanisms that protect users from inadequate model behavior.

Inference-time safety mechanisms are primarily associated with uncertainty quantification (UQ) techniques (Gal and Ghahramani, 2016) and selective classification (Geifman and El-Yaniv, 2017; Xin et al., 2021; Vazhentsev et al., 2022, 2023). Selective classification aims to enhance the reliability of ML-based applications by abstaining from unreliable predictions with high uncertainty. We suggest that the same approach could be applied to increase fairness.

In this work, we propose an inference-time safety mechanism that aims to increase the overall quality of models in terms of prediction performance and fairness in situations when model retraining is prohibitive. We call this approach *selective debiasing*. Instead of rejecting predictions of selected instances as in selective classification, we apply to them inference-time debiasing using post-processing debiasing techniques. To the best of our knowledge, this style of approach is novel to the NLP community.

Our main contributions are as follows:

- We propose selective debiasing, an inference-time safety mechanism that aims to improve both the performance and fairness of model predictions by applying a post-processing debiasing method to only a selected subset of predictions.
- We suggest a scoring criterion that aims to se-

[1]The code is available online at https://github.com/glkuzi/selective-debiasing

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 95–107
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

lect the most unreliable and biased predictions. Experiments demonstrate that this scoring criterion is generally better than UQ techniques in selective debiasing.

2 Background

Debiasing techniques can be categorized into three groups: at-training, pre-processing, and post-processing (Han et al., 2022b).

At-training and pre-processing methods. One of the most popular at-training methods is adversarial training (Adv) (Li et al., 2018). It aims to solve a minimax game between minimizing the loss for the primary task and maximizing the loss for predicting the protected attribute. The diverse adversaries method (DAdv) (Han et al., 2021) extends Adv by using an ensemble of multiple diverse discriminators instead of just one. In the pre-processing category, one of the most remarkable methods is Balanced Training with Equal Opportunity (BTEO) (Han et al., 2022a). It rebalances the dataset to minimize the True Positive Rate (TPR) gap between two protected groups. In the same category, Balanced Training with Joint balance (BTJ) (Lahoti et al., 2020) aims to improve the worst-case performance over all unobserved protected groups by focusing on the computationally identifiable regions of error.

Post-processing methods. There are two well-known approaches to post-processing debiasing: Iterative Null-space Projection (INLP) (Ravfogel et al., 2020) and LEAst-squares Concept Erasure (LEACE) (Belrose et al., 2023).

INLP is an iterative method that involves finding an orthogonal projection of a linear classifier matrix, which is initially learned to predict protected attributes from representations (e.g. hidden states of the standard model). This orthogonal projection is then iteratively used to remove all relevant information from these representations, which was used by the classifier to predict protected attributes.

LEACE is a concept erasure technique that renders representations impervious to the prediction of a specific concept while minimizing changes to the original representations. To construct a transformation matrix, it first whitens the data by equalizing the variance across all directions in the representation space. Next, the data is orthogonally projected onto the subspace that captures correlations between representations and protected attributes. Finally, the data is unwhitened using the same covariance matrix. This resulting transformation matrix is subtracted from the original representations (see the formal definition for LEACE in Appendix A).

At-training and pre-processing methods require retraining the model from scratch and access to the whole training set. They also cannot be selectively applied to a subset of predictions. Post-processing techniques do not involve changes to the model itself, can be trained on a subset of data, and can be applied to predictions selectively. However, their performance is usually worse.

In our work, we propose a method that combines the advantages of both post-processing and at-training / pre-processing methods. While it does not need access to the whole training dataset or retraining the model from scratch, it also has better performance than the standard post-processing techniques.

3 Proposed Method

We propose a selective approach, based on applying debiasing only to predictions with the highest bias score. This section introduces the general concept of selective debiasing and presents the bias quantification method underlying this approach.

Selective debiasing. Selective classification is a widely recognized safety mechanism that safeguards against using unreliable model predictions. In this approach, predictions flagged as unreliable due to high uncertainty scores are handled differently, e.g. they are rejected or are escalated to human operators for further review.

Instead of rejecting instances completely as in selective classification, we apply debiasing to selected predictions. In particular, we identify the potentially most biased instances using a bias quantification method $\mathcal{B}(x_i, p_i)$ and replace the original prediction $p_i = f(x_i)$ with a prediction debiased using a post-processing method d: $\hat{p}_i = d(f(x_i))$:

$$\bar{p_i} = \begin{cases} p_i = f(x_i), \text{ if } \mathcal{B}(x_i, p_i) < h \\ \hat{p}_i = d(f(x_i)), \text{ if } \mathcal{B}(x_i, p_i) \geq h, \end{cases} \quad (1)$$

where h is a predefined threshold selected on a validation set.

We note that the proposed approach is different from the standard post-processing debiasing methods since we change predictions for only some instances. While debiasing all predictions might significantly reduce model performance, modifying only predictions likely to be of low quality or

biased is less risky in terms of worsening outcomes and has the potential to correct errors. Such an approach also allows tuning the accuracy–fairness trade-off for debiasing methods (Han et al., 2022b; Kuzmin et al., 2023).

Bias quantification method. Selective classification is usually based on UQ methods. However, uncertainty on its own does not reflect the presence of bias; it simply highlights potentially erroneous predictions. Figure 1 presents a motivational example. It shows the rejection plots for oracle rejection strategies in selective classification for both accuracy and fairness (see the exact definition of fairness in Appendix E). We can see that the fairness oracle outperforms the UQ oracle in terms of fairness while keeping the same performance in terms of accuracy. These results illustrate that it is possible to improve fairness without penalty to accuracy by changing the order of instances being eliminated, i.e. using a different selection criterion.

Consider a multi-label classification model with classes $c \in C$. To quantify how biased a model prediction is for a given instance, we suggest using the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the originally predicted probability distribution p_i^c and distribution $\hat{p_i}^c$ after debiasing:

$$\mathcal{B}^i_{KL} = \sum_{c \in C} p_i^c \log \left(\frac{p_i^c}{\hat{p_i}^c} \right). \quad (2)$$

KL divergence measures the difference in predictions between the standard and the debiased model. The greater the difference, the more information about the protected attribute is removed from the original representation of the instance. This approach could be used with various post-processing methods. In particular, we suggest using LEACE, but also present results with INLP.

Note that applying a post-processing method to a model is a matter of one or two matrix multiplications. An additional prediction step requires inferring only the last layer of a model, which is very fast. Therefore, the runtime overhead introduced by bias quantification is very small (see Appendix H).

4 Experiments

4.1 Experimental Setup

Datasets. For our experiments, we use two English text classification datasets that, in addition to target variables, provide explicit protected attributes. The first is MOJI (Blodgett et al., 2016), a

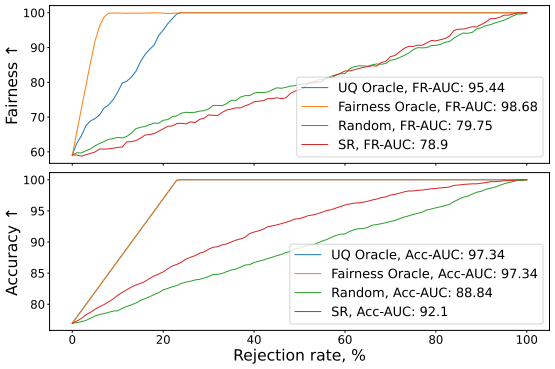

Figure 1: Rejection results for fairness and accuracy with oracle scores on a synthetic dataset with a LogReg model; the FR-AUC and Acc-AUC are the areas under fairness– and accuracy–rejection curves correspondingly. The details are presented in Appendix B.

dataset for sentiment analysis with a binary class ("happy" and "sad") and a binary protected attribute, which corresponds to the author's ethnicity (African American English (AAE) vs. Standard American English (SAE)). The second is a version of the widely used BIOS dataset (De-Arteaga et al., 2019) for occupation classification with a binary gender as the protected attribute. BIOS-2 (Subramanian et al., 2021) is a two-class subsample of the original BIOS dataset with a highly-skewed joint distribution of classes and protected attribute values. As it has been shown to be beneficial to report results for both "balanced" and "imbalanced" versions of datasets (Kuzmin et al., 2023), we conduct experiments on both versions. Detailed information and statistics of the datasets are presented in Appendix C. Due to the limited availability of datasets with annotated protected attributes, most research on debiasing and fairness has been conducted on these few datasets (Han et al., 2022b).

Metrics. We employ several metrics to evaluate the predictive performance and fairness of the model. To evaluate the performance, we use accuracy. For fairness, we consider the widely used equal opportunity criterion (Hardt et al., 2016; Han et al., 2022a,b). We also use two aggregated metrics to evaluate the performance in terms of both accuracy and fairness. The first one is the distance to the optimal point (DTO) (Han et al., 2021):

$$\text{DTO} = \sqrt{(1-\text{Accuracy})^2 + (1-\text{Fairness})^2}. \quad (3)$$

The second one is the Fairness F-score (FF) – a smoothed minimum of accuracy and fairness:

$$\text{FF-score} = \frac{2 \cdot \text{Accuracy} \cdot \text{Fairness}}{\text{Accuracy} + \text{Fairness}}. \quad (4)$$

Debiasing method type		No debiasing	At-training		Pre-processing		Post-processing & Selective					
Dataset	**Metric**	**Standard**	**Adv**	**DAdv**	**BTEO**	**BTJ**	**LEACE-last**	**LEACE-last+SR, opt. perc.**	**LEACE-last+KL, opt. perc.**	**LEACE-cls**	**LEACE-cls+SR, opt. perc.**	**LEACE-cls+KL, opt. perc.**
MOJI imbalanced	Fairness ↑	61.8±0.7	73.7±0.6	73.4±0.4	**75.2±0.6**	74.8±0.6	**75.8±2.6**	68.6±1.4	75.7±0.8	75.2±3.0	68.4±1.1	**77.2±0.7**
	Accuracy ↑	**79.1±0.7**	72.0±0.7	72.4±0.5	73.6±0.6	73.2±0.4	68.3±2.6	**77.6±0.9**	72.7±1.2	66.8±3.0	**77.6±1.0**	71.8±1.2
	DTO ↓	43.6±0.6	38.4±0.5	38.3±0.4	**36.2±0.1**	36.7±0.4	39.9±3.6	38.6±0.7	**36.6±0.5**	41.4±4.1	38.8±0.6	**36.2±1.2**
	FF-score ↑	69.4±0.4	72.8±0.4	72.9±0.3	**74.4±0.1**	**74.0±0.3**	71.8±2.6	72.8±0.5	**74.1±0.3**	70.8±3.0	72.7±0.4	**74.4±0.8**
MOJI balanced	Fairness ↑	69.5±0.2	83.8±0.8	84.7±1.5	85.5±0.5	**85.6±0.6**	79.7±3.9	77.1±0.9	**86.6±0.5**	77.6±4.2	77.0±0.8	**87.5±0.5**
	Accuracy ↑	71.9±0.4	74.0±0.4	74.1±0.6	**74.8±0.3**	74.5±0.4	73.6±0.8	**74.0±0.3**	74.0±0.2	73.0±1.2	**74.0±0.4**	73.7±0.5
	DTO ↓	41.5±0.4	30.7±0.7	30.1±0.7	**29.0±0.1**	29.3±0.4	33.4±3.0	34.7±0.7	**29.3±0.3**	35.2±3.7	34.7±0.6	**29.1±0.6**
	FF-score ↑	70.7±0.3	78.6±0.5	79.1±0.6	**79.8±0.1**	79.6±0.3	76.5±2.2	75.5±0.5	**79.8±0.2**	75.2±2.6	75.5±0.4	**80.0±0.4**
BIOS-2 imbalanced	Fairness ↑	90.4±0.8	**97.2±0.8**	96.4±0.4	95.8±1.0	96.6±0.8	92.8±9.3	93.0±2.3	**94.5**±4.4	77.3±6.5	94.8±2.3	**96.7±0.9**
	Accuracy ↑	**96.7±0.1**	94.8±0.4	95.0±0.3	95.2±0.3	95.0±0.5	60.5±3.6	**94.6±0.2**	92.0±0.4	64.0±5.5	**94.6±0.1**	93.2±0.3
	DTO ↓	10.1±0.7	**5.9±0.2**	6.2±0.2	6.5±0.6	6.1±0.3	41.3±2.1	**9.0±1.7**	10.3±2.8	43.4±2.4	7.7±1.7	**7.6±0.5**
	FF-score ↑	93.5±0.4	**96.0±0.2**	95.7±0.1	95.5±0.4	95.8±0.2	72.8±2.3	**93.8±1.2**	93.2±2.3	69.6±1.7	94.7±1.2	**94.9±0.4**
BIOS-2 balanced	Fairness ↑	89.7±0.6	97.8±0.8	**98.0±0.8**	95.9±0.8	96.4±0.3	90.6±9.8	93.7±2.6	**94.6**±4.2	74.8±9.2	96.6±1.8	**97.5±0.9**
	Accuracy ↑	92.4±0.3	91.9±0.6	91.9±1.5	92.6±0.5	**92.9±0.6**	49.9±9.4	**90.9±1.3**	89.3±1.8	63.8±10.1	**91.9±0.7**	90.6±1.3
	DTO ↓	12.8±0.6	8.5±0.4	8.4±1.4	8.5±0.2	**8.0±0.6**	52.4±6.0	**11.1±2.4**	12.4±3.4	46.0±4.3	**8.9±1.4**	9.7±1.5
	FF-score ↑	91.1±0.4	94.7±0.1	**94.9±0.7**	94.2±0.2	94.6±0.3	63.0±4.6	**92.3±1.9**	91.9±2.9	67.5±3.0	**94.2±1.2**	93.9±1.1

Table 1: Comparison of debiasing methods and selective debiasing. The best results in the group are in bold, and the best results overall are underlined. The results are averaged over 5 random seeds. The gray color corresponds to the results with p-value > 0.05 with respect to standard model.

Details of the equal opportunity fairness calculation are presented in Appendix E.

Models. For the BIOS-2 dataset, we use BERT ("bert-base-cased") (Devlin et al., 2019). For the MOJI dataset, we use the domain-specific BERTweet model (Nguyen et al., 2020) which is good for processing data from social media sources. For both models, we add a three-layer MLP as a classification head, following Han et al. (2022b). Model hyperparameters are described in Appendix D.

Baselines. We compare the proposed selective debiasing approach to inference-time debiasing of all predictions using LEACE and INLP, as well as to at-training and pre-processing debiasing techniques: Adv, DAdv, BTEO, BTJ. We also compare the proposed KL-based bias quantification score with a UQ baseline: Softmax Response (SR: Geifman and El-Yaniv (2017)), calculated as $\mathcal{B}_{SR}(x_i) = 1 - \max_{c \in C} p_i^c$.

Details of debiasing methods. Pre-processing and at-training debiasing methods were applied while training the model from scratch on the full dataset, whereas post-processing methods were trained using only 20% of the data. The optimal threshold for selective debiasing was chosen based on the first 15% of the validation set. "LEACE-last" in our experiments represents LEACE applied to the outputs of the last hidden layer of the classifier, while "LEACE-cls" is LEACE applied to each linear layer of the classification head of the model. The hyperparameters of debiasing methods are provided in Appendix D.

4.2 Results

Table 1 presents results for various at-training and pre-processing debiasing methods, post-processing debiasing methods, selective debiasing based on LEACE with SR, and selective debiasing using the proposed KL-based bias quantification score. Here, we show results only for the threshold that gives an optimal selection percentage. The full results with various selection percentages are presented in Appendix F. The results for selective debiasing using INLP are provided in Appendix F.

In the majority of cases, the best results are unsurprisingly achieved by at-training and pre-processing debiasing techniques, as these methods retrain the models from scratch on the full training data. Nevertheless, the proposed selective debiasing approach based on LEACE substantially enhances the results of inference-time debiasing using post-processing techniques in terms of metrics that take into account both fairness and performance: FF-score and DTO. Inference time debiasing becomes competitive with at-training and pre-processing techniques. For LEACE-cls with KL selection, selective debiasing even outperforms these methods on MOJI-balanced. The results in Tables 15 to 17 also show that selective debiasing consistently outperforms standard inference-time debiasing in terms of FF-score.

LEACE-cls generally achieves better fairness than LEACE-last and slightly better joint fairness–

performance in terms of DTO and FF-score.

When comparing the results of the proposed bias quantification method based on the KL distance with SR, we can see that our method notably outperforms SR on the MOJI datasets and is on par with SR on BIOS-2. We further explore other distance-based bias quantification methods (Euclidean and cosine distances) in Appendix G. Results in Tables 15 to 17 show that in most cases, selection by KL works comparably or better than other distance-based measures. Moreover, KL scores are easier to compute than distance-based scores.

5 Conclusion and Future Work

We proposed selective debiasing – a new simple inference-time safety mechanism for increasing model performance and fairness. We showed that it is helpful in the case when re-training a model from scratch for better fairness is prohibitive or there is no access to full training data. Additionally, for the selection of problematic predictions, we suggest a bias quantification approach based on KL divergence that achieves better results than the standard UQ method. The proposed mechanism fills the gap for efficient techniques that can be applied at inference time and opens the door for safer ML-based systems. In future work, we aim to investigate a deeper integration between UQ and debiasing methods.

Limitations

In this work, we considered only group fairness (equal opportunity criterion), where there exist many other fairness definitions. However, this research is focused particularly on group fairness, and the equal opportunity criterion is the metric of choice in previous work on the same datasets. During all experiments, we assume that we have access to the protected attributes, which is not always the case. But this is a common assumption for any work on debiasing; moreover, it is necessary for the calculation of the fairness metric. Finally, all of the experiments were conducted on the English language, but the used methods are language-independent, so we do not expect significant differences in results for other languages.

Ethical Considerations

In this work, we consider group fairness and instance-level bias quantification. We used only publicly available datasets and models, and only for the intended use. In our research, we used protected attributes to apply debiasing methods and to compute metrics; however, this is necessary for all debiasing methods. To avoid possible harm, we used only attributes that users self-disclosed for the experiments.

Acknowledgments

We appreciate the anonymous reviewers for their valuable suggestions that helped enhance this paper. This research was supported in part through the computational resources of HPC facilities at HSE University.

References

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063. Curran Associates, Inc.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4878–4887.

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

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

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

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Gleb Kuzmin, Artem Vazhentsev, Artem Shelmanov, Xudong Han, Simon Suster, Maxim Panov, Alexander Panchenko, and Timothy Baldwin. 2023. Uncertainty estimation for debiased models: Does fairness hurt reliability? In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–770, Nusa Dua, Bali. Association for Computational Linguistics.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc.

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

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

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

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, Online. Association for Computational Linguistics.

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

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

A LEAst-Squares Concept Erasure

LEACE removes information about a concept Z from the representation space X. To formally describe LEACE, we firstly introduce the following notions. Let $\mathbf{x} \in X$ be an instance from X (e.g. embedding from the last layer in the case of LEACE-last), $\mathbf{\Sigma_{XX}}$ is the covariance matrix for X, $\mathbf{\Sigma_{XZ}}$ is the covariance matrix between X and Z, and $W_\perp$ stands for the pseudoinverse of the matrix W. The $\mathbf{W}$ and $\mathbf{P_{W\Sigma}}_{XZ}$ defined as follows:

$$\mathbf{W} = (\mathbf{\Sigma_{XX}}^{1/2})_\perp, \tag{5}$$

$$\mathbf{P_{W\Sigma}}_{XZ} = (\mathbf{W\Sigma_{XZ}})(\mathbf{W\Sigma_{XZ}})_\perp. \tag{6}$$

Then the final LEACE transformation is defined as follows:

$$\hat{\mathbf{y}}(\mathbf{x}) = \mathbf{x} - \mathbf{W}_\perp \cdot \mathbf{P_{W\Sigma}}_{XZ} \cdot \mathbf{W}(\mathbf{x} - \mathbb{E}[X]) \tag{7}$$

B Fairness and UQ Oracles

In this section, we describe in detail oracle strategies for fairness and accuracy. For both strategies, we assume access to the ground-truth labels, while for fairness oracle we also use protected attributes. Accuracy oracle is built as follows – we find all erroneously classified instances and replace predictions on these instances with ground-truth labels while keeping all other predictions unchanged. This oracle shows the best possible UQ strategy that allows the detection of all erroneous predictions and gives the maximal increase in accuracy. The same idea is behind fairness oracle, but instead of accuracy, we use fairness as a target metric. For fairness, we first replace predictions for instances, which gives the maximal increase in fairness. These predictions are chosen greedily from the erroneous ones. To measure the quality of these oracle strategies and to compare them with other scores, we calculated several metrics: FR-AUC, Acc-AUC, and FF-score-AUC. Each corresponds to the area under the target metric-rejection curve, where the target metric is fairness, accuracy, or FF-score; the area under the curve is calculated on binarized over 100 points target metric values.

C Datasets Statistics

The synthetic dataset was generated as a random 2 classes classification task using `make_classification` function from Scikit-learn library (Pedregosa et al., 2011) with the following parameters: `n_features=10`, `n_informative=5`, `n_clusters_per_class=2`, `random_state=42`, `n_redundant=2`. The protected attribute for the synthetic dataset is designed as a condition over the first informative feature and equals 1 if this feature is greater than 0, and 0 otherwise. The overall statistics for each dataset are presented in Table 2. Tables 3 and 4 shows the joint distribution of the target variable and protected attributes.

Dataset	Num. of classes/attributes	Protected attribute	Train/Val/Test
Synthetic	2/2	Geometric	6k/2k/2k
Moji (balanced)	2/2	Race	100k/8k/8k
Moji (imbalanced)	2/2	Race	100k/5k/5k
Bios-2 (imbalanced)	2/2	Gender	21k/3k/8k
Bios-2 (balanced)	2/2	Gender	21k/1k/2k

Table 2: Dataset statistics.

Split	Gender	Profession		
		Nurse	Surgeon	Total
Train	Female	53.34	5.74	59.08
	Male	5.50	35.42	40.92
	All	58.84	41.16	100.00
Val	Female	53.32	5.08	58.40
	Male	5.52	36.08	41.60
	All	58.83	41.17	100.00
Test	Female	53.82	7.51	61.33
	Male	5.01	33.66	38.67
	All	58.83	41.17	100.00
Val (balanced)	Female	26.02	23.98	50.00
	Male	26.02	23.98	50.00
	All	52.05	47.95	100.00
Test (balanced)	Female	20.02	29.98	50.00
	Male	20.02	29.98	50.00
	All	40.04	59.96	100.00

Table 3: Joint distribution for the BIOS-2 dataset.

Split	Ethnicity	Target		
		Sad	Happy	Total
Train	SA	40.00	10.00	50.00
	AA	10.00	40.00	50.00
	All	50.00	50.00	100.00
Val	SA	40.02	9.98	50.00
	AA	9.98	40.02	50.00
	All	50.00	50.00	100.00
Test	SA	40.02	9.99	50.01
	AA	9.99	40.00	49.99
	All	50.01	49.99	100.00
Val (balanced)	SA	25.00	25.00	50.00
	AA	25.00	25.00	50.00
	All	50.00	50.00	100.00
Test (balanced)	SA	25.01	25.01	50.01
	AA	24.99	24.99	49.99
	All	50.00	50.00	100.00

Table 4: Joint distribution for the MOJI dataset.

D Training Setup and Hyperparameters

To find an optimal set of hyperparameters, we conducted a grid search on the validation set. We used accuracy as an optimization target for standard models, and DTO for models with debiasing. The grid and optimal parameters for the standard models are described in Table 5. For each debiasing method, we tuned the method's parameters and kept the training parameters of the base model – the grid and optimal values for debiasing methods presented in Table 6. The training was conducted on a cluster with Nvidia V100 GPUs. An approximate number of GPU hours spent during the experiments is presented in Table 7.

Dataset	Num. Epochs	Batch Size	Learning Rate	Weight Decay	Dropout Rate
MOJI imbalanced	20	32	1e-6	0	0.1
MOJI balanced	20	32	1e-6	0	0.1
BIOS-2 imbalanced	20	16	1e-6	0	0.1
BIOS-2 balanced	20	32	1e-6	1e-4	0.1

Table 5: Optimal training hyperparameters for BERTweet on MOJI and BERT on BIOS-2 for standard model. We use a grid search with the following grid values: batch size: [16, 32], learning rate: [1e-6, 5e-6, 1e-5, 3e-5, 5e-5], weight decay: [0, 1e-4]. The number of epochs is determined by early-stopping.

Dataset	Debiasing Method	Adv. Lambda	Adv. Diverse Lambda	INLP by Class	INLP Discriminator Reweighting
Moji (imbalanced)	Adv	1.0	-	-	-
	DAdv	1.0	1.0	-	-
	INLP	-	-	False	True
Moji (balanced)	Adv	1.0	-	-	-
	DAdv	1.0	1.0	-	-
	INLP	-	-	False	False
BIOS-2 (imbalanced)	Adv	1.0	-	-	-
	DAdv	1.0	1.0	-	-
	INLP	-	-	False	True
BIOS-2 (balanced)	Adv	1.0	-	-	-
	DAdv	1.0	1.0	-	-
	INLP	-	-	False	True

Table 6: Optimal debiasing hyperparameters for BERTweet on MOJI and BERT on BIOS-2 for various debiasing methods. The base training parameters are the same as for the vanilla model. We use a grid search with the following grid values: Adv. Lambda/Adv. Diverse Lambda: [1e-4, 1e-3, 1e-2, 1e-1, 1, 1e2, 1e3], INLP by Class/INLP Discriminator Reweighting: [False, True]. The remaining parameters for each method used default values from (Han et al., 2022b). For DAdv Adv. Lambda/Adv. Diverse Lambda parameters were tuned jointly, as in (Han et al., 2022b).

Dataset	Model	GPU hours	Num. of Params
Moji	BERTweet	339	135m
Bios-2	BERT	119	110m

Table 7: Overall computation statistics. GPU hours specify the approximate number of GPU hours spent for training and evaluating the corresponding model for all experiments on both imbalanced and balanced sets. The column Num. of Params contains the number of parameters of a single model.

E Equal Opportunity

There are a numerous amount of group fairness definitions; to avoid any mismatches, we are presenting the step-by-step process of equal opportunity criterion calculation. This criterion is based on recall values, or true positive rates (TPR) for each class and protected group.

- TPR (recall) for each protected group defined as follows:

$$TPR = \frac{TP}{TP + FN}, \tag{8}$$

  where TP, FN – is true positives and false negatives for specific group.

- After we calculate TPR-gap:

$$\delta = \sqrt{\frac{1}{C} \sum_{c} \sum_{g} |TPR_{c,g} - \overline{TPR_c}|^2}, \tag{9}$$

  here g is group index, c - class index, $\overline{TPR_c}$ - TPR averaged across all groups for class c.

- Finally, we calculate fairness with the following equation:

$$Fairness = 100 \cdot (1 - \delta). \tag{10}$$

F Additional Experiments

To check how stable the proposed methods are, we compare selective debiasing results over 5%, 10%, and 15% of selection for random, SR, and KL scores. The results are presented in Tables 9 to 11. The optimal percentage selected on the validation set from values from 1% to 15%; results for each dataset-method pair in Tables 12 and 13. In general, optimal scores are better or comparable with results on various percentages, which allows us to use this approach to detect the optimal percentage of selection.

Table 8 shows the performance of selective debiasing and post-processing debiasing methods trained on a full training set. As one can see, the performance on the full set is comparable with the results on only 20% from Table 1.

The results for selective debiasing with INLP trained on 20% of data are presented in Table 14. INLP-based selective debiasing improves the FF-score only on MOJI-balanced, while on other datasets, it is consistent with the base inference-time debiasing method. INLP-based approaches overall fall behind the corresponding LEACE-based techniques.

Debiasing method type		**No debiasing**	**At-training**		**Pre-processing**		**Post-processing & Selective**								
Dataset	**Metric**	**Standard**	**Adv**	**DAdv**	**BTEO**	**BTJ**	**LEACE-last**	**LEACE-last+SR, opt. perc.**	**LEACE-last+KL, opt. perc.**	**LEACE-cls**	**LEACE-cls+SR, opt. perc.**	**LEACE-cls+KL, opt. perc.**	**INLP**	**INLP+SR, opt. perc.**	**INLP+KL, opt. perc.**
MOJI imbalanced	Fairness ↑	61.8±0.7	73.7±0.6	73.4±0.4	**75.2±0.6**	74.8±0.6	75.7±2.6	68.5±1.3	**75.9±1.3**	74.5±2.4	68.4±1.2	**77.0±0.9**	**88.2±6.3**	64.1±1.7	73.2±1.3
	Accuracy ↑	**79.1±0.7**	72.0±0.7	72.4±0.5	73.6±0.6	73.2±0.4	68.3±2.3	**77.7±1.0**	72.2±1.1	66.8±2.5	**77.6±1.0**	71.6±1.1	59.9±7.3	**77.6±1.3**	71.6±1.9
	DTO ↓	43.6±0.6	38.4±0.5	38.3±0.4	**36.2±0.1**	36.7±0.4	40.0±3.4	38.6±0.7	**36.8±0.7**	41.9±3.4	38.8±0.7	**36.6±1.0**	42.6±5.1	42.4±1.3	**39.0±1.8**
	FF-score ↑	69.4±0.4	72.8±0.4	72.9±0.3	**74.4±0.1**	74.0±0.3	71.8±2.4	72.8±0.5	**74.0±0.5**	70.4±2.4	72.7±0.5	**74.2±0.7**	70.8±3.4	70.2±0.9	**72.4±1.3**
MOJI balanced	Fairness ↑	69.5±0.2	83.8±0.8	84.7±1.5	85.5±0.5	**85.6±0.6**	79.7±3.5	77.0±0.9	**86.7±0.6**	77.0±3.4	77.0±0.8	**87.3±0.7**	**77.3±5.6**	70.4±1.3	74.0±2.8
	Accuracy ↑	71.9±0.4	74.0±0.4	74.1±0.6	**74.8±0.3**	74.5±0.4	73.6±0.7	**74.0±0.4**	73.9±0.2	73.0±0.9	**74.0±0.4**	73.6±0.5	65.9±4.6	**71.8±0.4**	69.0±1.8
	DTO ↓	41.5±0.4	30.7±0.7	30.1±0.7	**29.0±0.1**	29.3±0.4	33.4±2.7	34.7±0.7	**29.3±0.4**	35.5±3.0	34.7±0.6	**29.3±0.5**	41.3±4.5	40.9±0.9	40.6±2.3
	FF-score ↑	70.7±0.3	78.6±0.5	79.1±0.6	**79.8±0.1**	79.6±0.3	76.5±2.0	75.5±0.5	**79.8±0.3**	74.9±2.1	75.5±0.4	**79.9±0.4**	71.0±3.4	71.1±0.7	71.4±1.6
BIOS-2 imbalanced	Fairness ↑	90.4±0.8	**97.2±0.8**	96.4±0.4	95.8±1.0	96.6±0.8	**93.3±8.1**	93.1±2.3	92.9±2.1	78.0±5.5	95.2±2.5	**96.4±1.0**	91.6±1.6	**91.9±0.8**	91.5±1.5
	Accuracy ↑	**96.7±0.1**	94.8±0.4	95.0±0.3	95.2±0.3	95.0±0.5	61.1±4.0	94.6±0.3	**94.7±0.2**	65.4±5.6	**94.6±0.1**	93.2±0.3	**95.9±0.8**	**95.9±0.6**	95.9±0.8
	DTO ↓	10.1±0.7	**5.9±0.2**	6.2±0.2	6.5±0.6	6.1±0.3	40.4±2.2	**8.9±1.7**	9.0±1.6	41.7±2.0	**7.4±1.8**	7.7±0.6	9.5±1.1	**9.1±0.4**	9.5±1.1
	FF-score ↑	93.5±0.4	**96.0±0.2**	95.7±0.1	95.5±0.4	95.8±0.2	73.5±2.0	**93.8±1.2**	93.7±1.1	70.7±1.2	**94.9±1.3**	94.8±0.6	93.7±0.6	**93.8±0.2**	93.6±0.6
BIOS-2 balanced	Fairness ↑	89.7±0.6	97.8±0.8	**98.0±0.8**	95.9±0.8	96.4±0.3	91.2±10.2	93.2±2.6	**94.3±3.6**	74.7±9.2	96.7±1.6	**97.6±1.1**	**91.8±1.1**	91.4±0.9	**91.8±1.1**
	Accuracy ↑	92.4±0.3	91.9±0.6	91.9±1.5	92.6±0.5	**92.9±0.6**	50.4±8.8	**90.7±1.2**	90.0±1.8	64.0±9.7	**91.9±0.9**	90.6±1.4	90.7±1.2	**91.1±1.1**	90.7±1.1
	DTO ↓	12.8±0.6	8.5±0.4	8.4±1.4	8.5±0.2	**8.0±0.6**	51.9±5.1	**11.6±2.4**	11.7±3.3	45.8±3.5	**8.8±1.4**	9.7±1.6	12.5±1.2	**12.4±1.1**	12.4±1.1
	FF-score ↑	91.1±0.4	94.7±0.1	**94.9±0.7**	94.2±0.2	94.6±0.3	63.7±3.6	91.9±1.9	**92.1±2.6**	67.7±2.4	**94.2±1.2**	94.0±1.1	91.2±0.9	**91.3±0.8**	91.3±0.8

Table 8: Comparison of debiasing methods and selective debiasing; the post-processing methods trained on full training set. The best results in the group are in bold, and the best results overall are underlined. The gray color corresponds to the results with p-value > 0.05 with respect to standard model.

Dataset	**Standard**	**LEACE**	**Random, 5%**	**SR, 5%**	**KL, 5%**	**Random, 10%**	**SR, 10%**	**KL, 10%**	**Random, 15%**	**SR, 15%**	**KL, 15%**	**Random, optimal percentage**	**SR, optimal percentage**	**KL, optimal percentage**
MOJI imbalanced	69.4±0.4	71.8±2.6	70.6±0.3	70.8±0.5	72.5±0.6	71.4±0.3	71.8±0.6	73.7±0.7	72.0±0.4	72.8±0.5	**74.1±0.3**	72.0±0.4	72.8±0.5	**74.1±0.3**
MOJI balanced	70.7±0.3	76.5±2.2	71.8±0.2	72.4±0.1	75.4±0.2	72.5±0.0	74.0±0.2	78.2±0.3	73.3±0.1	75.5±0.5	**79.8±0.2**	73.3±0.1	75.5±0.5	**79.8±0.2**
BIOS-2 imbalanced	93.5±0.4	72.8±2.3	92.9±0.5	93.7±1.0	93.3±2.0	92.0±0.8	93.7±1.3	90.3±2.1	91.1±1.2	93.1±1.1	86.3±2.2	93.4±0.4	**93.8±1.2**	93.2±2.3
BIOS-2 balanced	91.1±0.4	63.0±4.6	90.6±0.3	91.6±1.0	92.0±2.1	89.3±0.9	92.2±1.9	89.6±2.4	88.7±1.2	**92.6±2.2**	85.4±2.4	90.9±0.3	92.3±1.9	91.9±2.9

Table 9: FF-score of selective debiasing for LEACE on the last layer for various percentages.

Dataset	**Standard**	**LEACE**	**Random, 5%**	**SR, 5%**	**KL, 5%**	**Random, 10%**	**SR, 10%**	**KL, 10%**	**Random, 15%**	**SR, 15%**	**KL, 15%**	**Random, optimal percentage**	**SR, optimal percentage**	**KL, optimal percentage**
MOJI imbalanced	69.4±0.4	70.8±3.0	70.7±0.4	70.8±0.5	72.5±0.8	71.5±0.4	71.7±0.5	74.2±0.6	72.1±0.5	72.7±0.4	**74.4±0.8**	72.1±0.5	72.7±0.4	**74.4±0.8**
MOJI balanced	70.7±0.3	75.2±2.6	71.8±0.3	72.4±0.1	75.5±0.2	72.7±0.2	74.0±0.2	78.7±0.3	73.5±0.1	75.5±0.4	**80.0±0.4**	73.5±0.1	75.5±0.4	**80.0±0.4**
BIOS-2 imbalanced	93.5±0.4	69.6±1.7	93.4±0.7	94.6±1.0	**95.1±0.4**	93.1±1.0	94.7±0.9	91.0±1.0	92.9±1.3	93.2±0.6	85.1±1.1	93.5±0.4	94.7±1.2	94.9±0.4
BIOS-2 balanced	91.1±0.4	67.5±3.0	91.2±0.5	92.4±0.6	93.9±1.1	90.8±1.1	93.7±1.2	90.1±2.3	90.8±1.3	**94.4±0.9**	83.4±2.2	91.2±0.7	94.2±1.2	93.9±1.1

Table 10: FF-score of selective debiasing for LEACE on the classifier level for various percentages.

Dataset	Standard	INLP	Random, 5%	SR, 5%	KL, 5%	Random, 10%	SR, 10%	KL, 10%	Random, 15%	SR, 15%	KL, 15%	Random, optimal percentage	SR, optimal percentage	KL, optimal percentage
MOJI imbalanced	69.4$_{\pm 0.4}$	**71.9**$_{\pm 2.2}$	70.0$_{\pm 0.4}$	69.5$_{\pm 0.3}$	71.0$_{\pm 0.8}$	70.1$_{\pm 0.4}$	69.6$_{\pm 0.5}$	71.8$_{\pm 1.0}$	70.2$_{\pm 0.5}$	70.1$_{\pm 0.5}$	**71.9**$_{\pm 1.3}$	70.2$_{\pm 0.5}$	70.0$_{\pm 0.6}$	**71.9**$_{\pm 1.3}$
MOJI balanced	70.7$_{\pm 0.3}$	71.9$_{\pm 3.2}$	71.0$_{\pm 0.4}$	71.2$_{\pm 0.1}$	72.0$_{\pm 0.6}$	71.2$_{\pm 0.4}$	71.6$_{\pm 0.4}$	72.5$_{\pm 0.9}$	71.3$_{\pm 0.5}$	71.8$_{\pm 0.5}$	**72.8**$_{\pm 1.0}$	71.3$_{\pm 0.5}$	71.8$_{\pm 0.4}$	**72.8**$_{\pm 1.0}$
BIOS-2 imbalanced	93.5$_{\pm 0.4}$	93.8$_{\pm 0.7}$	93.5$_{\pm 0.4}$	**93.9**$_{\pm 0.7}$	93.8$_{\pm 0.8}$	93.5$_{\pm 0.4}$	93.8$_{\pm 0.7}$	93.8$_{\pm 0.8}$	93.6$_{\pm 0.4}$	93.8$_{\pm 0.7}$	93.8$_{\pm 0.8}$	93.6$_{\pm 0.4}$	**93.9**$_{\pm 0.7}$	93.8$_{\pm 0.8}$
BIOS-2 balanced	91.1$_{\pm 0.4}$	91.8$_{\pm 1.1}$	91.1$_{\pm 0.4}$	91.5$_{\pm 0.9}$	91.6$_{\pm 1.1}$	91.1$_{\pm 0.5}$	91.8$_{\pm 1.1}$	91.7$_{\pm 1.1}$	91.2$_{\pm 0.5}$	**91.9**$_{\pm 1.2}$	91.7$_{\pm 1.2}$	91.2$_{\pm 0.5}$	**91.9**$_{\pm 1.2}$	91.6$_{\pm 1.1}$

Table 11: FF-score of selective debiasing for INLP for various percentages.

Dataset	LEACE-last			LEACE-cls			INLP		
	Random	SR	KL	Random	SR	KL	Random	SR	KL
MOJI imbalanced	15	15	14	15	15	15	15	14	15
MOJI balanced	15	15	15	15	15	15	15	13	15
BIOS-2 imbalanced	1	6	6	1	6	4	15	6	14
BIOS-2 balanced	1	11	7	7	12	5	8	15	5

Table 12: Optimal selection percentages for various debiasing methods.

Dataset	LEACE-last			LEACE-cls			INLP		
	Random	SR	KL	Random	SR	KL	Random	SR	KL
MOJI imbalanced	15	15	15	15	15	15	15	15	14
MOJI balanced	15	15	15	15	15	15	15	13	11
BIOS-2 imbalanced	1	6	3	1	6	4	9	12	8
BIOS-2 balanced	1	11	6	7	12	5	8	15	13

Table 13: Optimal selection percentages for various debiasing methods, the post-processing methods trained on full training set.

Debiasing method type		No debiasing	At-training		Pre-processing		Post-processing & Selective		
Dataset	Metric	Standard	Adv	DAdv	BTEO	BTJ	INLP	INLP+SR, opt. perc.	INLP+KL, opt. perc.
MOJI imbalanced	Fairness ↑	61.8$_{\pm 0.7}$	73.7$_{\pm 0.6}$	73.4$_{\pm 0.4}$	**75.2**$_{\pm 0.6}$	74.8$_{\pm 0.6}$	**77.3**$_{\pm 7.3}$	63.5$_{\pm 1.1}$	70.5$_{\pm 2.4}$
	Accuracy ↑	**79.1**$_{\pm 0.7}$	72.0$_{\pm 0.7}$	72.4$_{\pm 0.5}$	73.6$_{\pm 0.6}$	73.2$_{\pm 0.4}$	68.4$_{\pm 6.8}$	**78.0**$_{\pm 0.6}$	73.5$_{\pm 1.4}$
	DTO ↓	43.6$_{\pm 0.6}$	38.4$_{\pm 0.5}$	38.3$_{\pm 0.4}$	**36.2**$_{\pm 0.1}$	36.7$_{\pm 0.4}$	40.0$_{\pm 3.5}$	42.6$_{\pm 0.8}$	**39.7**$_{\pm 1.8}$
	FF-score ↑	69.4$_{\pm 0.4}$	72.8$_{\pm 0.4}$	72.9$_{\pm 0.3}$	**74.4**$_{\pm 0.1}$	**74.0**$_{\pm 0.3}$	**71.9**$_{\pm 2.2}$	70.0$_{\pm 0.6}$	**71.9**$_{\pm 1.3}$
MOJI balanced	Fairness ↑	69.5$_{\pm 0.2}$	83.8$_{\pm 0.8}$	84.7$_{\pm 1.5}$	85.5$_{\pm 0.5}$	**85.6**$_{\pm 0.6}$	**85.8**$_{\pm 8.3}$	71.7$_{\pm 0.6}$	77.9$_{\pm 4.4}$
	Accuracy ↑	71.9$_{\pm 0.4}$	74.0$_{\pm 0.4}$	74.1$_{\pm 0.6}$	**74.8**$_{\pm 0.3}$	74.5$_{\pm 0.4}$	63.0$_{\pm 6.9}$	**71.9**$_{\pm 0.4}$	68.5$_{\pm 2.0}$
	DTO ↓	41.5$_{\pm 0.4}$	30.7$_{\pm 0.7}$	30.1$_{\pm 0.7}$	**29.0**$_{\pm 0.1}$	29.3$_{\pm 0.4}$	40.9$_{\pm 4.4}$	39.9$_{\pm 0.6}$	**38.8**$_{\pm 1.2}$
	FF-score ↑	70.7$_{\pm 0.3}$	78.6$_{\pm 0.5}$	79.1$_{\pm 0.6}$	**79.8**$_{\pm 0.1}$	79.6$_{\pm 0.3}$	71.9$_{\pm 3.2}$	71.8$_{\pm 0.4}$	**72.8**$_{\pm 1.0}$
BIOS-2 imbalanced	Fairness ↑	90.4$_{\pm 0.8}$	**97.2**$_{\pm 0.8}$	96.4$_{\pm 0.4}$	95.8$_{\pm 1.0}$	96.6$_{\pm 0.8}$	**92.0**$_{\pm 1.6}$	91.7$_{\pm 1.4}$	91.9$_{\pm 1.7}$
	Accuracy ↑	**96.7**$_{\pm 0.1}$	94.8$_{\pm 0.4}$	95.0$_{\pm 0.3}$	95.2$_{\pm 0.3}$	95.0$_{\pm 0.5}$	95.8$_{\pm 0.6}$	**96.2**$_{\pm 0.3}$	95.8$_{\pm 0.6}$
	DTO ↓	10.1$_{\pm 0.7}$	**5.9**$_{\pm 0.2}$	6.2$_{\pm 0.2}$	6.5$_{\pm 0.6}$	6.1$_{\pm 0.3}$	**9.1**$_{\pm 1.3}$	**9.1**$_{\pm 1.3}$	9.2$_{\pm 1.4}$
	FF-score ↑	93.5$_{\pm 0.4}$	**96.0**$_{\pm 0.2}$	95.7$_{\pm 0.1}$	95.5$_{\pm 0.4}$	95.8$_{\pm 0.2}$	93.8$_{\pm 0.7}$	**93.9**$_{\pm 0.7}$	93.8$_{\pm 0.8}$
BIOS-2 balanced	Fairness ↑	89.7$_{\pm 0.6}$	97.8$_{\pm 0.8}$	**98.0**$_{\pm 0.8}$	95.9$_{\pm 0.8}$	96.4$_{\pm 0.3}$	**91.8**$_{\pm 2.0}$	91.6$_{\pm 1.9}$	91.3$_{\pm 1.9}$
	Accuracy ↑	92.4$_{\pm 0.3}$	91.9$_{\pm 0.6}$	91.9$_{\pm 1.5}$	92.6$_{\pm 0.5}$	**92.9**$_{\pm 0.6}$	91.9$_{\pm 1.2}$	**92.2**$_{\pm 1.0}$	91.9$_{\pm 1.2}$
	DTO ↓	12.8$_{\pm 0.6}$	8.5$_{\pm 0.4}$	8.4$_{\pm 1.4}$	8.5$_{\pm 0.2}$	**8.0**$_{\pm 0.6}$	11.7$_{\pm 1.7}$	**11.5**$_{\pm 1.7}$	12.0$_{\pm 1.6}$
	FF-score ↑	91.1$_{\pm 0.4}$	94.7$_{\pm 0.1}$	**94.9**$_{\pm 0.7}$	94.2$_{\pm 0.2}$	94.6$_{\pm 0.3}$	91.8$_{\pm 1.1}$	**91.9**$_{\pm 1.2}$	91.6$_{\pm 1.1}$

Table 14: Comparison of debiasing methods and selective debiasing using INLP. The best results in the group are in bold, and the best results overall are underlined. The results averaged over 5 random seeds. The gray color corresponds to the results with p-value > 0.05 with respect to standard model.

G Comparison with other Distances

We also conducted additional experiments to compare how proposed selection strategies differ from other similarity measures. Here, we consider several measures, calculated over the output from the last hidden layer of the model, and compare them with SR and KL strategies. The results are presented in Tables 15 to 17. In most cases, selection by KL works comparably or better than the best-performing distance-based measure. Moreover, KL scores are easier to compute than distance-based scores. However, in some cases, cosine distance could serve as a replacement for the KL score due to its similar performance.

Dataset	Standard	LEACE	SR, 5%	KL, 5%	Euclidean, 5%	Cosine, 5%	SR, 10%	KL, 10%	Euclidean, 10%	Cosine, 10%	SR, 15%	KL, 15%	Euclidean, 15%	Cosine, 15%
MOJI imbalanced	69.4$_{\pm 0.4}$	71.8$_{\pm 2.6}$	70.8$_{\pm 0.5}$	72.5$_{\pm 0.6}$	71.4$_{\pm 0.7}$	72.2$_{\pm 0.6}$	71.8$_{\pm 0.6}$	73.7$_{\pm 0.7}$	72.5$_{\pm 0.8}$	73.8$_{\pm 0.6}$	72.8$_{\pm 0.5}$	74.1$_{\pm 0.3}$	73.0$_{\pm 0.8}$	**74.3$_{\pm 0.8}$**
MOJI balanced	70.7$_{\pm 0.3}$	76.5$_{\pm 2.2}$	72.4$_{\pm 0.1}$	75.4$_{\pm 0.2}$	74.0$_{\pm 0.4}$	75.0$_{\pm 0.2}$	74.0$_{\pm 0.2}$	78.2$_{\pm 0.3}$	76.2$_{\pm 0.4}$	78.1$_{\pm 0.3}$	75.5$_{\pm 0.5}$	**79.8$_{\pm 0.2}$**	77.5$_{\pm 0.4}$	79.2$_{\pm 0.5}$
BIOS-2 imbalanced	93.5$_{\pm 0.4}$	72.8$_{\pm 2.3}$	**93.7$_{\pm 1.0}$**	93.3$_{\pm 2.0}$	93.1$_{\pm 1.5}$	92.0$_{\pm 2.0}$	**93.7$_{\pm 1.3}$**	90.3$_{\pm 2.1}$	90.1$_{\pm 1.0}$	88.7$_{\pm 1.5}$	93.1$_{\pm 1.1}$	86.3$_{\pm 2.2}$	86.6$_{\pm 1.5}$	85.6$_{\pm 1.9}$
BIOS-2 balanced	91.1$_{\pm 0.4}$	63.0$_{\pm 4.6}$	91.6$_{\pm 1.0}$	92.0$_{\pm 2.1}$	90.3$_{\pm 1.9}$	89.5$_{\pm 1.6}$	92.2$_{\pm 1.9}$	89.6$_{\pm 2.4}$	87.9$_{\pm 3.0}$	86.0$_{\pm 1.7}$	**92.6$_{\pm 2.2}$**	85.4$_{\pm 2.4}$	84.4$_{\pm 1.9}$	82.8$_{\pm 2.1}$

Table 15: Comparison of FF-score of distance-based scores for LEACE-last for various percentages.

Dataset	Standard	LEACE	SR, 5%	KL, 5%	Euclidean, 5%	Cosine, 5%	SR, 10%	KL, 10%	Euclidean, 10%	Cosine, 10%	SR, 15%	KL, 15%	Euclidean, 15%	Cosine, 15%
MOJI imbalanced	69.4$_{\pm 0.4}$	70.8$_{\pm 3.0}$	70.8$_{\pm 0.5}$	72.5$_{\pm 0.8}$	71.9$_{\pm 0.5}$	72.2$_{\pm 0.5}$	71.7$_{\pm 0.5}$	74.2$_{\pm 0.6}$	73.3$_{\pm 0.8}$	73.7$_{\pm 1.1}$	72.7$_{\pm 0.4}$	**74.4$_{\pm 0.8}$**	73.9$_{\pm 0.9}$	74.3$_{\pm 1.3}$
MOJI balanced	70.7$_{\pm 0.3}$	75.2$_{\pm 2.6}$	72.4$_{\pm 0.1}$	75.5$_{\pm 0.2}$	74.7$_{\pm 0.3}$	74.4$_{\pm 0.5}$	74.0$_{\pm 0.2}$	78.7$_{\pm 0.3}$	77.2$_{\pm 0.4}$	77.3$_{\pm 1.0}$	75.5$_{\pm 0.4}$	**80.0$_{\pm 0.4}$**	78.7$_{\pm 0.4}$	78.9$_{\pm 1.1}$
BIOS-2 imbalanced	93.5$_{\pm 0.4}$	69.6$_{\pm 1.7}$	94.6$_{\pm 1.0}$	**95.1$_{\pm 0.4}$**	94.3$_{\pm 0.6}$	94.1$_{\pm 0.6}$	94.7$_{\pm 0.9}$	91.0$_{\pm 1.0}$	89.9$_{\pm 1.2}$	90.3$_{\pm 1.2}$	93.2$_{\pm 0.6}$	85.1$_{\pm 1.1}$	83.9$_{\pm 1.0}$	84.4$_{\pm 1.0}$
BIOS-2 balanced	91.1$_{\pm 0.4}$	67.5$_{\pm 3.0}$	92.4$_{\pm 0.6}$	93.9$_{\pm 1.1}$	93.3$_{\pm 1.4}$	92.6$_{\pm 1.1}$	93.7$_{\pm 1.2}$	90.1$_{\pm 2.3}$	87.2$_{\pm 1.8}$	87.0$_{\pm 1.3}$	**94.4$_{\pm 0.9}$**	83.4$_{\pm 2.2}$	80.6$_{\pm 2.0}$	80.9$_{\pm 1.5}$

Table 16: Comparison of FF-score of distance-based scores for LEACE-cls for various percentages.

Dataset	Standard	LEACE	SR, 5%	KL, 5%	Euclidean, 5%	Cosine, 5%	SR, 10%	KL, 10%	Euclidean, 10%	Cosine, 10%	SR, 15%	KL, 15%	Euclidean, 15%	Cosine, 15%
MOJI imbalanced	69.4$_{\pm 0.4}$	**71.9$_{\pm 2.2}$**	69.5$_{\pm 0.3}$	71.0$_{\pm 0.8}$	69.5$_{\pm 0.5}$	69.5$_{\pm 0.5}$	69.6$_{\pm 0.5}$	71.8$_{\pm 1.0}$	69.7$_{\pm 0.7}$	69.6$_{\pm 0.6}$	70.1$_{\pm 0.5}$	**71.9$_{\pm 1.3}$**	69.8$_{\pm 0.9}$	69.7$_{\pm 0.8}$
MOJI balanced	70.7$_{\pm 0.3}$	71.9$_{\pm 3.2}$	71.2$_{\pm 0.1}$	72.0$_{\pm 0.6}$	70.9$_{\pm 0.4}$	71.1$_{\pm 0.5}$	71.6$_{\pm 0.4}$	72.5$_{\pm 0.9}$	71.2$_{\pm 0.5}$	71.3$_{\pm 0.6}$	71.8$_{\pm 0.5}$	**72.8$_{\pm 1.0}$**	71.4$_{\pm 0.6}$	71.5$_{\pm 0.6}$
BIOS-2 imbalanced	93.5$_{\pm 0.4}$	93.8$_{\pm 0.7}$	**93.9$_{\pm 0.7}$**	93.8$_{\pm 0.8}$	93.5$_{\pm 0.4}$	93.5$_{\pm 0.4}$	93.8$_{\pm 0.7}$	93.8$_{\pm 0.8}$	93.5$_{\pm 0.4}$	93.5$_{\pm 0.4}$	93.8$_{\pm 0.7}$	93.8$_{\pm 0.8}$	93.5$_{\pm 0.4}$	93.5$_{\pm 0.4}$
BIOS-2 balanced	91.1$_{\pm 0.4}$	91.8$_{\pm 1.1}$	91.5$_{\pm 0.9}$	91.6$_{\pm 1.1}$	91.1$_{\pm 0.4}$	91.1$_{\pm 0.4}$	91.8$_{\pm 1.1}$	91.7$_{\pm 1.1}$	91.1$_{\pm 0.4}$	91.1$_{\pm 0.4}$	**91.9$_{\pm 1.2}$**	91.7$_{\pm 1.2}$	91.1$_{\pm 0.4}$	91.1$_{\pm 0.4}$

Table 17: Comparison of FF-score of distance-based scores for INLP for various percentages.

H Computational Efficiency

To estimate the computational efficiency of selective debiasing, we calculated the inference time of the standard model and the model with selective debiasing. The results are presented in Tables 18 and 19. Table 18 shows the inference time of models averaged for 10 runs, while Table 19 presents computational overhead for each debiasing method. The computational overhead is calculated as follows:

$$CompOverhead = 100 \cdot \left(\frac{T_{selective}}{T_{standard}} - 1 \right), \tag{11}$$

where T is the summary inference time of the debiasing method for all datasets. These experiments were conducted on one Nvidia H100 GPU. The proposed selective debiasing approach does not introduce much computational overhead – for LEACE-last and LEACE-cls it is less than 1%.

Table 20 shows a detailed comparison of debiasing methods. As one can see, at-training and pre-processing debiasing methods require a training model from scratch, while post-processing methods with selective debiasing do not require this. Hence, post-processing methods are especially beneficial when the full dataset or the model is unavailable, while selective debiasing allows for increasing the overall performance of these methods. On the other hand, there is some computational overhead for post-processing methods compared to other ones. However, this overhead is negligible in most cases.

Dataset	**LEACE-last**		**LEACE-cls**		**INLP**	
	Selective	**Standard**	**Selective**	**Standard**	**Selective**	**Standard**
MOJI imbalanced	3.738±0.011	3.737±0.020	3.762±0.009	3.749±0.035	3.775±0.008	3.730±0.008
MOJI balanced	5.978±0.023	5.96±0.014	6.008±0.014	5.971±0.024	6.053±0.017	5.974±0.016
BIOS-2 imbalanced	6.064±0.018	6.059±0.033	6.090±0.018	6.049±0.013	6.116±0.022	6.051±0.022
BIOS-2 balanced	1.526±0.007	1.525±0.006	1.544±0.024	1.527±0.025	1.542±0.004	1.525±0.004

Table 18: Inference time of standard model and model with applied selective debiasing (in seconds, averaged for 10 runs).

	LEACE-last	**LEACE-cls**	**INLP**
Overhead, %	0.14	0.62	1.19

Table 19: The computational overhead of selective debiasing for various methods.

Debiasing method type	**Base**	**At-training**		**Pre-processing**		**Selective**		
Debiasing method	**Standard**	**Adv**	**DAdv**	**BTEO**	**BTJ**	**LEACE-last selective**	**LEACE-cls selective**	**INLP selective**
Require model retraining from Standard model	×	✓	✓	✓	✓	×	×	×
At-training method	×	✓	✓	×	×	×	×	×
Pre-processing method	×	×	×	✓	✓	×	×	×
Post-processing method	×	×	×	×	×	✓	✓	✓
Inference speed (relative to Standard model)	1.000	1.000	1.000	1.000	1.000	1.001	1.006	1.012

Table 20: Debiasing methods comparison. At-training and pre-processing debiasing methods can have the same inference speed, but require model training from scratch, which is impossible in some cases.