Uncertainty Estimation for Debiased Models: Does Fairness Hurt Reliability?

Gleb Kuzmin[3,5] **Artem Vazhentsev**[2,3] **Artem Shelmanov**[1] **Xudong Han**[1]
Simon Šuster[4] **Maxim Panov**[1,6*] **Alexander Panchenko**[2,3] **Timothy Baldwin**[1,4]
[1]MBZUAI [2]Skoltech [3]AIRI [4]The University of Melbourne [5]FRC CSC RAS [6]TII

{kuzmin, vazhentsev, panchenko}@airi.net simon.suster@unimelb.edu.au
{artem.shelmanov, xudong.han, maxim.panov}@mbzuai.ac.ae tb@ldwin.net

Abstract

When deploying a machine learning model, one should aim not only to optimize performance metrics such as accuracy but also care about model fairness and reliability. Fairness means that the model is prevented from learning spurious correlations between a target variable and socio-economic attributes, and is generally achieved by applying debiasing techniques. Model reliability stems from the ability to determine whether we can trust model predictions for the given data. This can be achieved using uncertainty estimation (UE) methods. Debiasing and UE techniques potentially interfere with each other, raising the question of whether we can achieve both reliability and fairness at the same time. This work aims to answer this question empirically based on an extensive series of experiments combining state-of-the-art UE and debiasing methods, and examining the impact on model performance, fairness, and reliability.[1]

1 Introduction

When deploying a machine learning (ML) model in production, care should be taken to look beyond prediction performance metrics such as accuracy or F1. We believe that modern ML-based applications should be evaluated along two additional critical dimensions: reliability and fairness.

A reliable system should not only perform well on average but also be capable of identifying situations when it is unable to make accurate predictions. By incorporating mechanisms to detect such cases, we can implement appropriate fallback mechanisms such as involving human operators or more advanced models for final decision-making (El-Yaniv et al., 2010; Geifman and El-Yaniv, 2017). This is especially crucial in safety-critical domains like medicine, where the cost of mistakes is high, or in high-load applications where it is impossible to rely solely on automatic decisions (e.g. user content moderation). The better these mechanisms work, the better the model is in terms of *reliability*. In a broader context, a reliable model is characterized by its consistent performance across decision-making tasks involving the uncertainty of predictions (Tran et al., 2022b). In this work, we consider two tasks: selective classification (Geifman and El-Yaniv, 2017), i.e. the ability to abstain from potentially erroneous predictions; and out of distribution (OOD) detection (Hendrycks and Gimpel, 2017), i.e. recognizing instances that are different from the domain of the training set.

Fairness is another critical dimension that needs careful consideration. ML models often exhibit biases due to artifacts in training data or pre-trained language models that can cause, for example, unfair decisions correlated with race, gender, and other demographic and socio-economic factors (Díaz et al., 2019; Park et al., 2018; Badjatiya et al., 2019). It is important to address these biases as they can lead to inequity in user experience, perpetuate stereotypes, and cause other forms of representational harm to users (Blodgett et al., 2020).

While standard prediction performance metrics indicate how well a model is performing in general, they do not capture how a model may behave inconsistently across different conditions or protected groups (Dwork et al., 2012). Therefore, both reliability and fairness are not captured by standard prediction performance metrics, and generally cannot be achieved without deliberate effort.

Enhancing reliability can be accomplished through the use of advanced uncertainty estimation (UE) techniques (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016; Lee et al., 2018; Liu et al., 2020; Podolskiy et al., 2021; Xin et al., 2021; Yoo et al., 2022). Promoting model fairness entails defining fairness metrics and employing special debiasing techniques (Elazar and Goldberg, 2018;

[*]Research was conducted while working at TII.

[1]The code is available online at https://github.com/mbzuai-nlp/fairlib_uncertainty

Wang et al., 2019; Ravfogel et al., 2020; Han et al., 2021, 2022a,d; Baldini et al., 2022).

Just as debiasing methods have been observed to make models more vulnerable to adversarial attacks (Xu et al., 2021; Tran et al., 2022a), they have the potential to impact model reliability in terms of selective classification and OOD detection performance. While the majority of research has focused on the trade-off between model fairness and performance (Liang et al., 2021; Han et al., 2022b,d), no work has investigated the trade-off between fairness and reliability. As such, the main research questions of this work are: (a) whether there is interference between debiasing and UE techniques; (b) whether it is possible to achieve fairness and model reliability simultaneously; and (c) what combinations of techniques lead to the best trade-offs. We address these questions by conducting a large-scale empirical investigation that combines state-of-the-art UE and debiasing methods. For evaluation, we employ text classification datasets and various transformer-based models.

Our main findings are as follows: (a) debiasing can have a more substantial negative impact on selective classification performance than accuracy; (b) the results depend on the distribution of target classes and protected attributes in the test set; (c) rejecting predictions in selective classification can impact relative fairness between different debiasing methods; and (d) OOD detection is also vulnerable to debiasing, but can be mitigated with proper UE techniques. On the basis of our experiments, we suggest best-practice approaches to achieve a good trade-off between model reliability and fairness.

2 Background

2.1 Debiasing and Fairness

Consider we have a labeled dataset $D = (x_i, y_i, g_i)_{i=1}^{n}$, where $x_i \in X$ is an input text, $y_i \in Y$ is a label of a target variable (e.g., sentiment), and $g_i \in G$ is a private attribute associated with x_i (e.g., author gender). In the context of debiasing, our objective is to train a model using the dataset D that not only achieves high accuracy in predicting Y but also exhibits fairness. Specifically, we aim to minimize the disparity in true positive rates (TPR) across different protected attributes, which is known as equal opportunity fairness (Hardt et al., 2016): $\text{GAP}^{\text{TPR}} = |\text{TPR}_g - \text{TPR}_{\neg g}|$, where TPR_g and $\text{TPR}_{\neg\text{g}}$ designate TPR within protected groups g and $\neg g$.

2.2 Debiasing Methods

For our experiments, we selected state-of-the-art techniques from extrinsic debiasing methods available in the Fairlib library (Han et al., 2022d): "preprocessing", "at-training", and "post-processing". For quick reference, all the methods and their corresponding acronyms are summarized in Table 1 in Appendix A.

Pre-processing methods adjust the training set to be balanced across protected groups via resampling or reweighting instances. **Balanced Training with Joint balance** (**BTJ**; Lahoti et al. (2020)) reweights training instances to balance the joint distribution of protected attributes and target labels. A similar method, **Balanced Training with Equal Opportunity** (**BTEO**; Han et al. (2022a)), balances the protected attributes within "advantage" classes through resampling instances based on equal opportunity objectives. BTEO and BTJ are equivalent when the target distribution is inherently balanced. However, when it is not, the approach adopted in BTEO helps to mitigate the susceptibility of BTJ to small-sized minority groups.

At-training methods modify the training procedure or objective. **Adversarial Training** (**Adv**; Elazar and Goldberg (2018)) extends the training objective with a discriminator component responsible for making the model unlearn the protected attributes. The **Diverse Adversaries approach** (**DAdv**; Han et al. (2021)) strengthens Adv by adding an ensemble of adversaries to the loss and subjects them to a diversity constraint for learning orthogonal hidden representations from one another. This approach improves the stability of the training and reduces bias compared to the Adv. **Group Difference** (**GD**; Shen et al. (2022)) adds a loss component that minimizes the loss gap between different groups. We use the variant of this method GD_{diff} that minimizes the differences across protected groups within each class. **Fair Batch Selection** (**FairBatch**; Roh et al. (2021)) dynamically adjusts the probability of resampling instances in each minibatch during training to achieve loss disparity across protected groups.

A post-processing method, **Iterative Null-space Projection** (**INLP**; Ravfogel et al. (2020)), removes protected information from an already trained model by iteratively projecting its hidden representations to the null-space of protected attribute discriminators. The purified representations are subsequently employed for classification.

2.3 Uncertainty Estimation and Reliability

Model reliability refers to its capacity to perform well across a wide range of uncertainty-related tasks (Tran et al., 2022b), such as selective classification, OOD detection, and adversarial attack detection. Uncertainty is a score that quantifies the amount of our trust in a model prediction on a given instance and is intended to correlate with the chance of making a mistake. Estimated uncertainty scores are commonly used as a decision rule in the aforementioned tasks. For example, in selective classification, instances x with a high uncertainty score $u(x)$ are rejected or replaced with predictions of human experts or more advanced systems. Similarly, uncertainty exceeding a given threshold $u(x) > u_{\text{ood}}$ indicates a high likelihood that the instance is OOD. In information theory, uncertainty has a concrete manifestation as the entropy of some distribution (e.g., a predictive distribution $u(x) = \mathcal{H}[p(y|x)]$). However, in a general sense, any score that demonstrates commendable performance in the aforementioned tasks can be considered as a measure of uncertainty. It is common to distinguish two types of uncertainty that arise from two different sources: (a) aleatoric uncertainty arises from irreducible noise in data and inherent ambiguity in tasks that persists even with perfect knowledge and modeling techniques; and (b) epistemic uncertainty reflects the lack of knowledge about optimal model parameters, and can be mitigated by gathering more training data. Their sum (total uncertainty) is commonly used as an indicator of mistakes in selective classification; epistemic uncertainty is also crucial for OOD detection.

2.4 Uncertainty Estimation Methods

We experiment with various widely used UE methods that capture different types of uncertainty: aleatoric, epistemic, and total uncertainty. As a baseline, we use **Softmax Response** (**SR**; Cordella et al. (1995); Geifman and El-Yaniv (2017)), which simply uses maximum probability from the softmax layer as a confidence score.

A widely-used computationally intensive approach to UE is based on **Monte-Carlo dropout** (**MC**; Gal and Ghahramani (2016)). In this work, we use the following UE scores: Bayesian Active Learning with Disagreement (BALD; Houlsby et al. (2011)) and Sampled Maximum Probability (SMP; Gal et al. (2017)). BALD captures epistemic uncertainty, while SMP captures the total uncertainty.

Methods based on the modeling probability density of hidden instance representations are computationally efficient alternatives that have been shown to be effective for epistemic UE. One robust method of this type is **Mahalanobis Distance** (**MD**; Lee et al. (2018); Podolskiy et al. (2021)), which is based on estimating the minimal class-conditional probability of an input instance x that follows a Gaussian distribution. The uncertainty score is computed as the Mahalanobis distance between latent instance representations of x and the closest centroid of a class.

To measure aleatoric uncertainty, we use **Deep Fool** (Ducoffe and Precioso, 2018), whereby we compute the l_2 norm of the minimum perturbation vector that is required to apply to a latent representation to change the prediction of a model. The smaller the norm, the higher the uncertainty.

In addition, we combine MD and DeepFool into a single score, which we call **Hybrid Uncertainty Quantification** (**HUQ**; Vazhentsev et al. (2023)). Depending on whether the instance lies close to the out-of-distribution area of the feature space, or around the discriminative border between classes, we use different types of uncertainty. Details of the Hybrid UQ method are presented in Appendix H.

For quick reference, all the methods and their corresponding acronyms are summarized in Table 1 in Appendix A.

3 Experimental Setup

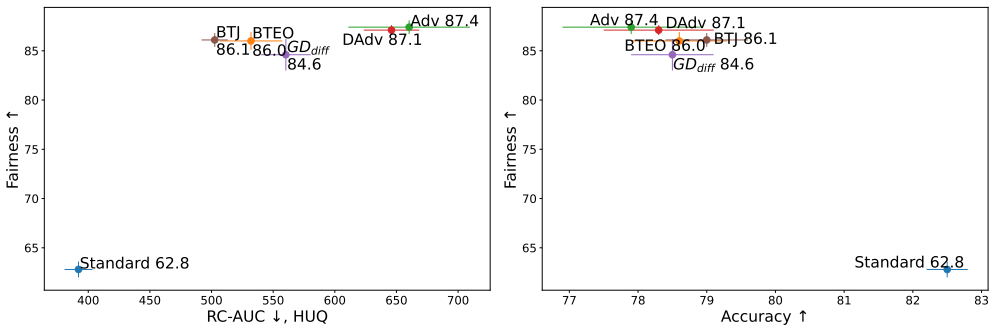
We evaluate the performance of UE techniques over the tasks of selective classification and OOD detection and compare standard models with models where we have applied a debiasing method.

3.1 Datasets

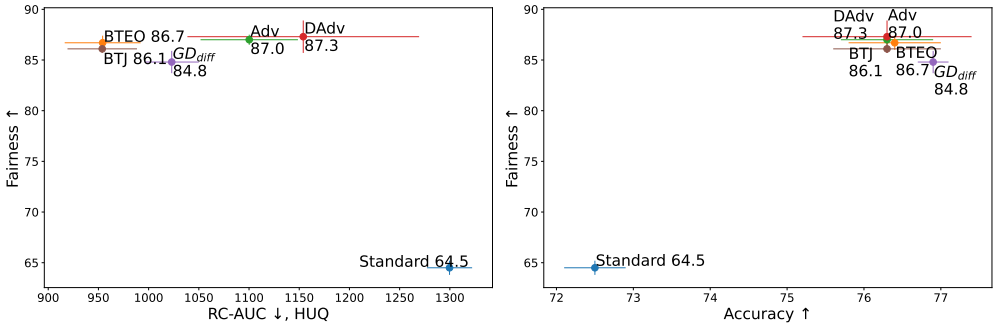
We experiment with two text classification datasets widely used for the evaluation of debiasing techniques in previous work: Moji (Blodgett et al., 2016) and Bios (De-Arteaga et al., 2019).

The Moji dataset is a collection of English tweets paired with a binary protected attribute that represents the ethnicity of the tweet author. It captures the usage of English in two registers: Standard American English (SAE) and African American English (AAE). The target variable is a binary sentiment of tweets (HAPPY and SAD).

Bios comprises biographies annotated with 28 profession classes as the target variable. Due to the extreme scarcity of some classes in the dataset, we



(a) Moji with the imbalanced test and validation sets.



(b) Moji with the balanced test and validation sets.

Figure 1: Trade-off between RC-AUC of selective classification with HUQ and fairness (left) and between accuracy and fairness (right) on Moji (the BERTweet model). We removed results for INLP and FairBatch from this figure due to extremely high RC-AUC values for these debiasing methods. The fairness scores are presented alongside each method for better comparison.

have chosen to use a subsample that focuses on the nine most prevalent classes. The protected attribute is the binary gender.

As debiasing aims to remove the discrepancy between the distribution learned by the model from the training set and a "desired" distribution that is pure from the influence of stereotypes and prejudices in the data, we consider it important to report evaluation results for two versions of the datasets. In the first version, which we call "imbalanced", the distributions $p(g|y)$ are the same in train, validation, and test sets. This reflects the common ML setting, where the test data is similar to training data and inherits all biases present in it. In the second version, which we call "balanced", the distribution $p(g|y)$ in the test and validation sets is balanced, which means there is no preferable protected attribute within each class. In both cases, the training distribution is not changed. The statistics of the datasets and theoretical motivation behind the various test distributions are presented in Appendix C.

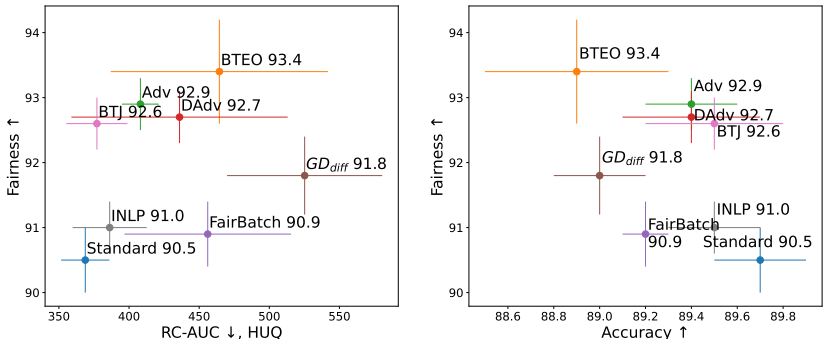
3.2 Models

For experiments, we use three models that were employed in previous work on fairness and demonstrated strong performance on the respective tasks: pre-trained BERT model ("bert-base-cased"; Devlin et al. (2019)) for Bios, BERTweet (Nguyen et al., 2020) for Moji, and a frozen DeepMoji encoder (Felbo et al., 2017) with a three-layer perceptron (MLP) as a classification head from Shen et al. (2022) also for Moji. All parameters of BERT and BERTweet are fine-tuned on the training sets, while for DeepMoji+MLP, we fine-tune only the MLP head. The hyperparameter optimization process is discussed in detail in Section 3.4.

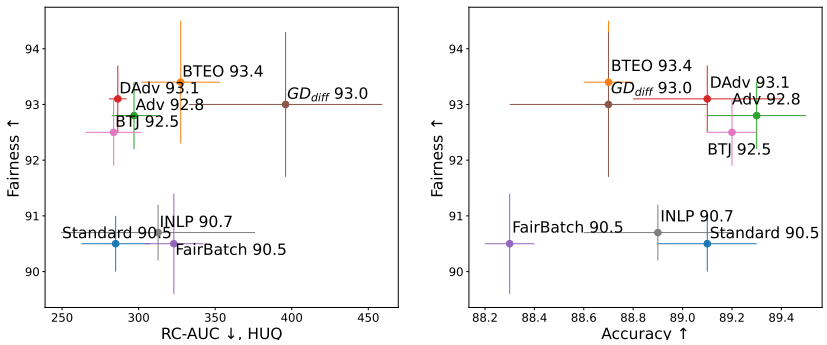
3.3 Metrics

The models are evaluated according to their performance on the classification task via accuracy, a gap-based fairness metric, and the quality of UE.

Debiasing. The quality of debiasing methods is evaluated according to the equal opportunity fairness metric. It measures a lack of disparity in true positive rate across groups formed by the protected attribute (Han et al., 2023). Since the details of this



(a) Bios with the imbalanced test and validation sets.



(b) Bios with the balanced test and validation sets.

Figure 2: Trade-off between RC-AUC of selective classification with HUQ and fairness (left), and between accuracy and fairness (right) on Bios (the BERT model). The fairness scores are presented alongside each method for better comparison.

metric vary in the literature, we provide a step-by-step algorithm for its calculation in Appendix D.

Uncertainty Estimation. UE methods are evaluated on selective classification and OOD detection tasks. In selective classification, we test the ability to detect and reject model mistakes using uncertainty scores as predictors. The standard metric for this task is RC-AUC (El-Yaniv et al., 2010) – the area under the risk–coverage curve, where the coverage is the percentage of retained instances with the lowest uncertainty, and the risk is the average loss over these instances; lower is better. Following Xin et al. (2021), we use a binary loss for calculating the risk. And example of a risk–coverage curve is presented in Figure 6 in Appendix E.

To evaluate the quality of OOD detection, we mix the test set of the target dataset (Moji or Bios) with a series of datasets considered to be OOD. Then, we calculate the ROC-AUC metric, considering the uncertainty score as a predictor of an instance from an OOD dataset. We obtain ROC-AUC for each OOD dataset and average metric values across them all. This is a standard approach adopted in the UE literature (Hendrycks et al., 2019; Hu and Khan, 2021; Zhou et al., 2021). We cannot use the CLINC (Larson et al., 2019) and ROSTD (Schuster et al., 2019) datasets, which are other commonly used benchmarks for OOD detection since these datasets do not provide annotation of protected attributes. The details of the datasets used to represent an OOD domain are described in Appendix C.3.

3.4 Hyperparameter Optimization

We split the hyperparameter selection into two steps. In the first step, we select hyperparameters for training models without debiasing by optimizing the model accuracy. We tune learning rate, batch size, and weight decay via grid search (see the grid and the optimal values in Appendix B).

In the second step, we select hyperparameters of debiasing methods on a fixed grid. We optimize a multi-criteria objective Distance To the Optimum (DTO; Marler and Arora (2004); Han et al. (2022a)). The optimum is a utopia point assumed to be a model that achieves 100% performance in terms of accuracy and fairness:

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

To mitigate the problem with different absolute values of performance and fairness metrics achievable

for the considered task, we use a balanced version of DTO, where evaluation scores are normalized by their maximum values in the set of experiments (e.g. for checkpoints from different epochs, in the case of training or for models with different hyperparameters, in the case of hyperparameter optimization).

Having optimized the hyperparameters, we conduct experiments with five random seeds to report mean and confidence intervals. The hyperparameter grid for each debiasing method and computational resources involved in experiments are presented in Appendix B.

4 Results

4.1 Selective Classification

First of all, consider the selective classification performance (RC-AUC) individually without relation to fairness. Complete results for Bios are presented in Tables 13 and 14, the results for Moji are presented in Tables 15 to 18 in Appendix F. TPR values for each individual class and protected attribute are presented in Tables 21 to 26 in Appendix I.

On Bios, the SR baseline is substantially outperformed by DeepFool, Monte-Carlo dropout, and HUQ, while density-based methods MD and DDU usually do not provide any improvements. The poor performance of the latter methods might be due to the Bios test set not having a marked covariate shift, and subsequently, not containing many OOD instances that could be spotted by density-based methods. The hybrid method, which mixes multiple uncertainty scores, most often outperforms other UE methods for both versions of the dataset. For example, on Bios with an imbalanced test set, HUQ outperforms other UE methods for all debiasing techniques except DAdv and FairBatch, where it also has substantial improvements over the SR baseline. For the standard model and Adv, it improves RC-AUC by more than 14% compared to the SR baseline, for GD_{diff} and INLP, by around 30%, and for BTEO and BTJ, by around 10%.

On Moji, none of the UE techniques are able to outperform the SR baseline, except in the case of INLP, where the baseline has a very high RC-AUC. In this case, density-based methods and HUQ substantially improve the result, though it is much worse than other methods.

Since HUQ often achieves the best results for selective classification, we use it to perform further analysis of debiasing techniques.

Results on the Imbalanced Test Sets. In Figures 1a and 2a, we present the trade-off between RC-AUC and fairness and between accuracy and fairness for models debiased using various methods. From these figures, we can see that on both Moji and Bios, higher fairness results in worse selective classification performance over the imbalanced test sets. Comparing the results for RC-AUC and accuracy, we see that in some cases the malignant increase in RC-AUC is much more substantial than the loss in accuracy. For example, while accuracy for the BTEO and GD_{diff} methods on Bios is reduced by only 0.9% and 1.5% in relative terms, RC-AUC increases by more than 25% and 34%, respectively.

On both Moji and Bios, the best trade-off between fairness and reliability is achieved by BTJ. On Moji, it gives the smallest increase in RC-AUC, while giving a boost in fairness comparable with other methods. On Bios, this method does not increase RC-AUC at all, while also giving a substantial improvement in fairness. Considering the results on Bios, it is also worth noting that Adv, DAdv, and BTEO also achieve a good trade-off: while they worsen RC-AUC, they also lead to a big improvement in fairness. INLP affects both fairness and RC-AUC only slightly.

Results on the Balanced Test Sets. Figures 1b and 2b present accuracy and RC-AUC obtained on the balanced test sets. Since in these test sets, the protected attribute ratios for each of the target classes are balanced, the results are very different from the previous case. In this setting, we can see that on Moji, debiasing positively affects both the accuracy and the selective classification performance (Figure 1b and Figure 8 in Appendix F). All debiasing methods while improving fairness also substantially improve accuracy and RC-AUC, with BTEO and BTJ offering the best trade-offs. These results are strictly opposite to the results obtained on the imbalanced test sets (Figure 1a and Figure 7 in Appendix F). This phenomenon could be attributed to the fact that, during the process of removing bias accumulated from the training set, the modeled probability is adjusted to align more closely with the "desired" distribution. Therefore, testing on the dataset that is closer to this "desired" distribution also demonstrates the improvement in model performance due to debiasing (see theoretical justification in Appendix C.1).

On Bios, debiasing still does not help obtain no-

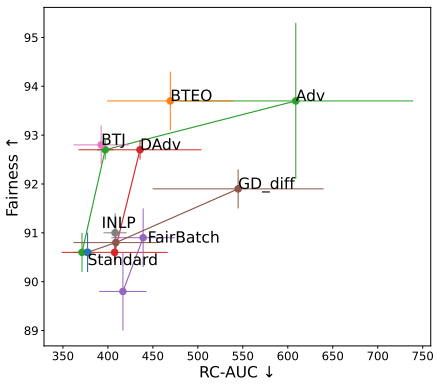


Figure 3: Effect of varying hyperparameters for debiasing methods on Bios with the imbalanced test and validation sets (BERT model).

table improvements in RC-AUC or accuracy. However, in this setting, some debiasing methods do not diminish the performance. While we saw performance degradation for Adv, DAdv, and BTEO in the setting with an imbalanced test set, in this case, there is no gap of note. Overall, for Bios, the best trade-off is achieved by BTJ, DAdv, Adv, and BTEO as they lie in the upper part of the chart, providing high fairness with little or no degradation of RC-AUC.

Effect of Various Hyperparameters in Debiasing Methods. Figure 3 presents the results with varying hyperparameters of the debiasing methods. The hyperparameters related to model training are still optimal. We keep only Pareto optimal points in the chart: points where both fairness and RC-AUC deteriorate are not shown. The standard model and models debiased with BTEO and BTJ have only one point since they do not have variable hyperparameters. The presented results show that for some methods like Adv, DAdv, GD_{diff} it is possible to further improve fairness in exchange for the heavily deteriorated RC-AUC.

How Does Rejecting Predictions Affect Fairness? Figures 4a and 4b depict the dependence of fairness on the rejection rate, in presenting the percentage of predictions that were replaced by the ground-truth. This setting could be considered as an emulation of a human–machine system, where most uncertain instances are processed by humans. As expected, with a greater rejection rate, fairness increases, because the overall performance improves and the average performance gap reduces. However, we note that these charts reveal a discrepancy between the results of different debiasing methods on different rejection rates. On Bios, we see that GD_{diff} in the setting without rejection (0% rejection rate) demonstrates a similar level of fairness with other methods, but starting from 15%, it falls well behind them and the standard model. BTJ on the contrary, demonstrates slightly inferior fairness at the beginning of the curve and starts to outperform other debiasing methods after the rejection rate exceeds 20%. Together with BTEO, it achieves higher performance than other methods on Moji along almost the whole range of rejection rates.

4.2 Out-of-Distribution Detection

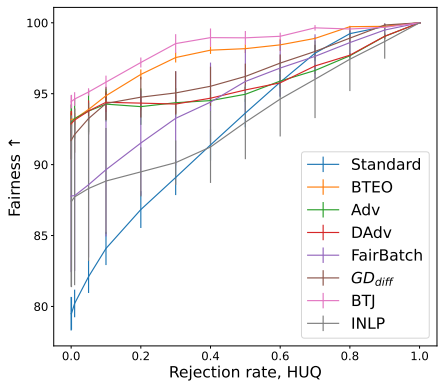
The detailed results for OOD detection on both datasets are presented in Tables 19 and 20 in Appendix G. As expected, DDU substantially improves the OOD detection performance compared to SR in most cases. Figures 5a and 5b demonstrate how the OOD detection performance changes after applying debiasing techniques compared to the standard model. We see that some debiasing methods have a strong negative impact on OOD detection performance. For the SR baseline, applying any debiasing technique on Moji results in substantial performance losses, with the biggest drops of around 15% points for INLP and FairBatch despite the increase in accuracy (Figure 1b). On Bios, we see a large decrease in OOD detection performance for INLP and GD_{diff}.

At the same time, we see that the more advanced DDU method is much less vulnerable to debiasing. On Moji, the performance drop can be seen only for methods that modify the training loss function: Adv, DAdv, and GD_{diff}. Note also that for Adv and DAdv the drop is slightly smaller than the drop for SR. On Bios, a small drop can be seen for FairBatch, while for other methods there is no performance drop at all. Moreover, for BTJ, DAdv, and BTEO there is a small improvement.

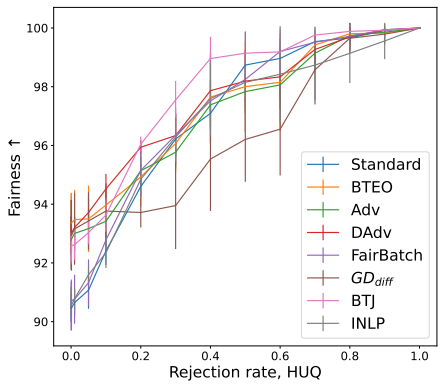
Overall, we can see that the combinations of BTJ, BTEO, and INLP with DDU perform consistently well on OOD detection across both datasets, without any deterioration due to debiasing.

5 Related Work

Great interest in the problem of model fairness within the NLP community in recent years has spurred the development of numerous debiasing techniques (Han et al., 2022a; Elazar and Goldberg, 2018; Shen et al., 2022; Ravfogel et al., 2020). In

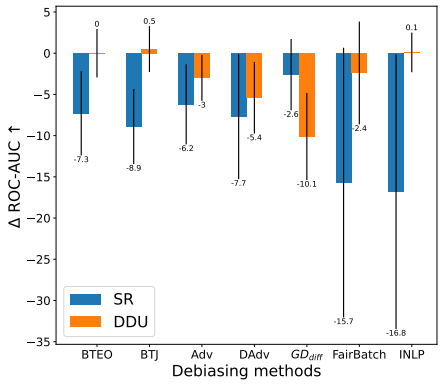


(a) Moji (BERTweet model).

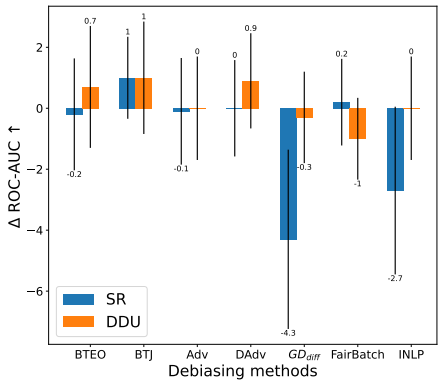


(b) Bios (BERT model).

Figure 4: Dependence of fairness from a rejection rate with HUQ on the balanced test and validation sets.



(a) Moji (BERTweet model).



(b) Bios (BERT model).

Figure 5: The ROC-AUC difference between debiased and standard models for OOD detection on various datasets with imbalanced test and validation sets.

our study, we conduct experiments with state-of-the-art methods selected across the methodological spectrum. There is also ongoing research related to analyzing debiasing methods under various conditions, including different synthetic distributions of protected attributes in the training set (Han et al., 2022c). The approach used in this paper differs from previous work since it analyzes performance, fairness, and reliability over various distributions of protected attributes in the test set, which has not been done before.

NLP models have been increasingly deployed in safety-critical applications in healthcare, finance, and legal domains. This has led to a notable research interest in UE. The most notable UE techniques investigated in NLP are Monte Carlo dropout (Malinin and Gales, 2021), training loss regularization (Xin et al., 2021; Zhang et al., 2019), and density-based methods (Liu et al., 2020; Mukhoti et al., 2023; Podolskiy et al., 2021; Yoo et al., 2022). We conduct experiments with recent widely-used UE methods and also with a prominent hybrid technique that mixes multiple uncertainty scores.

The interference between debiasing and UE techniques recently has been noted in adversarial attack detection. Tran et al. (2022a) show that debiased models are more vulnerable to attacks because debiasing reduces the distance to the classifier decision boundary. Xu et al. (2021) also find that promoting robustness using adversarial training tends to introduce disparity of performance between different protected groups. However, to our knowledge, debiasing in conjunction with reliability tasks such as selective classification and OOD detection has not been investigated before.

6 Further Discussion and Conclusion

In this work, we have investigated the influence of debiasing techniques on model reliability. We discovered that debiasing can substantially reduce the quality of selective classification, particularly

when the test set is imbalanced, i.e. it has a similar biased distribution $p(g|y)$ to the training set. The worsening of selective classification performance is more pronounced than in the case of accuracy. Furthermore, we demonstrate that the decrease in performance due to debiasing can be eliminated (results on Bios) or even turned into improvements (results on Moji) when evaluation is conducted on the balanced test sets. The discrepancy in results demonstrates the importance of conducting an evaluation of debiasing methods not only on various training distributions but also on balanced and imbalanced test distributions. Our experiments reveal that the best trade-off between fairness and selective classification performance is achieved by methods based on instance reweighting: BTJ (Lahoti et al., 2020) and BTEO (Han et al., 2022a).

We also found that rejecting predictions in selective classification can impact relative fairness between various methods and models. For example, after a certain percentage of rejections, the fairness of a debiased model may decrease even below the fairness level of a standard model. This is similar to increased accuracy disparities observed in a rejection scenario (Jones et al., 2021) but has not been shown in a debiasing setup before. However, it is worth noting that BTJ consistently demonstrates robustly good fairness across all rejection rates.

Lastly, our experiments reveal that OOD detection is also vulnerable to debiasing when using the baseline UE technique of softmax response. However, applying a more advanced approach such as DDU alleviates this issue providing similar performance with the standard model for most debiasing methods. BTJ and BTEO combined with DDU also demonstrate very robust performance in this scenario. These combinations result in better fairness with no penalties for OOD detection compared to the standard model.

Overall, methods based on instance reweighting emerge as the most favorable choices for simultaneously obtaining fairness, good performance, and high reliability. Returning to the main research question, when using the right combination of techniques, it is possible to achieve both model fairness and reliability.

7 Limitations

- We focused exclusively on equal opportunity fairness in this paper, despite the myriad of different definitions of fairness in the literature, such as demographic parity. Therefore, in Appendix D, we provide a comprehensive description of the calculation process for the fairness metric. Although different fairness criteria may yield slightly different results, we hypothesize that these variations would not significantly alter the relationship between fairness and reliability. We leave further investigation of this matter to future research.
- We conducted experiments only on English. However, all methods are language-agnostic and are compatible with any transformer-based model. We do not expect there to be major deviations in results for other languages.
- We investigated group fairness under the assumption that we have an access to protected attributes, which is not always true for real-world datasets. On the other hand, this is a common assumption in work in the debiasing literature.

8 Ethical Considerations

In this work, we consider the trade-off between the performance, fairness, and reliability of a model. We used only publicly-available models and datasets, and only according to the intended use; to avoid any harm to users, we used only attributes that users have self-disclosed.

Acknowledgements

We are grateful to Trevor Cohn for his suggestions and help regarding the theoretical justification for experimenting with various test distributions (Appendix C.1). The work of Alexander Panchenko (Sections 3 and 4) was supported by the Russian Science Foundation grant 20-71-10135.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 49–59. ACM.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland. Association for Computational Linguistics.

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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Luigi P. Cordella, Claudio De Stefano, Francesco Tortorella, and Mario Vento. 1995. A method for improving classification reliability of multilayer perceptrons. *IEEE Transactions on Neural Networks*, 6 5:1140–7.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2019. Addressing age-related bias in sentiment analysis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6146–6150. ijcai.org.

Melanie Ducoffe and Frédéric Precioso. 2018. Adversarial active learning for deep networks: a margin based approach. *ArXiv preprint*, abs/1802.09841.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

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.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. 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, Timothy Baldwin, and Trevor Cohn. 2022b. Towards equal opportunity fairness through adversarial learning. *ArXiv preprint*, abs/2203.06317.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2023. Fair enough: Standardizing evaluation and model selection for fairness research in NLP. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–312.

Xudong Han, Aili Shen, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022c. Systematic evaluation of predictive fairness. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 68–81.

Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. 2022d. fairlib: A unified framework for assessing and improving classification fairness. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022) Demo Session*, Abu Dhabi, UAE.

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

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations*. OpenReview.net.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. 2019. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations*. OpenReview.net.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Yibo Hu and Latifur Khan. 2021. Uncertainty-aware reliable text classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 628–636.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. 2021. Selective classification can magnify disparities across groups. In *9th International Conference on Learning Representations*. OpenReview.net.

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 33: Annual Conference on Neural Information Processing Systems 2020*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. 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 6402–6413.

Ken Lang. 1995. NewsWeeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339. Morgan Kaufmann.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 7167–7177.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Jeremiah Z. Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human*

Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations*. OpenReview.net.

R Timothy Marler and Jasbir S Arora. 2004. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26:369–395.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. 2023. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394.

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.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 13675–13682. AAAI Press.

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.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. Fairbatch: Batch selection for model fairness. In *9th International Conference on Learning Representations*. OpenReview.net.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Optimising equal opportunity fairness in model training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4073–4084, Seattle, United States. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Cuong Tran, Keyu Zhu, Ferdinando Fioretto, and Pascal Van Henternyck. 2022a. Fairness increases adversarial vulnerability. *ArXiv preprint*, abs/2211.11835.

Dustin Tran, Jeremiah Zhe Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda E Mariet, Huiyi Hu, et al. 2022b. Plex: Towards reliability using pretrained large model extensions. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*.

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.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5309–5318. IEEE.

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.

Han Xu, Xiaorui Liu, Yaxin Li, Anil K. Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11492–11501. PMLR.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density

estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. 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 3126–3136, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Acronyms

Acronym	Full name	Description
Debiasing methods		
BTJ	Balanced Training with Joint balance	Balances the joint distribution of protected attributes and target labels through instance reweighting
BTEO	Balanced Training with Equal Opportunity	Balances the protected attributes within "advantage" classes through resampling instances based on equal opportunity objectives
Adv	Adversarial Training	Extends the training objective with a discriminator component responsible for making the model unlearn the protected attributes
DAdv	Diverse Adversaries approach	Similar to Adv, but add an ensemble of adversaries to the loss and subjects them to a diversity constraint for learning orthogonal hidden representations from one another
GD_{diff}	Group Difference	Adds a loss component that minimizes the loss gap across protected groups within each class
FairBatch	Fair Batch Selection	Dynamically adjusts the probability of resampling instances in each minibatch during training to achieve loss disparity across protected groups
INLP	Iterative Null-space Projection	Removes protected information from an already trained model by iteratively projecting its hidden representations to the null-space of protected attribute discriminators, and after uses these representations for classification
UE methods & UE scores		
SR	Softmax Response	Uses maximum probability from the softmax layer as a confidence score
MC	Monte-Carlo dropout	Makes N stochastic forward passes using dropout, obtained predictions aggregated with various UE scores (e.g. BALD, SMP, PV).
MD	Mahalanobis Distance	Estimates minimal class-conditional probability of an input instance x that follows a Gaussian distribution. The uncertainty score is computed as the Mahalanobis distance between latent instance representations of x and the closest centroid of a class.
DeepFool	Deep Fool	Computes the l_2 norm of the minimum perturbation vector that is required to apply to a latent representation to change the prediction of a model. The smaller the norm, the higher the uncertainty.
HUQ	Hybrid Uncertainty Quantification	Hybrid method, combining various UE methods. For more details refer to Appendix H.
BALD	Bayesian Active Learning with Disagreement	$\frac{1}{N}\sum_{c,n} p_n^c \log p_n^c - \sum_{n=1}^{N} \frac{1}{N}\sum_n p_n^c \frac{1}{N}\sum_n \log p_n^c$, where C = number of classes, N = number of stochastic passes, and p_n^c = probability of class c during the stochastic pass n.
SMP	Sampled Maximum Probability	$1 - \max_{c \in C} \sum_{n=1}^{N} p_n^c$
PV	Probability Variance	$\frac{1}{C}\sum_{c=1}^{C}(\frac{1}{N}\sum_{n=1}^{N}(p_n^c - \frac{1}{N}\sum_n \log p_n^c)^2)$

Table 1: Short-list of acronyms used in the paper.

Dataset	Debiasing Method	Num. Epochs	Batch Size	Learning Rate	Weight Decay	Debiasing Parameter
Bios (imbalanced)	Standard	20	16	5e-6	0	-
	BTEO	20	16	5e-6	0	-
	Adv	20	16	5e-6	0	1.0
	DAdv	20	16	5e-6	0	1.0/1.0
	INLP	20	16	5e-6	0	True/False
	FairBatch	20	16	5e-6	0	0.05
	GD_{diff}	20	16	5e-6	0	0.5
	BTJ	20	16	5e-6	0	-
Bios (balanced)	Standard	20	16	5e-6	0	-
	BTEO	20	16	5e-6	0	-
	Adv	20	16	5e-6	0	1.0
	DAdv	20	16	5e-6	0	1.0/1.0
	INLP	20	16	5e-6	0	False/True
	FairBatch	20	16	5e-6	0	0.05
	GD_{diff}	20	16	5e-6	0	0.5
	BTJ	20	16	5e-6	0	-

Table 2: Optimal training hyperparameters for BERT on Bios with various debiasing methods. 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]. For all models, dropout rate is 0.1. The number of epochs is determined by early-stopping.

Dataset	Debiasing Method	Num. Epochs	Batch Size	Learning Rate	Weight Decay	Debiasing Parameter
Moji (imbalanced)	Standard	20	32	1e-6	0	-
	BTEO	20	32	1e-6	0	-
	Adv	20	32	1e-6	0	1.0
	DAdv	20	32	1e-6	0	1.0/1.0
	INLP	20	32	1e-6	0	False/True
	FairBatch	20	32	1e-6	0	0.5
	GD_{diff}	20	32	1e-6	0	0.5
	BTJ	20	32	1e-6	0	-
Moji (balanced)	Standard	20	32	1e-6	0	-
	BTEO	20	32	1e-6	0	-
	Adv	20	32	1e-6	0	1.0
	DAdv	20	32	1e-6	0	1.0/1.0
	INLP	20	32	1e-6	0	False/False
	FairBatch	20	32	1e-6	0	0.5
	GD_{diff}	20	32	1e-6	0	0.5
	BTJ	20	32	1e-6	0	-

Table 3: Optimal training hyperparameters for BERTweet on Moji with various debiasing methods. 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]. For all models, dropout rate is 0.1. The number of epochs is determined by early-stopping.

B Hyperparameters and Computational Resources

For hyperparameter optimization, we employed the standard grid-search with accuracy on the validation set as an optimization target for the standard model and with DTO for the debiased models. The grid and the best parameters are described in Tables 2 to 4. For each debiasing method we tuned method-specific parameters, namely: adv_lambda for Adv, adv_lambda/adv_diverse_lambda for DAdv, INLP_discriminator_reweighting/INLP_by_class for INLP, DyBTalpha for FairBatch and GD_{diff}. The remaining parameters are given by default in the Fairlib framework (Han et al., 2022d).

All experiments were conducted on a cluster with Nvidia V100 GPUs. The total amount of GPU hours and the number of model parameters are specified in Table 6.

Dataset	Debiasing Method	Num. Epochs	Batch Size	Learning Rate	Dropout Rate	Debiasing Parameter
Moji (imbalanced)	Standard	100	512	1e-4	0	-
	BTEO	100	512	1e-4	0	-
	Adv	100	512	1e-4	0	1.0
	DAdv	100	512	1e-4	0	1.0/1.0
	INLP	100	512	1e-4	0	False/False
	FairBatch	100	512	1e-4	0	0.1
	GD_{diff}	100	512	1e-4	0	0.5
	BTJ	100	512	1e-4	0	-
Moji (balanced)	Standard	100	128	3e-3	0.1	-
	BTEO	100	128	3e-3	0.1	-
	Adv	100	128	3e-3	0.1	10.0
	DAdv	100	128	3e-3	0.1	1.0/1.0
	INLP	100	128	3e-3	0.1	False/False
	FairBatch	100	128	3e-3	0.1	0.1
	GD_{diff}	100	128	3e-3	0.1	0.5
	BTJ	100	128	3e-3	0.1	-

Table 4: Optimal training hyperparameters for MLP+DeepMoji on Moji with various debiasing methods. We use a grid search with the following grid values: batch size: [64, 128, 256, 512, 1024], learning rate: [1e-2, 5e-3, 3e-3, 1e-3, 5e-4, 1e-4], dropout rate: [0.0, 0.1, 0.2, 0.3]. The number of epochs is determined by early-stopping.

Method	Parameter	Search Range
Adv	adv_lambda	[1e-4, 1e-3, 1e-2, 1e-1, 1, 1e2, 1e3]
DAdv	adv_lambda/adv_diverse_lambda	[1e-4, 1e-3, 1e-2, 1e-1, 1, 1e2, 1e3]
INLP	INLP_discriminator_reweighting	[True, False]
INLP	INLP_by_class	[True, False]
FairBatch	DyBTalpha	[1e-4, 1e-3, 1e-2, 5e-2, 1e-1, 5e-1, 1]
GD_{diff}	DyBTalpha	[1e-4, 1e-3, 1e-2, 5e-2, 1e-1, 5e-1, 1]

Table 5: The hyperparameter grid for debiasing methods. For DAdv we jointly tuned both parameters as in Han et al. (2022d).

Dataset	Model	GPU hours	Num. of Params
Moji	BERTweet	737	135m
Moji	MLP	89	0.3m
Bios	BERT	1134	110m

Table 6: Overall computation statistics. GPU hours specify the approximate number of GPU hours spent for training and evaluating the corresponding model for all debiasing and UE methods on both imbalanced and balanced sets. The column Num. of Params contains the number of parameters of a single model.

C Dataset Statistics and Test Set Distributions

In this section, we present the statistics of the datasets used in our experiments, including a joint probability distribution of the target value and protected attribute: Tables 7 to 10. We note that in Moji, the original distribution in test and validation is balanced, so we manipulated this distribution to create the "imbalanced" version. In Bios, on the contrary, the test and validation follow the training distribution, so we modify them to create the "balanced" version. Below, we provide a theoretical justification for performing such manipulations of the test distributions. We also present statistics of datasets used as OOD domains in Tables 11 and 12.

C.1 Theoretical Justification for Experimenting with Various Test Distributions

Consider a dataset consisting of n instances $\mathcal{D} = \{(\boldsymbol{x}_i, \mathrm{y}_i, \mathrm{z}_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i$ is an input vector to the classifier, $\mathrm{y}_i \in [0, 1]$ represents binary target class label, and $\mathrm{z}_i \in [\mathrm{g}, \neg\mathrm{g}]$ is the binary group label, such as gender. $n_{\mathrm{c,g}}$ denotes the number of instances in a subset with target label c and protected label g. A vanilla model (m) makes prediction, $\hat{\mathrm{y}} = m(\boldsymbol{x})$.

When evaluating the accuracy of a model m,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{TP}_\text{g} + \text{TP}_{\neg\text{g}} + \text{TN}_\text{g} + \text{TN}_{\neg\text{g}}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Since the denominator in Equation (1) is a constant number for a particular dataset (which is the total number of instances in the test set), to simplify the analysis, we will focus on the numerator hereafter $\text{TP}_\text{g} + \text{TP}_{\neg\text{g}} + \text{TN}_\text{g} + \text{TN}_{\neg\text{g}}$.

How does Equation (1) related to Equal Opportunity Fairness? Recall that equal opportunity fairness is measured by the equality of true positive rate (TPR), e.g., the TPR gap, $|\text{TPR}_\text{g} - \text{TPR}_{\neg\text{g}}|$, between two demographic groups.

Let * denote the results after bias mitigation, e.g. TP^*_g is the TP of group g after debiasing, and **assuming** that bias mitigation w.r.t. equal opportunity fairness only changes the prediction w.r.t. positive instances,

$$\text{Accuracy} - \text{Accuracy}^* = \frac{1}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}(\text{TP}_\text{g} + \text{TP}_{\neg\text{g}} - \text{TP}^*_\text{g} - \text{TP}^*_{\neg\text{g}}) \tag{2}$$

Moreover, by definition, $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$, therefore, Equation (2) can be expressed based on TPR:

$$n(\text{Accuracy} - \text{Accuracy}^*) = (n_{1,\text{g}}\text{TPR}_\text{g} + n_{1,\neg\text{g}}\text{TPR}_{\neg\text{g}} - n_{1,\text{g}}\text{TPR}^*_\text{g} - n_{1,\neg\text{g}}\text{TPR}^*_{\neg\text{g}}) \tag{3}$$

where, as introduced before, $n = \text{TP} + \text{TN} + \text{FP} + \text{FN}$, $n_{1,\text{g}} = \text{TP}_\text{g} + \text{FN}_\text{g}$, and $n_{1,\neg\text{g}} = \text{TP}_{\neg\text{g}} + \text{FN}_{\neg\text{g}}$. By grouping TPR by groups, we can see that

$$n(\text{Accuracy} - \text{Accuracy}^*) = n_{1,\text{g}}(\text{TPR}_\text{g} - \text{TPR}^*_\text{g}) + n_{1,\neg\text{g}}(\text{TPR}_{\neg\text{g}} - \text{TPR}^*_{\neg\text{g}}) \tag{4}$$

Let $\Delta_\text{g} = \text{TPR}_\text{g} - \text{TPR}^*_\text{g}$ and $\Delta_{\neg\text{g}} = \text{TPR}_{\neg\text{g}} - \text{TPR}^*_{\neg\text{g}}$ denote the changes in TPR after debiasing for group g and $\neg$g, respectively. Although debiasing may decrease the TPR of the majority group, a good debiasing method should result in a larger increase in terms of the TPR of the minority group, i.e., $\Delta_\text{g} + \Delta_{\neg\text{g}} < 0$. For example, the vanilla model achieves $\text{TPR}_\text{g} = 0.8$ and $\text{TPR}_{\neg\text{g}} = 0.2$, and a debiased model achieves $\text{TPR}^*_\text{g} = 0.7$ and $\text{TPR}^*_{\neg\text{g}} = 0.5$. In this example,

$$\Delta_\text{g} + \Delta_{\neg\text{g}} = 0.1 + (-0.3) = -0.2$$

How does the test set distribution affect the accuracy evaluation? If the test set is balanced (i.e., $n_{1,\text{g}} = n_{1,\neg\text{g}} = \frac{1}{2}n_1$), Equation (4) can be simplified as:

$$\text{Accuracy} - \text{Accuracy}^* = \frac{n_1}{2n}(\Delta_\text{g} + \Delta_{\neg\text{g}}),$$

showing that the debiased method could **improve** the accuracy score by $-\frac{n_1}{2n}(\Delta_\text{g} + \Delta_{\neg\text{g}})$.

Split	Gender \ Profession	Attorney	Dentist	Journalist	Nurse	Photographer	Physician	Psychologist	Surgeon	Teacher	Total
Train	Female	6.44	2.52	5.06	8.88	4.60	8.38	5.81	0.97	5.03	47.68
	Male	10.05	4.76	5.03	0.87	7.93	11.35	3.45	5.81	3.07	52.32
	All	16.49	7.28	10.09	9.74	12.53	19.73	9.26	6.78	8.10	100.00
Val	Female	6.86	3.27	5.42	8.65	4.99	8.16	6.34	0.83	5.26	49.79
	Male	9.74	4.32	4.41	0.89	7.50	11.00	3.08	6.05	3.21	50.21
	All	16.60	7.59	9.83	9.55	12.49	19.16	9.42	6.88	8.47	100.00
Test	Female	5.84	2.63	4.76	8.61	4.00	14.76	5.44	1.19	4.75	51.97
	Male	10.50	4.59	5.33	0.82	8.56	5.21	3.84	5.59	3.58	48.03
	All	16.33	7.22	10.10	9.43	12.55	19.97	9.28	6.78	8.33	100.00

Table 7: Bios distribution (imbalanced test and validation sets) over target variable and protected attribute for all subsets.

Split	Gender \ Profession	Attorney	Dentist	Journalist	Nurse	Photographer	Physician	Psychologist	Surgeon	Teacher	Total
Train	Female	6.44	2.52	5.06	8.88	4.60	8.38	5.81	0.97	5.03	47.68
	Male	10.05	4.76	5.03	0.87	7.93	11.35	3.45	5.81	3.07	52.32
	All	16.49	7.28	10.09	9.74	12.53	19.73	9.26	6.78	8.10	100.00
Val	Female	9.60	4.58	6.18	1.25	6.99	11.43	4.31	1.17	4.50	50.00
	Male	9.60	4.58	6.18	1.25	6.99	11.43	4.31	1.17	4.50	50.00
	All	19.21	9.16	12.35	2.50	13.97	22.85	8.62	2.33	8.99	100.00
Test	Female	9.16	4.12	7.47	1.28	6.27	8.17	6.03	1.87	5.62	50.00
	Male	9.16	4.12	7.47	1.28	6.27	8.17	6.03	1.87	5.62	50.00
	All	18.31	8.24	14.94	2.57	12.54	16.35	12.06	3.74	11.24	100.00

Table 8: Bios distribution (balanced test and validation sets) over target variable and protected attribute for all subsets.

Split	Ethnicity \ Sentiment score	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

Table 9: Moji distribution (imbalanced test and validation sets) over target variable and protected attribute for all subsets.

On the other hand, if the test set is imbalanced, let's explore the condition when debiasing does not affect accuracy:

$$n_{1,\mathrm{g}}\Delta_{\mathrm{g}} + n_{1,\neg\mathrm{g}}\Delta_{\neg\mathrm{g}} = 0 \Leftrightarrow \frac{n_{1,\mathrm{g}}}{n_{1,\neg\mathrm{g}}} = -\frac{\Delta_{\neg\mathrm{g}}}{\Delta_{\mathrm{g}}}$$

In our previous example, where $\Delta_{\mathrm{g}} = 0.1$ and $\Delta_{\neg\mathrm{g}} = -0.3$, the debiasing method does not affect accuracy if $\frac{n_{1,\mathrm{g}}}{n_{1,\neg\mathrm{g}}} = -\frac{-0.3}{0.1} = 3$.

In general, if there much more instances in group g than group ¬g for class 1 (i.e., $\frac{n_{1,\mathrm{g}}}{n_{1,\neg\mathrm{g}}} > -\frac{\Delta_{\neg\mathrm{g}}}{\Delta_{\mathrm{g}}}$), debiasing would decrease accuracy.

C.2 Main Datasets with Balanced and Imbalanced Test Distributions

Split	Ethnicity \ Sentiment score	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	25.00	25.00	50.00
	AA	25.00	25.00	50.00
	All	50.00	50.00	100.00
Test	SA	25.01	25.01	50.01
	AA	24.99	24.99	49.99
	All	50.00	50.00	100.00

Table 10: Moji distribution (balanced test and validation sets) over the target variable and protected attribute for all subsets.

Dataset	Num. of classes	Protected attribute	Num. of attributes	Train/Val/Test
Moji (balanced)	2	Race	2	100k/8k/8k
Moji (imbalanced)	2	Race	2	100k/5k/5k
Bios (imbalanced)	9	Gender	2	64k/10k/25k
Bios (balanced)	9	Gender	2	64k/7k/16k

Table 11: Dataset statistics. This table presents overall dataset statistics with the number of samples for each split and task-specific parameters, such as the number of classes and protected attributes. We keep only 9 prevalent professions except "professor" since it already has a balanced distribution.

Dataset	Num. of classes	Train/Test
IMDB (Maas et al., 2011)	2	20K/25K
20 News Groups (Lang, 1995)	20	11.3K/7.5K
TREC-10 (Li and Roth, 2002; Hovy et al., 2001)	6	5.5K/0.5K
SST-2 (Socher et al., 2013)	2	67.3K/0.9K
WMT-16 (Bojar et al., 2016)	-	4500K/3K

Table 12: OoD dataset statistics. The table presents the number of samples for the training and test parts of the datasets. For the SST-2 dataset, we used the available validation set as the test set. From these datasets, we use only the entire test part as OoD instances.

C.3 Datasets Used as Out-of-distribution Domains

D Details of Fairness Metric Calculation

1. We calculate the true positive rate (TPR) for each of the protected groups in a dataset:

$$TPR = \frac{TP}{TP + FN}. \tag{5}$$

2. We group-wise aggregate TPR gap according to the following formula:

$$\beta_c = \sum_g |TPR_{c,g} - \overline{TPR_c}|. \tag{6}$$

   Here, $\overline{TPR_c}$ stands for $TPR_{c,g}$ averaged across groups.
3. We aggregate acquired β_c class-wise:

$$\delta = \sqrt{\frac{1}{C}\sum_c \beta_c^2}. \tag{7}$$

4. Finally, we subtract δ from 1 to align fairness with accuracy. Optionally, we multiply it by 100 for easy comparison to other metrics:

$$Fairness = 100 \cdot (1 - \delta). \tag{8}$$

E Example of a Risk–Coverage Curve

Figure 6 presents the risk–coverage curve for the Bios dataset with the standard BERT model. In this example, HUQ is used as an uncertainty estimation method for rejecting instances.

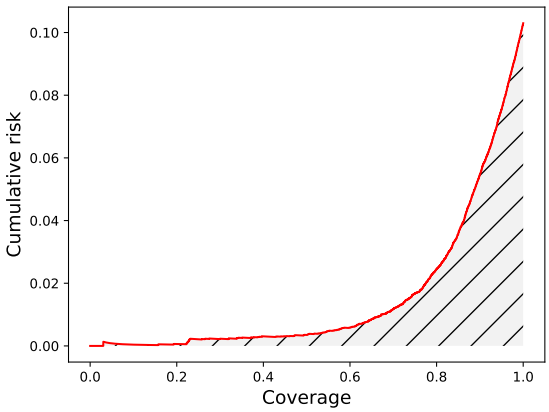


Figure 6: The example of the RC curve for the Bios dataset with the standard model.

F Additional Experimental Results for Selective Classification

Method	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
-	Fairness ↑	90.5±0.5	93.4±0.8	92.9±0.4	92.7±0.4	90.9±0.5	91.8±0.6	92.6±0.4	91.0±0.4
-	Accuracy ↑	89.7±0.2	88.9±0.4	89.4±0.2	89.4±0.3	89.2±0.1	89.0±0.2	89.5±0.3	89.5±0.2
-	DTO ↓	14.0±0.2	13.0±0.3	12.8±0.3	12.8±0.4	14.1±0.2	13.8±0.3	12.8±0.4	13.8±0.3
MD	RC-AUC ↓	430.8±15.8	595.0±85.9	517.3±70.0	575.8±203.3	478.4±84.2	640.4±120.2	504.2±38.5	455.6±27.6
MC (SMP)	RC-AUC ↓	379.5±11.0	480.1±72.2	421.7±14.6	**416.1±21.6**	**416.7±14.9**	619.7±48.5	391.1±22.2	455.8±38.4
MC (PV)	RC-AUC ↓	379.0±17.6	469.0±60.2	438.3±4.0	438.9±35.7	427.6±8.2	545.1±60.2	404.5±24.9	451.2±22.7
MC (BALD)	RC-AUC ↓	373.7±15.4	469.7±62.9	444.6±16.0	460.0±73.7	420.1±7.6	567.3±88.6	416.3±27.7	437.0±24.6
HUQ (DeepFool + MD)	RC-AUC ↓	**368.8±17.2**	**464.4±77.5**	**408.1±13.1**	436.0±77.1	456.1±59.4	**525.2±55.3**	**377.1±21.8**	**386.2±26.4**
DeepFool	RC-AUC ↓	402.7±17.3	466.3±53.4	430.7±48.4	428.7±33.4	447.3±41.6	628.0±64.9	389.8±21.4	919.8±272.4
DDU	RC-AUC ↓	719.3±31.3	915.1±92.4	706.3±133.4	812.8±91.8	586.4±106.7	877.2±93.8	785.2±54.1	709.4±26.5
Baseline (SR)	RC-AUC ↓	429.1±18.9	509.5±66.3	478.0±54.3	463.3±50.9	527.9±76.0	738.4±54.9	419.6±26.8	587.7±66.6

Table 13: Performance of selective classification (RC-AUC) for various debiasing methods on Bios with imbalanced test and validation sets (BERT model). The best results for each debiasing method are highlighted with bold font.

Method	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
-	Fairness ↑	90.5±0.5	93.4±1.1	92.8±0.6	93.1±0.6	90.5±0.9	93.0±1.3	92.5±0.6	90.7±0.5
-	Accuracy ↑	89.1±0.2	88.7±0.1	89.3±0.2	89.1±0.3	88.3±0.1	88.7±0.4	89.2±0.1	88.9±0.3
-	DTO ↓	14.5±0.4	13.2±0.6	12.9±0.3	12.9±0.5	15.1±0.6	13.4±0.7	13.1±0.4	14.5±0.2
MD	RC-AUC ↓	334.8±36.7	397.3±19.8	349.6±20.7	385.8±104.3	340.4±42.0	422.8±65.5	347.1±32.3	347.1±37.2
MC (SMP)	RC-AUC ↓	293.2±16.4	347.9±28.8	283.9±14.1	**286.1±8.7**	**305.3±9.9**	462.4±53.0	293.7±19.8	328.7±42.9
MC (PV)	RC-AUC ↓	290.3±16.9	337.8±26.4	288.8±8.8	301.4±16.6	316.6±7.6	**378.7±35.4**	299.1±16.7	327.8±36.8
MC (BALD)	RC-AUC ↓	287.6±21.7	338.0±33.7	291.8±7.4	312.9±36.8	310.5±8.0	390.2±46.9	302.9±14.8	319.3±35.7
HUQ (DeepFool + MD)	RC-AUC ↓	**285.0±22.4**	**327.5±25.8**	297.0±14.6	286.4±5.8	323.0±19.3	396.0±63.1	**283.7±18.1**	**312.7±63.3**
DeepFool	RC-AUC ↓	301.3±4.9	339.5±17.2	**279.0±13.5**	293.1±17.1	332.9±15.5	453.6±43.4	290.1±12.0	489.4±137.2
DDU	RC-AUC ↓	497.8±33.5	557.4±40.2	478.0±23.5	500.6±44.5	405.5±59.0	557.1±54.0	516.7±40.3	487.0±29.9
Baseline (SR)	RC-AUC ↓	326.2±14.3	369.3±21.9	309.3±20.6	312.3±15.0	362.2±32.8	519.6±37.1	310.8±15.9	452.2±123.8

Table 14: Performance of selective classification (RC-AUC) for various debiasing methods on Bios with balanced test and validation sets (BERT model). The best results for each debiasing method are highlighted with bold font.

Method	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
-	Fairness ↑	61.5±0.4	76.7±0.7	86.4±1.4	85.1±0.6	85.2±1.4	91.0±1.1	85.2±0.4	62.2±0.6
-	Accuracy ↑	82.9±0.1	79.5±0.6	77.8±1.6	78.6±0.2	78.6±0.5	76.1±0.5	78.5±0.1	82.3±0.8
-	DTO ↓	42.1±0.3	31.1±0.4	26.1±0.8	26.1±0.3	26.0±0.5	25.5±0.3	26.1±0.3	41.8±0.3
MD	RC-AUC ↓	874.4±7.8	1052.4±21.1	1177.7±47.2	1157.0±14.3	1095.1±24.1	1239.4±34.7	1118.7±15.9	950.6±70.9
MC (SMP)	RC-AUC ↓	344.6±2.3	457.8±21.6	580.7±78.7	**530.3±8.3**	537.4±27.8	669.1±24.8	**529.2±5.9**	358.7±28.1
MC (PV)	RC-AUC ↓	408.5±6.8	570.8±48.6	662.8±196.6	560.5±10.9	634.3±19.9	1226.6±81.6	688.6±29.8	458.6±53.5
MC (BALD)	RC-AUC ↓	779.9±39.4	969.8±165.1	775.7±323.5	615.4±24.9	952.6±88.7	1559.3±44.6	1191.7±69.9	846.2±49.8
HUQ (SR + MD)	RC-AUC ↓	**342.3±2.4**	**452.8±20.8**	**578.5±74.4**	530.5±8.1	**535.8±27.4**	**668.9±24.8**	**529.2±5.9**	**357.2±27.4**
DeepFool	RC-AUC ↓	344.6±2.1	457.8±21.6	580.5±78.8	530.6±8.0	537.3±27.9	**668.9±24.8**	**529.2±5.9**	757.6±97.6
DDU	RC-AUC ↓	818.2±6.8	986.8±20.2	1077.2±50.6	1048.9±9.6	1024.9±22.9	1152.0±29.9	1044.8±11.7	867.6±61.8
Baseline (SR)	RC-AUC ↓	344.6±2.1	457.8±21.6	580.5±78.8	530.6±8.0	537.3±27.9	**668.9±24.8**	**529.2±5.9**	358.1±27.6

Table 15: Performance of selective classification for various debiasing methods on Moji with the imbalanced test and validation sets (DeepMoji+MLP model). The best results for each debiasing method are highlighted with bold font.

Method	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
-	Fairness ↑	61.4±2.0	83.2±5.8	89.8±1.0	89.9±1.1	91.1±1.0	90.8±1.0	89.6±1.9	86.6±5.2
-	Accuracy ↑	71.4±0.7	71.4±1.8	73.4±0.7	74.9±0.7	74.0±1.0	74.2±0.9	75.1±0.6	68.4±3.2
-	DTO ↓	48.0±2.0	33.6±1.5	28.5±0.7	27.1±0.4	27.5±0.6	27.4±0.7	27.1±0.7	34.6±3.6
MD	RC-AUC ↓	2086.3±212.7	2115.1±126.9	2151.5±145.2	2190.7±213.9	2325.3±147.2	2537.9±145.8	1869.7±313.8	2314.5±361.5
MC (SMP)	RC-AUC ↓	**1358.9±50.4**	1327.4±77.6	1252.8±117.4	1088.7±31.3	1127.5±69.8	**1158.2±140.9**	1049.3±41.0	1814.7±464.8
MC (PV)	RC-AUC ↓	1504.2±77.6	1397.7±131.1	1662.5±243.2	1256.8±104.5	1216.5±79.6	1902.0±124.3	1128.2±54.9	1901.9±461.5
MC (BALD)	RC-AUC ↓	1697.8±127.9	1499.2±218.0	1756.1±212.5	1476.8±169.1	1364.8±100.2	2171.6±82.9	1249.6±123.4	2013.5±471.9
HUQ (SR + MD)	RC-AUC ↓	1360.6±50.2	**1325.9±78.1**	**1236.5±132.3**	**1086.5±35.1**	**1125.3±68.9**	1158.9±144.4	**1049.0±41.2**	**1719.0±434.9**
DeepFool	RC-AUC ↓	1360.6±50.2	1326.1±78.2	1255.6±118.7	1089.1±32.8	1128.7±69.3	1159.5±143.8	1049.2±41.6	2555.2±353.1
DDU	RC-AUC ↓	1997.2±167.6	1860.2±110.6	2083.7±385.3	2079.0±199.6	2231.2±167.4	2430.6±157.5	1800.6±343.9	2247.3±275.0
Baseline (SR)	RC-AUC ↓	1360.6±50.2	1326.1±78.2	1255.8±118.8	1089.1±32.8	1128.7±69.3	1159.5±143.8	1049.2±41.6	1823.3±468.2

Table 16: Performance of selective classification for various debiasing methods on Moji with the balanced test set (DeepMoji+MLP model). The best results for each debiasing method are highlighted with bold font.

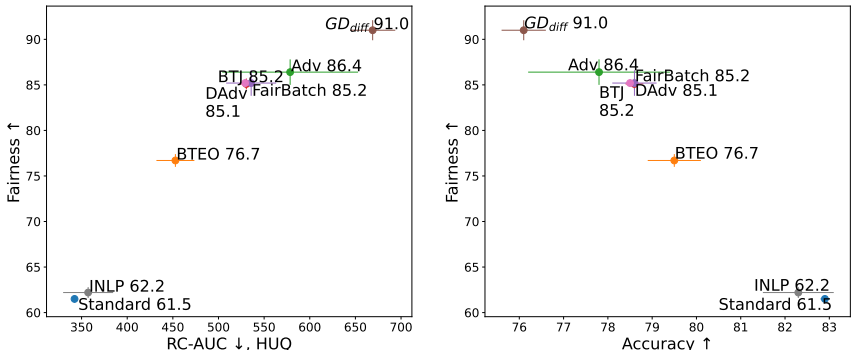


Figure 7: Trade-off between RC-AUC of selective classification with HUQ and fairness (left) and between accuracy and fairness (right) on Moji with the imbalanced test set with MLP model. The fairness scores are presented alongside each method for better comparison.

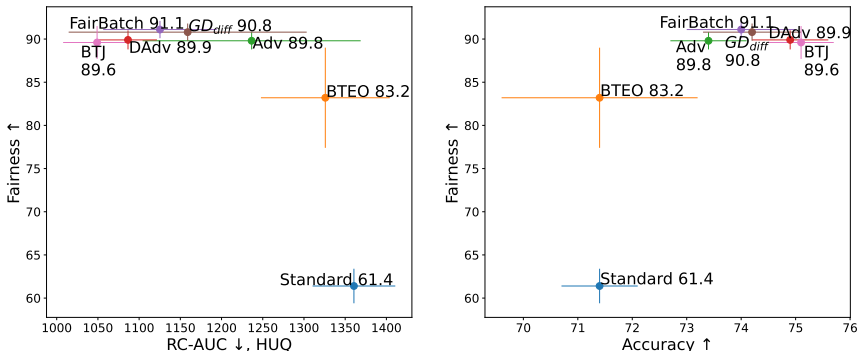


Figure 8: Trade-off between RC-AUC of selective classification with HUQ and fairness (left) and between accuracy and fairness (right) on Moji with the balanced test set with MLP model. We removed results for INLP from this figure due to the high RC-AUC value for this debiasing method. The fairness scores are presented alongside each method for better comparison.

Method	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
-	Fairness ↑	62.8±0.8	86.0±0.9	87.4±0.7	87.1±0.5	81.4±12.8	84.6±1.6	86.1±0.7	74.3±12.8
-	Accuracy ↑	82.5±0.3	78.6±0.7	77.9±1.0	78.3±0.8	75.6±5.2	78.5±0.6	79.0±0.6	72.6±11.1
-	DTO ↓	41.2±0.7	25.6±0.3	25.5±1.0	25.3±0.8	32.3±7.8	26.5±0.8	25.2±0.3	40.3±3.4
MD	RC-AUC ↓	530.0±44.6	842.8±48.4	777.1±23.5	762.0±18.0	993.2±290.0	824.6±67.6	751.8±49.1	1010.3±649.8
MC (SMP)	RC-AUC ↓	**377.9±10.8**	**524.3±23.8**	**631.0±41.1**	624.3±19.9	758.7±295.6	568.8±21.6	**488.1±8.6**	1189.5±860.1
MC (PV)	RC-AUC ↓	410.0±21.8	582.5±32.5	637.4±45.5	**621.9±24.7**	958.4±308.8	794.4±123.4	521.6±11.2	1428.4±883.0
MC (BALD)	RC-AUC ↓	447.7±31.0	631.7±34.9	655.8±47.8	638.7±28.4	1078.1±315.2	847.2±119.5	558.9±14.2	1517.1±814.3
HUQ (SR + MD)	RC-AUC ↓	392.1±11.3	532.0±25.4	660.3±49.3	646.0±22.5	**681.7±218.9**	**560.3±20.2**	502.6±10.7	877.4±682.5
DeepFool	RC-AUC ↓	392.6±11.9	532.6±24.7	682.3±50.2	664.3±22.5	703.7±240.4	586.2±18.8	503.1±10.9	1455.6±906.5
DDU	RC-AUC ↓	521.2±59.1	811.9±39.9	723.6±26.4	710.8±20.5	992.8±300.5	775.9±76.4	716.1±63.7	**841.2±429.9**
Baseline (SR)	RC-AUC ↓	392.7±11.5	532.4±25.2	681.5±49.3	664.6±22.6	702.5±246.5	585.9±19.0	502.9±10.8	1086.6±882.9

Table 17: Performance of selective classification for various debiasing methods on Moji with the imbalanced test set (BERTweet model). The best results for each debiasing method are highlighted with bold font.

Method	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
-	Fairness ↑	64.5±0.7	86.7±0.7	87.0±0.5	87.3±1.6	85.2±6.5	84.8±1.1	86.1±0.4	78.4±13.3
-	Accuracy ↑	72.5±0.4	76.4±0.6	76.3±0.6	76.3±1.1	72.9±2.6	76.9±0.2	76.3±0.7	64.3±8.9
-	DTO ↓	44.9±0.8	27.1±0.3	27.0±0.6	27.0±0.5	31.2±5.1	27.7±0.4	27.5±0.4	44.0±3.6
MD	RC-AUC ↓	1709.0±117.3	1535.9±75.7	1342.2±53.1	1327.0±65.0	1850.9±284.2	1471.8±101.0	1429.5±41.4	2333.3±720.9
MC (SMP)	RC-AUC ↓	**1294.6±25.4**	**935.5±35.1**	**1077.4±40.7**	1157.4±158.1	1391.9±355.8	1034.9±10.5	**918.7±26.9**	2488.7±1066.2
MC (PV)	RC-AUC ↓	1440.3±58.4	1077.8±35.7	1113.8±53.1	1239.2±191.9	1715.4±185.1	1448.2±205.9	1021.3±22.3	2977.7±1346.9
MC (BALD)	RC-AUC ↓	1565.5±82.4	1187.7±37.7	1152.2±57.8	1276.7±182.0	1921.6±203.1	1526.9±181.3	1111.0±21.2	3082.1±1253.9
HUQ (SR + MD)	RC-AUC ↓	1299.7±22.5	954.2±37.8	1100.2±48.5	**1154.0±115.4**	**1290.9±315.5**	**1023.0±27.0**	953.8±34.6	**2096.0±884.9**
DeepFool	RC-AUC ↓	1299.3±22.7	954.5±37.6	1109.4±48.9	1157.7±109.2	1334.7±329.1	1045.7±14.8	954.2±35.1	3323.6±1923.2
DDU	RC-AUC ↓	1709.7±179.9	1541.7±106.2	1266.5±50.7	1271.3±69.4	1913.1±337.7	1455.0±113.3	1379.6±33.4	2158.5±593.6
Baseline (SR)	RC-AUC ↓	1299.7±22.5	954.2±37.8	1108.4±49.8	1160.2±115.5	1339.0±331.7	1045.6±14.8	953.8±34.6	2430.2±1184.2

Table 18: Performance of selective classification for various debiasing methods on Moji with the balanced test set (BERTweet model). The best results for each debiasing method are highlighted with bold font.

G Additional Experimental Results for Out-of-distribution Detection

Method	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP	Mean Δ
-	Fairness ↑	64.5±0.7	86.7±0.7	87.0±0.5	87.3±1.6	85.2±6.5	84.8±1.1	86.1±0.4	78.4±13.3	20.6±3.7
-	Accuracy ↑	72.5±0.4	76.4±0.6	76.3±0.6	76.3±1.1	72.9±2.6	76.9±0.2	76.3±0.7	64.3±8.9	1.7±2.2
-	DTO ↓	44.9±0.8	27.1±0.3	27.0±0.6	27.0±0.5	31.2±5.1	27.7±0.4	27.5±0.4	44.0±3.6	-14.7±1.9
DDU	ROC-AUC ↑	**95.4±1.7**	**95.4±2.4**	92.4±2.2	90.0±4.0	**93.0±6.0**	**85.3±5.0**	**95.9±2.2**	95.5±1.7	**-2.9±3.8**
MD	ROC-AUC ↑	**95.7±1.7**	95.3±2.3	**92.7±2.3**	**90.3±4.1**	92.5±6.2	84.1±5.4	**96.1±2.2**	**95.9±1.7**	-3.3±3.9
SR	ROC-AUC ↑	76.6±2.9	69.3±4.2	70.4±3.9	68.9±7.0	60.9±16.1	74.0±3.2	67.7±3.5	59.8±16.4	-9.3±8.5

Table 19: Performance of OoD detection for various debiasing and UE methods over the Moji dataset with the balanced test set (BERTweet model). The best results for each debiased model are highlighted in bold.

Method	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP	Mean Δ
-	Fairness ↑	90.5±0.5	93.4±1.1	92.8±0.6	93.1±0.6	90.5±0.9	93.0±1.3	92.5±0.6	90.7±0.5	1.8±1.0
-	Accuracy ↑	89.1±0.2	88.7±0.1	89.3±0.2	89.1±0.3	88.3±0.1	88.7±0.4	89.2±0.1	88.9±0.3	-0.2±0.3
-	DTO ↓	14.5±0.4	13.2±0.6	12.9±0.3	12.9±0.5	15.1±0.6	13.4±0.7	13.1±0.4	14.5±0.2	-0.9±0.6
DDU	ROC-AUC ↑	**93.8±1.6**	**94.9±1.8**	**94.2±1.1**	**94.8±0.9**	**92.6±0.8**	**93.8±1.4**	95.5±1.4	**93.8±1.6**	**0.4±2.1**
MD	ROC-AUC ↑	93.5±1.5	94.6±2.1	93.6±1.4	94.3±1.3	91.9±1.0	**93.8±1.6**	**95.6±1.4**	93.6±1.5	**0.4±2.1**
SR	ROC-AUC ↑	91.6±1.1	91.2±1.8	92.1±0.9	91.9±1.2	91.6±1.2	87.7±2.0	92.0±1.1	89.8±1.7	-0.7±1.8

Table 20: Performance of OoD detection for various debiasing and UE methods over the Bios dataset with the balanced test set (BERT model). The best results for each debiased model are highlighted in bold.

H Details of the Hybrid Uncertainty Quantification Method

We combine aleatoric and epistemic uncertainty in a single score, which we call **Hybrid Uncertainty Quantification (HUQ)**. Consider we have a training dataset $\mathcal{D}$. We define $\mathcal{D}_{\text{ID}} = \{\mathbf{x} \in \mathcal{D} \colon U_{\text{E}}(\mathbf{x}) \leq \delta_{\min}\}$ as in-distribution instances from $\mathcal{D}$; $\mathcal{X}_{\text{ID}} = \{\mathbf{x} \colon U_{\text{E}}(\mathbf{x}) \leq \delta_{\min}\}$ as arbitrary in-distribution instances; $\mathcal{X}_{\text{IDA}} = \{\mathbf{x} \in \mathcal{X}_{\text{ID}} \colon U_{\text{A}}(\mathbf{x}) > \delta_{\max}\}$ as ambiguous in-distribution instances (instances that lie on the discriminative border of the trained classifier). Here, $\delta_{\min}$, $\delta_{\max}$ are thresholds selected on the validation dataset. Consider we are given measures of aleatoric $U_{\text{A}}(\mathbf{x})$ and epistemic $U_{\text{E}}(\mathbf{x})$ uncertainty. To make different UE scores comparable, we define a ranking function $R(\mathbf{u}, \mathfrak{D})$ as a rank of $\mathbf{u}$ over a sorted dataset $\mathfrak{D}$, where $\mathbf{u}_1 > \mathbf{u}_2$ implies $R(\mathbf{u}_1, \mathfrak{D}) > R(\mathbf{u}_2, \mathfrak{D})$. For a given measure of aleatoric and epistemic uncertainty, we compute total uncertainty $U_{\text{T}}(\mathbf{x})$ as a linear combination $U_{\text{T}}(\mathbf{x}) = (1-\alpha) R(U_{\text{E}}(\mathbf{x}), \mathcal{D}) + \alpha R(U_{\text{A}}(\mathbf{x}), \mathcal{D})$, where α is a hyperparameter selected on the validation dataset. Finally, we define HUQ as follows:

$$U_{\text{HUQ}}(\mathbf{x}) = \begin{cases} R(U_{\text{A}}(\mathbf{x}), \mathcal{D}_{\text{ID}}), \forall \mathbf{x} \in \mathcal{X}_{\text{ID}} \setminus \mathcal{X}_{\text{AID}}, \\ R(U_{\text{A}}(\mathbf{x}), \mathcal{D}), \forall \mathbf{x} \in \mathcal{X}_{\text{AID}}, \\ U_{\text{T}}(\mathbf{x}), \forall \mathbf{x} \notin \mathcal{X}_{\text{ID}}. \end{cases}$$

I Disaggregated Experimental Results

Group	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
SAE	TPR	76.8±0.2	74.2±1.1	79.9±0.7	79.9±0.2	80.6±0.2	78.2±0.5	81.0±0.3	75.9±1.9
AAE	TPR	66.7±0.2	70.8±0.3	70.1±1.0	70.4±0.2	70.3±0.5	70.9±0.5	70.1±0.3	66.8±0.5
SAE Sad	TPR	89.3±0.2	92.6±0.8	78.6±0.5	79.4±0.9	81.6±0.9	75.5±2.8	80.2±0.8	89.2±1.1
AAE Sad	TPR	42.1±0.4	66.2±2.4	59.5±2.0	58.4±0.6	60.7±1.7	63.1±4.1	59.4±0.6	43.5±3.3
SAE Happy	TPR	64.2±0.3	55.9±2.9	81.3±1.5	80.4±1.2	79.7±1.2	80.8±2.9	81.8±0.8	62.5±4.9
AAE Happy	TPR	91.3±0.2	75.5±2.2	80.7±3.9	82.3±0.4	79.8±1.5	78.8±3.3	80.8±0.7	90.0±2.4

Table 21: Disaggregated TPR values for various debiasing methods on Moji with the imbalanced test and validation sets (DeepMoji+MLP model).

Group	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
SAE	TPR	75.5±1.7	73.1±1.5	78.2±1.0	78.8±1.0	78.0±1.1	77.8±1.3	79.4±0.6	72.5±2.7
AAE	TPR	67.2±0.9	69.8±2.5	68.7±0.9	70.9±0.5	69.9±0.8	70.6±0.5	70.8±0.8	64.2±4.3
SAE Sad	TPR	89.7±2.4	92.6±1.5	77.5±9.1	77.4±7.1	73.9±1.9	70.6±6.0	81.3±4.1	77.2±10.6
AAE Sad	TPR	43.9±2.8	73.5±7.0	66.7±10.1	63.2±6.3	64.9±3.8	57.7±5.5	67.3±6.3	59.1±14.6
SAE Happy	TPR	61.3±5.8	53.6±4.4	78.8±8.2	80.3±6.3	82.1±2.1	84.9±4.5	77.4±5.2	67.9±7.6
AAE Happy	TPR	90.6±1.2	66.1±11.7	70.7±11.4	78.7±5.7	75.0±5.0	83.6±5.0	74.2±6.8	69.3±8.0

Table 22: Disaggregated TPR values for various debiasing methods on Moji with the balanced test and validation sets (DeepMoji+MLP model).

Group	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
SAE	TPR	77.4±1.2	82.8±0.8	82.6±0.7	82.9±0.5	79.3±1.7	84.0±1.1	81.6±1.5	71.3±10.5
AAE	TPR	66.2±0.6	70.4±0.4	70.7±1.0	70.9±0.8	68.2±3.0	69.7±0.7	71.0±0.4	60.5±5.3
SAE Sad	TPR	92.5±0.8	85.1±1.9	85.1±2.2	84.5±1.4	80.6±9.7	83.7±1.7	89.2±0.8	84.0±6.8
AAE Sad	TPR	45.8±1.6	66.3±3.1	69.1±2.6	68.0±1.9	63.3±14.8	64.0±3.2	69.7±1.7	50.7±13.3
SAE Happy	TPR	62.3±3.0	80.6±3.3	80.1±3.4	81.4±2.0	78.0±8.4	84.4±3.5	74.0±3.8	58.6±17.1
AAE Happy	TPR	86.6±0.7	74.5±3.7	72.3±3.3	73.8±3.0	73.0±9.6	75.3±2.6	72.4±1.7	70.3±22.3

Table 23: Disaggregated TPR values for various debiasing methods on Moji with the imbalanced test set (BERTweet model).

Group	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
SAE	TPR	77.4±0.8	82.2±0.8	82.3±0.5	82.3±1.7	78.0±1.9	83.9±0.7	81.6±1.0	68.8±10.8
AAE	TPR	67.6±0.2	70.6±0.5	70.3±0.9	70.3±0.7	67.8±3.6	69.9±0.3	71.1±0.5	59.8±7.7
SAE Sad	TPR	92.4±0.6	85.4±2.2	84.6±1.8	81.7±3.2	75.8±10.5	82.4±0.9	88.8±0.8	88.0±8.7
AAE Sad	TPR	48.5±0.4	67.2±3.3	67.8±2.1	65.6±1.3	71.9±12.0	62.6±1.1	69.3±1.4	60.9±21.5
SAE Happy	TPR	62.5±2.0	79.0±3.7	80.1±2.5	82.8±0.8	80.1±10.3	85.3±1.2	74.3±2.8	49.6±25.5
AAE Happy	TPR	86.8±0.4	74.0±4.1	72.8±2.9	75.0±1.8	63.6±18.9	77.2±0.9	72.8±2.2	58.8±33.6

Table 24: Disaggregated TPR values for various debiasing methods on Moji with the balanced test set (BERTweet model).

Group	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
Male	TPR	86.1±0.4	85.6±0.7	86.4±0.3	86.4±0.5	85.5±0.1	85.4±0.4	86.7±0.7	86.0±0.4
Female	TPR	86.3±0.2	85.7±0.5	86.5±0.2	86.5±0.4	86.0±0.2	85.4±0.4	86.6±0.4	86.3±0.3
Male teacher	TPR	93.6±0.4	93.2±0.4	93.2±0.4	93.7±0.9	92.9±0.2	93.8±0.3	92.8±0.8	93.4±0.3
Female teacher	TPR	92.8±0.7	92.8±0.5	93.1±0.2	93.5±0.7	92.2±0.5	93.4±0.5	92.9±0.8	92.9±0.1
Male attorney	TPR	94.5±0.4	94.2±0.6	94.6±0.3	95.0±0.2	93.8±0.6	95.1±0.3	94.1±1.7	94.3±0.5
Female attorney	TPR	97.7±0.2	97.2±0.3	97.7±0.2	97.7±0.1	97.2±0.5	97.8±0.2	97.2±1.0	97.7±0.3
Male photographer	TPR	88.0±1.5	89.2±1.3	89.0±1.7	87.6±2.8	86.6±1.3	88.4±2.7	89.3±0.7	87.2±0.9
Female photographer	TPR	88.7±1.0	89.5±1.5	90.1±1.8	88.6±2.3	87.8±1.2	88.3±3.2	89.7±0.8	87.7±0.5
Male psychologist	TPR	74.5±1.1	77.0±1.3	78.3±1.9	78.3±2.0	75.0±0.5	74.5±1.4	78.9±3.4	75.7±2.0
Female psychologist	TPR	88.3±1.1	83.2±1.8	84.4±1.6	85.1±1.8	88.1±0.9	83.8±1.0	84.7±0.3	88.5±2.1
Male physician	TPR	94.2±0.7	94.0±0.4	94.4±0.4	93.6±0.3	94.3±0.4	94.6±0.8	94.2±0.9	94.2±0.5
Female physician	TPR	91.7±0.8	93.3±0.6	92.9±0.6	92.5±0.7	91.9±0.7	92.8±1.3	93.2±0.5	91.6±0.6
Male surgeon	TPR	91.7±1.1	91.0±1.3	89.8±1.2	90.2±0.6	91.3±0.3	91.9±1.2	90.1±1.4	91.9±0.9
Female surgeon	TPR	98.0±0.2	98.3±0.2	98.2±0.1	98.1±0.2	98.1±0.1	98.4±0.2	98.1±0.2	98.0±0.3
Male journalist	TPR	86.9±0.7	86.9±2.2	87.0±2.3	86.8±1.6	84.7±0.4	87.7±1.5	86.9±1.2	86.6±0.6
Female journalist	TPR	86.9±1.5	86.3±1.7	86.5±2.3	86.6±2.3	85.8±0.8	87.0±1.3	86.6±1.4	86.7±0.9
Male dentist	TPR	77.0±2.2	71.0±5.1	75.9±1.9	76.0±1.8	76.7±1.4	70.1±1.8	77.0±1.3	77.1±1.6
Female dentist	TPR	53.8±0.9	54.2±3.3	58.0±1.9	57.3±2.0	54.5±2.1	49.1±1.7	58.0±2.1	55.1±0.6
Male nurse	TPR	74.4±1.5	74.1±2.2	75.5±2.2	76.6±3.7	74.3±0.7	72.8±3.5	76.6±2.0	73.9±2.0
Female nurse	TPR	78.4±1.5	76.3±1.5	77.5±1.7	78.9±2.5	77.9±0.8	77.6±3.2	79.4±1.8	78.1±1.1

Table 25: Disaggregated TPR values for various debiasing methods on Bios with imbalanced test and validation sets (BERT model).

Group	Metric	Standard	BTEO	Adv	DAdv	FairBatch	GD_{diff}	BTJ	INLP
Male	TPR	86.0±0.3	85.4±0.5	86.4±0.5	86.4±0.6	85.5±0.3	84.6±1.2	86.8±0.4	85.9±0.3
Female	TPR	86.1±0.2	85.3±0.3	86.2±0.2	86.1±0.5	85.6±0.2	84.6±0.9	86.6±0.4	86.2±0.3
Male teacher	TPR	94.5±0.3	93.5±1.4	94.1±0.3	93.9±0.9	93.5±0.4	94.8±0.2	93.2±0.9	94.4±0.3
Female teacher	TPR	92.9±0.8	92.5±1.5	93.3±0.3	93.2±0.8	92.1±0.5	93.6±0.4	92.4±1.0	92.8±0.4
Male attorney	TPR	94.0±0.6	93.6±0.4	93.7±0.4	94.2±0.3	92.6±0.6	94.2±0.4	93.6±0.8	93.9±0.6
Female attorney	TPR	97.7±0.3	97.2±0.3	97.7±0.2	97.6±0.3	97.2±0.3	97.6±0.3	97.4±0.4	97.9±0.3
Male photographer	TPR	87.7±1.4	90.2±1.2	89.2±2.4	88.8±3.7	86.4±0.9	89.6±1.6	89.9±0.8	87.8±0.9
Female photographer	TPR	88.9±1.1	90.0±1.4	89.9±1.8	89.4±2.9	87.7±0.9	89.5±1.5	90.3±1.1	88.6±1.5
Male psychologist	TPR	73.7±1.4	76.2±1.3	76.6±0.9	78.4±1.9	74.5±0.2	74.0±1.8	79.1±4.0	74.2±1.0
Female psychologist	TPR	86.2±2.8	82.2±0.8	82.5±1.2	83.8±2.2	87.7±1.0	82.1±1.8	83.7±2.0	87.4±1.9
Male physician	TPR	95.3±1.0	94.2±1.1	94.9±0.6	94.0±0.4	94.9±0.2	95.4±0.9	94.5±0.9	94.8±0.4
Female physician	TPR	93.0±1.4	93.1±1.2	93.1±0.8	92.4±0.7	92.1±0.7	93.8±1.7	93.4±0.7	92.1±0.9
Male surgeon	TPR	91.9±0.6	91.0±1.4	90.3±0.5	90.6±1.2	91.5±0.6	92.1±1.3	89.2±0.8	91.9±0.4
Female surgeon	TPR	98.2±0.1	98.3±0.3	98.2±0.1	97.9±0.4	97.8±0.3	98.4±0.3	97.9±0.1	98.2±0.2
Male journalist	TPR	86.7±0.5	85.9±1.9	87.2±2.0	86.6±1.8	84.6±0.7	87.7±2.4	86.9±1.4	85.9±0.7
Female journalist	TPR	85.6±1.9	84.5±1.7	85.8±1.7	85.0±2.2	84.6±1.9	86.5±1.9	86.0±1.3	85.0±1.1
Male dentist	TPR	76.6±2.1	70.0±5.8	74.7±3.3	74.5±4.5	77.0±2.9	62.9±8.7	78.1±1.7	76.5±1.0
Female dentist	TPR	53.0±1.1	53.3±2.6	56.3±1.4	56.6±3.4	53.7±2.1	45.5±6.2	58.8±1.9	54.1±1.5
Male nurse	TPR	73.6±2.4	74.0±2.7	76.4±1.8	76.9±3.5	74.2±1.0	70.9±2.9	76.4±1.2	73.9±1.7
Female nurse	TPR	79.0±2.5	76.8±2.2	79.4±2.0	79.2±2.5	77.8±1.1	74.8±3.1	79.1±0.8	79.3±1.7

Table 26: Disaggregated TPR values for various debiasing methods on Bios with balanced test and validation sets (BERT model).