Appendix

1 | DEBIASING APPROACHES

We selected a diverse list of 28 representative debiasing approaches that operate on the training data (*pre-processing*), model (*in-processing*), and predicted outcomes (*post-processing*). These approaches may involve certain assumptions or goals about desired fairness. For example, `Independence` and `Uniform Sampling` both assume the probability of an instance with label Y should be independent from the probability of an instance from group S, i.e., the expected probability of a pair (class label Y, group membership S) should be equal to the product of the probability of an instance having class label Y and the probability of being from group S. We summarize the involved fairness metrics in selected approaches in Table 1.

TABLE 1 Involved Fairness Metrics in Selected Debiasing Approaches.

Metrics Name	Metric Discription	Approach
Demographic Parity/ Statistical Parity/ Discrimination Score	The difference in proportions of positive predictions between two groups	Massaging (Kamiran and Calders, 2012); Preferential Sampling (Kamiran and Calders, 2012); Label Debias-Dp (Jiang and Nachum, 2020); Reduction-Dp Agarwal et al. (2018); FairBatch-Dp (Roh et al., 2021)
Equal Opportunity	The difference in true positive rates between two groups	Label Debias-EqOpp (Jiang and Nachum, 2020); FairBatch-EqOpp (Roh et al., 2021)
Equalized Odds	The difference in error rates (true positive and false positive) between two groups	Label Debias-EqOdds (Jiang and Nachum, 2020); Reduction-EqOdds Agarwal et al. (2018); FairBatch-EqOdds (Roh et al., 2021)
Independence	The expected probability for an instance with class label Y from group S should be equal to the product of the probability of an instance having class label Y and the probability of being from group S.	Independence (Kamiran and Calders, 2012); Uniform Sampling (Kamiran and Calders, 2012)

A detailed overview of selected debiasing approaches in terms of the stages they operate, categories, and how they work was provided in Table 2.

In terms of pre-processing approaches, each source of bias might be addressed in different ways, such as re-sampling or reweighing. We summarize the pre-processing approaches in Table 3.

2 | FAIRNESS METRICS

In this study, we use ABROCA Gardner et al. (2019) to evaluate the predictive fairness. As pointed out in Gardner et al. (2019), both positive and negative predictions may induce different impacts on students, e.g., being predicted to be at risk of dropping out or not may lead to different types or levels of support for students. Therefore, ABROCA measures the difference in overall accuracy across compared groups, i.e., the absolute value of the area between the ROC curves of the compared groups. Mathematically, ABROCA is computed from the predicted probabilities and true labels, i.e., $\int_0^1 |ROC_{g1}(t) - ROC_{g2}(t)|$, where $g1$ and $g2$ are the two groups compared, i.e., the privileged group and unprivileged group.

TABLE 2 Overview of debiasing approaches investigated in this study.

Stage	Category	RowID	Approach	How the Approach Works
First-Phase to tackle *Embedding Bias*		1	DebiasedBERT (Sha et al., 2022)	Continuing training the BERT with additional demographically balanced domain-specific data to remove bias in the embedding
Pre-processing	Balancing data to tackle *Class Imbalance*	2	SMOTE-Class (Chawla et al., 2002)	Oversampling with generated synthetic instances to balance distribution of class labels
		3	Balanced-Class (Han et al., 2022)	Assign weights to instances to equally represent samples with different class labels
		4	CB-Class (Han et al., 2022)	Assign weights to instances to balance class labels within each demographic group
	Balancing data to tackle *Representation Bias*	5	Balanced-Demo (Han et al., 2022)	Assign weights to instances to equally represent samples with different membership
	Balancing data to tackle *Local Stereotypical Bias*	6	Uniform Sampling (Kamiran et al., 2010)	Uniformly re-samples instances to ensure the independency between class labels and group membership
		7	Independence (Kamiran et al., 2010)	Assign weights to instances to ensure the independency between class labels and group membership
		8	SMOTE-Demo (Chawla et al., 2002)	Oversampling with generated synthetic instances to equally represent samples with different group membership
		9	CB-Demo (Han et al., 2022)	Assign weights to instances to ensure equal representation of samples from different demographic groups within each class label.
		10	FairBalance (Yan et al., 2020)	Apply SMOTE to balance class labels within the same cluster as they exhibit similar features and should have balanced class labels.
	Balancing data to tackle *Global Stereotypical Bias*	11	SMOTE-Joint (Chawla et al., 2002)	Oversampling with generated synthetic instances to equally represent samples with different membership and class labels
		12	CB-Joint (Han et al., 2021)	Assign weights to instances to ensure equal representation of samples with different class labels and demographic labels
	Correcting labels to tackle *Label Bias*	13	Massaging (Kamiran et al., 2010)	Change class labels of borderline instances (i.e., those near the decision boundary) to ensure same proportion of positive instances across different groups
		14	Preferential Sampling (Kamiran et al., 2010)	Preferentially duplicates or removes borderline instances
		15	Label Debias-Dp (Jiang and Nachum, 2020)	Assign weights to instances during training process based on the amount of bias measured by demographic parity
		16	Label Debias-EqOpp (Jiang and Nachum, 2020)	Assign weights to instances during training process based on the amount of bias measured by equal opportunity
		17	Label Debias-EqOdds (Jiang and Nachum, 2020)	Assign weights to instances during training process based on the amount of bias measured by equalized odds
	Transforming features to tackle *Proxy Discrimination*	18	Correlation Remover (Bird et al., 2020)	Learn a new feature representation similar to the original but orthogonal to sensitive attributes
In-processing	Constraints	19	Reduction-Dp (Agarwal et al., 2018)	Constrain the model with demographic parity
		20	Reduction-EqOdds (Agarwal et al., 2018)	Constrain the model with equalized odds
	Optimization	21	FairBatch-Dp (Roh et al., 2021)	Adjusting the batch sizes w.r.t. sensitive groups based on demographic parity for each training epoch
		22	FairBatch-EqOpp (Roh et al., 2021)	Adjusting the batch sizes w.r.t. sensitive groups based on equalized opportunity for each training epoch
		23	FairBatch-EqOdds (Roh et al., 2021)	Adjusting the batch sizes w.r.t. sensitive groups based on equalized odds for each training epoch
	Adversarial Leaning	24	Adversarial Debiasing (Zhang et al., 2018)	Reduce the sensitive information encoded in the trained model can mitigate unfairness
Post-processing	Score Transformation	25	CalibratedEqOdds-fpr (Pleiss et al., 2017)	Occasionally return the group's mean probability for a randomly chosen subset of the group to ensure equal false positive rates
		26	CalibratedEqOdds-fnr (Pleiss et al., 2017)	Occasionally return the group's mean probability for a randomly chosen subset of the group to ensure equal false negative rates
		27	CalibratedEqOdds-weighted (Pleiss et al., 2017)	Occasionally return the group's mean probability for a randomly chosen subset of the group to ensure equal false positive rates
	Optimization	28	Fair-Projection (Alghamdi et al., 2022)	Project a trained probabilistic classifier onto other classifiers satisfying target fairness constraint by solving a meta optimization problem

TABLE 3 Selected pre-processing approaches according to their assumptions and ways of manipulating data.

	Data Balancing	Label Correction	Feature Blinding
Relabeling/ Transformation	-	Massaging (Kamiran et al., 2010)	Correlation Remover (Bird et al., 2020)
Re-sampling	Uniform Sampling (Kamiran et al., 2010); SMOTE-Demo/Class/Joint (Chawla et al., 2002) FairBalance (Yan et al., 2020);	Preferential Sampling (Kamiran et al., 2010)	-
Reweighing	Independence (Kamiran et al., 2010); Balanced-Demo/Class (Han et al., 2022); CB-Demo/Class (Han et al., 2022); Joint Balance (Han et al., 2022)	Label-bias (Jiang and Nachum, 2020)	-

3 | EXPERIMENT SET-UP

3.1 | Training Details

Forum Post Classification. For the testbed, we first extract the text embedding of each post using BERT without fine-tuning, which would be then fed up as the input features to the Logistic Regression. For model training in all cases, we set the *batch size* as 32, and *learning rate* as 0.001. For all debiasing approaches, we select the best model according to the performance on the validation set, i.e., the model with minimum validation loss.

STEM Career Prediction. We adopted the model[1] open-sourced by the authors of the original paper Yeung and Yeung (2019).

3.1.1 | Hyperparameter Tuning

Pre-processing approaches that require parameter tuning include `FairBalance`, `Preferential Sampling`, `Massaging`, and `Debiasing BERT`, we detailed their parameters tuning as below:

- For `FairBalance`, we selected the clustering algorithms following original paper from the list *kmeans*, *agglomerative clustering*, and *spectral clustering*.
- For the `Debiasing BERT` we set the max iteration of continue training as 12, and for each iteration we use the validation set as the reference set to determine the number of instances to be sampled for each subgroup.
- For `Preferential Sampling`, we set the ranker, i.e., the classifier to identify borderline instances, as the logistic regression.
- For `Massaging`, the ranker is same as `Preferential Sampling`, i.e., logistic regression.
- For `Correlation Remover`, we tune the parameter α from $[0.1, 0.2, 0.3, ..., 0.9, 1]$.

For in-processing approaches, we detailed the parameter settings as below:

- For `Reduction`, we select the best grid size from $\{10, 15, 20, 25, 30\}$ for each variant using parameterized with demographic parity and equalized odds.
- For `FairBatch` we follow the original paper to select α, i.e., the learning rate of FairBatch, from $[1e-4, 3e-4, 5e-4, 7e-4, 1e-3, .., 0.03, 0.05]$, for each variant using parameterized with demographic parity, equal opportunity, and equalized odds.

[1] `https://github.com/ckyeungac/ADM2017`

- For `Adversarial Debiasing`, we select the adversary weight within the range $[0.1, 1]$ with a step size as 0.1.

  In terms of post-processing approaches, we detailed the parameter settings as below:

- For `CalibratedEqOdds`, we follow the original paper to select the error rate from *false positive rate*, *false negative rate*, and *weighted*.
- For `FairProjection`, we set the cross-entropy as the divergence function and follow the original paper to use mean equalized odds as the fairness constraint.

3.1.2 | Implementation Details

All the methods were implemented and trained using Tensorflow. For each debiasing approach, we check if they were publicly available and open-sourced, if so we adapted the code into our evaluation framework. Specifically, for approaches `Reduction` and `CalibratedEqOdds`, we used the implementations provided by fairlearn[2] Bird et al. (2020) and aif360[3] Bellamy et al. (2019). For re-sampling approach `SMOTE` we used the implementation from Lemaître et al. (2017).

references

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. and Wallach, H. (2018) A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P., Asoodeh, S. and Calmon, F. (2022) Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, **35**, 38747–38760.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. et al. (2019) Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, **63**, 4–1.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H. and Walker, K. (2020) Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002) Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, **16**, 321–357.

Gardner, J., Brooks, C. and Baker, R. (2019) Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*, 225–234.

Han, X., Baldwin, T. and Cohn, T. (2021) Balancing out bias: Achieving fairness through balanced training. *arXiv preprint arXiv:2109.08253*.

— (2022) Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11335–11350. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Jiang, H. and Nachum, O. (2020) Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, 702–712. PMLR.

[2] `https://fairlearn.org/v0.8/api_reference/fairlearn.reductions.html#fairlearn.reductions.GridSearch`

[3] `https://aif360.readthedocs.io/en/latest/modules/generated/aif360.sklearn.postprocessing.CalibratedEqualizedOdds.html`

Kamiran, F. and Calders, T. (2012) Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, **33**, 1–33.

Kamiran, F., Calders, T. and Pechenizkiy, M. (2010) Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, 869–874. IEEE.

Lemaître, G., Nogueira, F. and Aridas, C. K. (2017) Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, **18**, 559–563.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. and Weinberger, K. Q. (2017) On fairness and calibration. *Advances in neural information processing systems*, **30**.

Roh, Y., Lee, K., Whang, S. E. and Suh, C. (2021) Fairbatch: Batch selection for model fairness. In *9th International Conference on Learning Representations*. The International Conference on Learning Representations.

Sha, L., Li, Y., Gasevic, D. and Chen, G. (2022) Bigger data or fairer data? augmenting bert via active sampling for educational text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1275–1285.

Yan, S., Kao, H.-t. and Ferrara, E. (2020) Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1715–1724.

Yeung, C.-K. and Yeung, D.-Y. (2019) Incorporating features learned by an enhanced deep knowledge tracing model for stem/non-stem job prediction. *International Journal of Artificial Intelligence in Education*, **29**, 317–341.

Zhang, B. H., Lemoine, B. and Mitchell, M. (2018) Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.