Fair Classification via Transformer Neural Networks: Case Study of an Educational Domain

Modar Sulaiman
University of Tartu
Tartu, Estonia
modar.sulaiman@ut.ee

Kallol Roy
University of Tartu
Tartu, Estonia
kallol.roy@ut.ee

ABSTRACT

Educational technologies nowadays increasingly use data and Machine Learning (ML) models. This gives the students, instructors, and administrators support and insights for the optimum policy. However, it is well acknowledged that ML models are subject to bias, which raises concern about the fairness, bias, and discrimination of using these automated ML algorithms in education and its unintended and unforeseen negative consequences. The contribution of bias during the decision-making comes from datasets used for training ML models and the model architecture. This paper presents a preliminary investigation of fairness constraint in transformer neural networks on Law School and Student-Mathematics datasets. The used transformer models transform these raw datasets into a richer representation space of natural language processing (NLP) while solving fairness classification. We have employed fairness metrics for evaluation and check the trade-off between fairness and accuracy. We have reported the various metrics of F1, SPD, EOD, and accuracy for different architectures from the transformer model class.

Keywords

Fairness; Classification; Educational Data; Bias; Representation Learning; Machine Learning; Transformer

1. INTRODUCTION

Automated decision-making with ML models in education is increasingly used to aid and support teachers, educators, and other stakeholders for optimal policy formulation. Though this method holds immense potential to improve accuracy in prediction, the outcomes of the ML models show unfair disadvantage to some sections (e.g., under represented) of the society. For example, the ML models show unfair approval for student loans for African students or predict lower bar exam success of students of low socio-economic group [3]. In general, the notion of unfairness in ML broadly categorized as follow, (i) Disparate treatment: where the ML model classifies differently (unfairly) people with the same values of non-sensitive features but different values of sensitive features, (ii) Disparate impact: where the ML model classifies that benefits (or hurts) people who are sharing the value of a sensitive feature vector more frequently than the other group, (iii) Disparate mistreatment: where the ML model achieves different classification accuracy for groups of people sharing different values of a sensitive feature.

We investigated mainly in this paper the fairness in predictions using transformer models in tabular data in the educational domain. We used the non-textual tabular data to train the transformer neural network with a bias mitigation method to achieve fairness in the classification tasks. The representation of non-textual data in textual representation space via the transformer model is one of the main strengths of the transformer models. It follows the distributional hypothesis: a word is characterized by the company it keeps.

In this work, we showed that although there exists a transformer model, namely the SAINT model, which achieves perfect group fairness without requiring any explicit debiasing method, one of the transformer models, (Tab) Tab-Transformer model (with fairness constraint), we used in our work improves fairness for protected groups at a negligible cost in terms of accuracy compared to the other models in the Law School dataset. Additionally, we show a case using the Student-Mathematics dataset, where we do not recommend using the transformer models to mitigate the bias in the data. Finally, we demonstrate the possibility of empirically achieving a slight trade-off between the performance and fairness using transformer models.

To the best of our knowledge, our work is the first study that considers the fairness constraint via transformer-based models for tabular data in the educational and other domains.

2. RELATED WORK

A comprehensive studies of algorithmic fairness in education is reported in [10]). They have investigated how discrimination emerges in automated systems and how it can be mitigated through studying and calibrating: the measurement of input data, model learning, and output presentation. A comparative study of different fairness algorithms on multiple datasets are reported [4]).

The Law School Admission Council (LSAC) National Longi-

tudinal Study reported discrimination against Africans (and other minorities) examinees during the bar passage examinations[15]). The recent thrusts of using transformer models on tabular data is reported [6, 8, 13]. Flurry of interesting and insight works on fairness constraints are investigated recently by various research groups[16, 12, 1].

3. FAIR CLASSIFICATION THEORY

The binary classification in educational data can be formulated as finding the optimum mapping function $\mathcal{T}(x)$ from feature vectors $x \in \mathcal{R}^d$ and class labels $y \in (0, 1)$. For example, the label $y = 1$ could represents the application loan is getting accepted and $y = 0$ otherwise, where x represents the input features. This decision boundary-based classifier $\mathcal{T}(x)$ predicts a label y, by a set of optimum parameters of neural weights w, by minimizing a cross-entropy loss function:

$$\mathcal{L}(\mathcal{T}(x) = \hat{y}, y) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (1)$$

where $\hat{y}$ is the predicted label. The fairness classification comes with the additional constraints for fairness guarantees of a sensitive feature (race, color) z [1, 16]. The fairness constrained optimization can be expressed as the following:

$$\begin{aligned} &\arg\min_{w} \mathcal{L}_w \\ &\text{subject to} \quad P(.|z = 0) = P(.|z = 1) \end{aligned} \quad (2)$$

where the constraint in (2) can be formulated as a condition of *no disparate treatment* (3), *no disparate impact* (4) or *no disparate mistreatment* (5, 6, 7) as shown below.

$$P(\hat{y}|x, z) = P(\hat{y}|x) \quad (3)$$

$$P(\hat{y} = 1|z = 0) = P(\hat{y}|z = 1) \quad (4)$$

$$P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1) \quad (5)$$

$$P(\hat{y} \neq y|y = -1, z = 0) = P(\hat{y} \neq y|y = -1, z = 1) \quad (6)$$

$$P(\hat{y} \neq y|y = 1, z = 0) = P(\hat{y} \neq y|y = 1, z = 1) \quad (7)$$

We consider in our experiments only the condition of *no disparate treatment* (3) as the fairness constraint.

4. PROPOSED MODELS

This section highlights the models that we used in our work. Most of the models are adaptation of the Transformer architecture [14] and we used them for fair classification in educational data. However, finding an optimum fair classifier defined by2 is a non-trivial problem. The desired fair classifier (satisfying the constraints) may be of non-convex boundary-based type and thus finding the optimum weights w in some cases is hard. The first transformer-based model that is explicitly designed for tabular data is Tab-Transformer (Tab) model [8]. Tab model uses attention to embed only categorical features in the tabular data. A very similar model to the (Tab) model is FT-Transformer (Feature Tokenizer + Transformer) model [6]. FT-Transformer (FT) model is also designed specifically for tabular data. However, FT-Transformer transforms all features (categorical and numerical) to embeddings where it is projecting the continuous features into a $d-$dimensional space before passing their embedding in the transformer model. Our paper shows that the most critical transformer-based model regarding fairness in the tabular dataset is SAINT model [13], because of the way it works across the samples. The architecture of the SAINT model itself plays the role of a regularizer for fairness where it has a novel intersample attention block that computes attention across samples. Consequently, the SAINT model implicitly satisfies the condition (3) and thus, does not need to add these constraints in the loss function in some cases. The last transformer-based model we use is Perceiver [9]. Perceiver model is designed to be architecture agnostic of the nature of the input data. It handles arbitrary configurations of different modalities (audio, images, and text). In our experiment we check the performance and fairness of Perceiver model in tabular data modality. In addition to the previous transformer-based models, we use one of the classical ML models, Logistic Regression (LR) model. LR is one of the common statistical analysis methods to predict a binary outcome and used in different works when studying fairness in ML.

5. DATASETS AND FAIRNESS METRICS

In this section we describe the datasets for our experiments and define the fairness metrics used in this paper.

5.1 Datasets

We have used the two datasets of Law School (LSAC) and Student-Mathematics in our experiments. For both datasets, we consider using the same features and processing[1] that are used to conduct the experiments in the survey [11]. Law school dataset was developed by a Law School Admission Council (LSAC) survey across 163 law schools in the United States in 1991 [15]. The dataset is investigated in a variety of studies and is currently hosted in the database of Project SEAPHE[2]. After cleaning and processing the law school data, we got 20,798 samples of students. We used 12 attributes (3 categorical, 3 binary and 6 numerical attributes) for each sample in experiments[11]. The model is used for predicting whether an examiner would pass the bar exam. We label 'Non-white' to be the minority and unprivileged group in our experiments, 16% of the dataset. For the Student-Mathematics data set[3], the task is to predict the performance of secondary school student in mathematics subject[2]. The dataset has 395 samples of students and 33 features. The majority and privileged group in the dataset is labelled as female students, 52.7 % of the dataset.

5.2 Fairness Metrics

We have used different fairness metrics of Absolute Between-ROC Area (ABROCA) [5], Equal Opportunity Difference (EOD) and Statistical Parity Difference (SPD). We interpret more unfair a model is if, bigger is the difference of fairness metric value between the subgroups. ABROCA is based on the Receiver Operating Characteristics (ROC) curve while EOD measures the difference of True Positive Rates (TPR) for unprivileged group and privileged groups and finally SPD measures the difference between probability of unprivileged group gets favorable prediction and probability of privileged group gets favorable prediction.

[1] https://github.com/tailequy/fairness-dataset/tree/main/experiments

[2] http://www.seaphe.org/databases.php

[3] https://archive.ics.uci.edu/ml/datasets/student+performance

6. EXPERIMENTS AND RESULTS

We trained the models on Law school and Student-Mathematics datasets. The training and test data were split randomly. Our results show trade-offs between fairness criteria and the performance of transformer-based models. Table 1 shows that all the transformer-based models (without fairness constraint) have comparable performance on the Law School dataset. However, the (Tab) model (without fairness constraint) has shown marginally better performance scores on the Law School dataset than other transformer models, with an accuracy of 0.90016 and F1-score of 0.94664. Table 1 has shown LR model has better results in performance than Tab-Transformer (Tab) model by a very small margin, with an accuracy improvement by +0.00705 and F1-score improvement by +0.0032. Table 1 and Figure **??** shows SAINT transformer model outperforms other models in fairness metrics without requiring any explicit debiasing method. This ensures that candidates are treated equally based on different races (protected attributes in the Law School dataset) when predicting whether the candidate would pass the bar exam. In short, the SAINT model comes out as an ideal candidate for correctly identifying successful students at equal rates for different race subgroups. However, transformer models are nonlinear with a large number of model parameters. This requires a large enough dataset to train [7] for achieving a good performance. The number of training samples from Student-Mathematics (only 395 samples) is thus not sufficient enough for training these transformer neural networks. Table 1 shows that the logistic regression (LR) model outperforms the complex SAINT model on the Student-Mathematics dataset, with an accuracy of 0.93277 and F1-score of 0.91111. Nevertheless, same as in the Law School dataset, the SAINT model outperforms the LR model based on the fairness metrics (SPD, EOD) in the Student-Mathematics dataset without requiring an explicit debiasing technique. Using the mentioned bias mitigation method in Section 3, Table 2 and Figure **??** show that the models with fairness constraint successfully limit the bias in the Law School dataset at a minimal cost in term of performance for most of the transformer models. Intriguingly, the (Tab) model with fairness constraint shows better performance and fairness than the other transformer models, with an accuracy of 0.94371, F1-score of 0.89342, and perfect group fairness. Furthermore, the (Tab) model (with fairness constraint) outperformed the SAINT model (without fairness constraint) in terms of fairness and accuracy. Even though the SAINT model showed its usefulness in achieving perfect group fairness without requiring any explicit debiasing method, the (Tab) model (with fairness constraint) outperformed the SAINT model (without fairness constraint) in terms of fairness and accuracy. Additionally, whilst the LR model hardly shows a noticeable better performance than the other transformer models, Figure **??** indicates that LR model tends to fail to converge to a complete fair solution in the Law School dataset, ABROCA of 0.171 without fairness constraint, and ABROCA of 0.0889 with fairness constraint.

In general, previous results in Table 1 and Table 2 show that when we use a large enough dataset, such as the Law School dataset, there is a slight trade-off between the performance and fairness in the transformer models (with fairness constraint). Additionally, Table 1 indicates that the SAINT model approximates the ideal distribution of the given dataset, which has a negligible accuracy-fairness trade-off.

Dataset: Law School. Protected attribute: race.				
Model	F1	Accuracy	SPD	EOD
LR	**0.94984**	**0.90721**	0.189538	0.082670
FT	0.94504	0.89839	-0.215906	-0.124452
Tab	0.94664	0.90016	-0.112048	-0.049809
Perceiver	0.94590	0.89919	-0.151387	-0.081128
SAINT	0.94299	0.89214	**0**	**0**
Dataset: Student-Mathematics. Protected attribute: sex.				
Model	F1	Accuracy	SPD	EOD
LR	**0.91111**	**0.93277**	0.153193	-0.005847
SAINT	0.76041	0.61344	**0**	**0**

Table 1: Results of each model without applying any fairness constraint.

Model	F1	Acc	SPD	EOD
LR	**0.94643**	**0.89983**	-0.100764	-0.043942
FT	0.94305	0.89230	0.001328	0.000617
Tab	0.94371	0.89342	**0**	**0**
Perceiver	0.94237	0.89102	**0**	**0**
SAINT	0.94012	0.88701	**0**	**0**

Table 2: Results of each model with fairness constraints on Law School. Protected attribute: race.

7. CONCLUSION

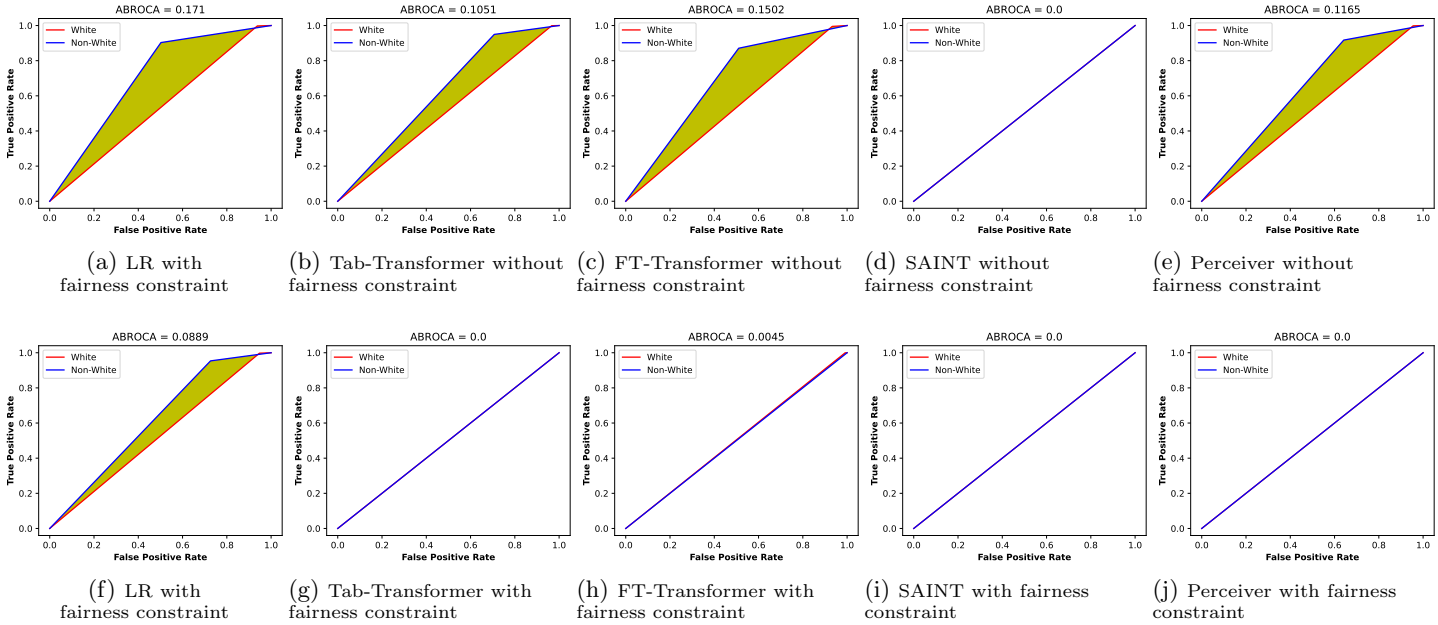
Improving fairness during training the model and preserving its sound performance is challenging in ML. The prior works have focused on enhancing fairness using classical ML models. Still, as we show, transformer models have advantages when we use them to study fairness in the educational domain. Our critical insight is that when trying to improve the fairness in final results using transformer models, it is valuable to check if the dataset is large enough and compatible with the number of the parameters in the transformer model. Additionally, we find that the (Tab) model (with fairness constraint) improves fairness for protected groups at a negligible cost in terms of accuracy compared to the other models in the Law School dataset. Consequently, we believe that this insight and the use of transformer models in the fairness domain provide a foundation for pursuing the fairness of artificial intelligence in the educational field and other areas. For our future work, we will comprehensively study the empirical performance of the transformer models for fair classification in different datasets. Additionally, there is a need to solve the problem of training the transformer models using a small dataset. Therefore, employing the idea of transfer learning can overcome such a problem.

ACKNOWLEDGMENTS

This work has received funding by the European Social Fund via IT Academy programme.

8. REFERENCES

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



(a) LR with fairness constraint

(b) Tab-Transformer without fairness constraint

(c) FT-Transformer without fairness constraint

(d) SAINT without fairness constraint

(e) Perceiver without fairness constraint

(f) LR with fairness constraint

(g) Tab-Transformer with fairness constraint

(h) FT-Transformer with fairness constraint

(i) SAINT with fairness constraint

(j) Perceiver with fairness constraint

Figure 1: ABROCA of each model on Law School Data. Figures (a, b, c, d, e) show ABROCA of each model without applying fairness constraints. Figures (f, g, h, i, j) show ABROCA of each model when applying fairness constraints.

[2] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. 2008.

[3] M. Fei and D.-Y. Yeung. Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 256–263. IEEE, 2015.

[4] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning, 2018.

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

[6] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[8] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

[9] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pages 4651–4664. PMLR, 2021.

[10] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. *CoRR*, abs/2007.05443, 2020.

[11] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, Mar. 2022.

[12] S. Liu and L. N. Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *arXiv preprint arXiv:2008.01132*, 2020.

[13] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] L. F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[16] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.