

Breaking Bias: How Optimal Transport Can Help to Tackle Gender Biases in NLP Based Job Recommendation Systems?

Fanny Jourdan^{1,2}, Titon Tshiongo-Kaninku^{1,3}, Nicholas Asher², Jean Michel Loubes¹ and Laurent Risser¹

¹*Institut de Mathématiques de Toulouse (UMR 5219), CNRS, Université de Toulouse, F-31062 Toulouse, France*

²*Institut de Recherche en Informatique de Toulouse (UMR 5505), CNRS, Université de Toulouse, F-31062 Toulouse, France*

³*AKKODIS Group-Hauts de France, F-59700 Lille, France*

Abstract

Automatic recommendation systems based on complex machine learning models, such as deep neural networks, have become popular during the last decade. Some of these systems, for instance those dedicated to online advertising, have a relatively limited impact on the users' life. However, such systems can also be used for applications which are ranked as High Risk by the European Commission in the A.I. act. Our contribution focuses on automatic job recommendation systems, which fall into this category of applications. We specifically work on the *Bios* dataset, for which the learning task consists in predicting the occupation of female and male individuals, based on their LinkedIn biography. This dataset therefore allows us to study the properties of NLP based job recommendation strategies in terms of gender biases. We first extend with a state-of-the-art Deep Neural-Network model existing experiments showing that the accuracy of trained decision rules may be significantly different for females and males looking for job opportunities in specific fields. Our main contribution is then to adapt a mathematically-grounded optimal transport strategy to ensure that the gender gaps are reasonable for all job categories which can be recommended. We finally show the effectiveness of our strategy on the *Bios* dataset.


Artificial intelligence (A.I.) is a field that has seen impressive growth over the past decade. In particular, most state-of-the-art Natural Language Processing (NLP) applications for translation or recommendations based on texts are now based on the use of Deep Neural Networks (DNNs). We focus here on recommendation systems, which recommend job candidates based on textual information in personal profiles that are held in social networks. Importantly, the so-called *A.I. act* ranks these applications as *High Risk*. The E.U. will therefore constrain such AI systems to have proper statistical properties with regard to potential discrimination they could engender if they are sold in or from the European Union (see articles 9.7, 10.2, 10.3 and 71.3). Being able to train decision rules with limited discrimination biases is therefore a critical concern for companies willing to put such A.I. systems on the E.U. market. On another hand, the field of NLP has also drastically changed in recent years with the now ubiquitous use of DNNs using transformer block layers [1]. The two most popular of transformer neural network architectures are now BERT [2] and GPT [3], and there are many variants of these models. These models

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

✉ fanny.jourdan@irit.fr (F. Jourdan); titon.tshiongo@math.univ-toulouse.fr (T. Tshiongo-Kaninku); nicholas.asher@irit.fr (N. Asher); loubes@math.univ-toulouse.fr (J. M. Loubes); laurent.risser@math.univ-toulouse.fr (L. Risser)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

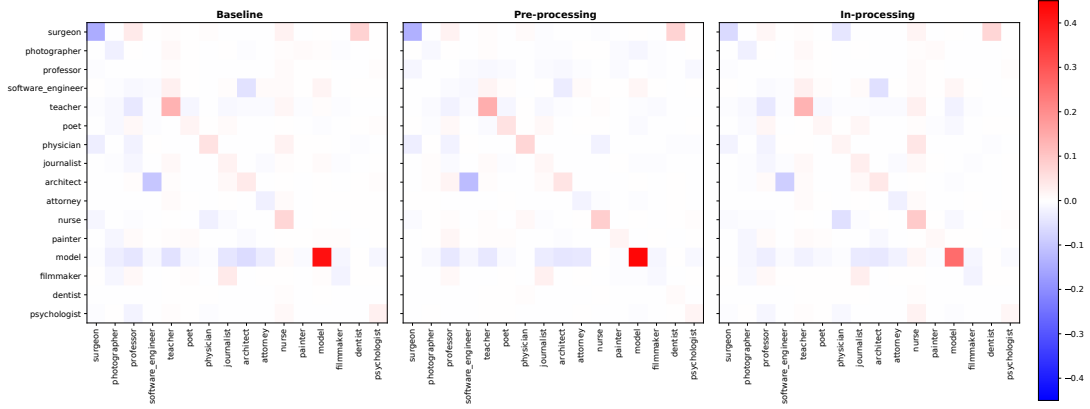


Figure 1: Average difference between the confusion matrices of occupation predictions for males and females. The baseline prediction model is the RoBERTa base model [6]. The pre-processing de-biasing method consists in removing explicit gender indicators in the biographies. The in-processing method consists in reducing the True Positive gender gaps using the W2reg penalty term of [7] when training the RoBERTa model.

have a much higher performance than their predecessors for NLP applications, in particular the LSTM models [4], but they are even less explainable, because of their complexity. A corollary of this gain of complexity, is that ensuring that these models learn decision rules that are free of discrimination biases requires the use of more and more advanced technical solutions.

In order to study the efficiency of bias discrimination techniques in NLP applications, the authors of [5] released the *Bios* dataset, which contains about 400K biographies (textual data), as well as the gender and the occupation (among 28 occupations) of its authors. The authors of [5] also show that training simple prediction models on this dataset leads to large True Positive Rate gender gap (TPRgp) when predicting some occupations. This means that for these occupations, an automatic job candidate recommendation system will clearly favor male (e.g. for Surgeon) or female (e.g. for Model) candidates. They also show that removing the impact of explicit gender indicators (e.g. he, she, her, ...) when making the predictions reduces the discrimination biases, but this effect is limited.

In the present work, we first reproduce these experiments using the RoBERTa base model [6]. This model was pretrained on a massive dataset (see [6]), and we specialized its predictions on the *Bios* dataset. To do so, we used 5 epochs with a batch size of 32 observations and a sequence length of 512 words. The optimizer was Adam with a learning rate of $1e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e6$. We performed five runs to evaluate the training procedure stability, and we split the dataset in 70% for train, 10% for validation and 20% for test. In order to assess whether removing the impact of explicit gender indicators with RoBERTa would have an impact, we reproduced the same protocol with a neutral *Bios* dataset. Note that we did not remove explicit gender indicators as in [5] because BERT models are sensitive to sentence structures. Our key contribution was finally to extend the bias mitigation method of [8], dedicated to binary classification on image, to multi-class classification based on NLP data.

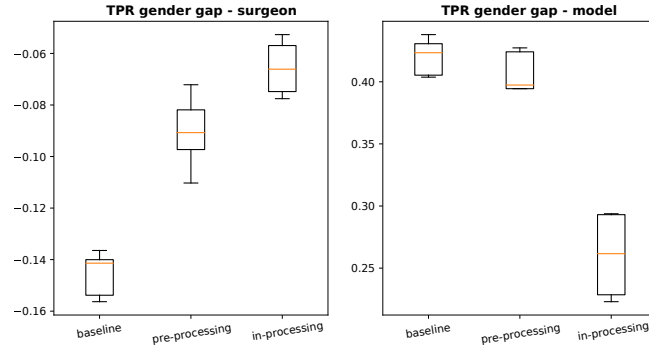


Figure 2: Detailed boxplots for the occupations which are originally the most impacted by gender gaps. Notations for the prediction models are the same as in Figure 1.

The mathematical contribution of [8] was to show how to compute pseudo-gradients of the Wasserstein-2 distance optimal transport metric in a mini-batch context, making it possible to mitigate undesirable algorithmic biases. This strategy was however not dedicated to multi-class classification. We then extended this solution by not only modifying the loss definition, but also by solving different technical locks related to (1) the fact it requires to solve a multivariate optimal transport problem, and (2) that the *Bios* dataset contains biographies related to extremely unbalanced occupations. After having extended the regularization strategy of [8], we used it to mitigate the gender biases considered as unacceptable (occupations with $> 10\%$ gender gap). Results are shown in Figures 1 and fig2. They first confirm that removing the gender information only slightly reduced the gender gaps for the occupations where it was particularly discriminatory. Using our optimal transport strategy has however made these biases much more reasonable. Importantly, the confusion matrices of Figure 1 also shed light on the fact that this bias reduction strategy additionally did not generate unacceptable gender gaps for other occupations. Our strategy is therefore likely to make the trained model certifiable following the A.I. act principles. More details about the multi-class extension of [8] and its application to NLP data can be found in [7]. Our regularization method is also freely available at the following address <https://github.com/lrisser/W2reg>.

Acknowledgements

This research was funded by the AI (Artificial Intelligence) Interdisciplinary Institute ANITI (Artificial and Natural Intelligence Institute.), which is funded by the French ‘Investing for the Future– PIA3’ program under the Grant agreement ANR-19-PI3A-0004. Titon Tshiongo Kaninku was funded by plan France Relance under Grant Agreement ANR-21-PRRD-0018.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [3] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).
- [4] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, *Advances in neural information processing systems* 27 (2014).
- [5] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. T. Kalai, 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*, 2019, pp. 120–128.
- [6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, *CoRR abs/1907.11692* (2019). URL: <http://arxiv.org/abs/1907.11692>. *arXiv:1907.11692*.
- [7] F. Jourdan, T. Tshiongo Kaninku, N. Asher, J.-M. Loubes, L. Risser, How optimal transport can tackle gender biases in multi-class neural-network classifiers for job recommendations?, *Algorithms* 16 (2023).
- [8] L. Risser, A. G. Sanz, Q. Vincenot, J.-M. Loubes, Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization, *Journal of Mathematical Imaging and Vision* (2022) 1–18.