Is a Fairness Metric Score Enough to Assess Discrimination Biases in Machine Learning?

Fanny Jourdan[1,2], Ronan Pons[3], Nicholas Asher[2], Jean Michel Loubes[1] and Laurent Risser[1]

[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] CNRS, Université de Toulouse, F-31062 Toulouse, France

Abstract

We present novel experiments shedding light on a potential limitation of common fairness metrics for assessing the undesirable biases made by machine learning algorithms. Our experiments are based on the *Bios* dataset, for which the learning task consists in predicting the occupation of individuals based on their LinkedIn biography. This dataset is then particularly suited to reproduce Natural Language Processing (NLP) solutions dedicated to automatic job recommendation, which was identified as a High-Risk application of A.I. in the A.I. act. We specifically address an important limitation of theoretical discussions dealing with group-wise fairness metrics in the machine learning literature: they focus on large datasets, although the norm in many commercial A.I. applications is to use reasonably small training and test sets. Data annotation, which is mandatory in supervised learning, is indeed a time consuming and costly task. It is therefore common practice to stop annotating the training data when they reach a sufficiently large size to get a desired level of prediction accuracy. This is typically the case when using active learning procedures, which have become very popular recently. We then question how reliable are different measures of bias when the size of the training and the test set is simply sufficient to learn reasonably accurate predictions. We conclude our study by emphasizing the crucial need to take into account the stability of the bias metrics for small variations of the test set when auditing high-risk A.I. systems.

1. Introduction

Potential biases introduced by Artificial Intelligence (AI) systems are now both an academic concern and a critical issue for industry, as the European Commission plans to regulate AI systems that could adversely affect individual users. The *AI act*[1] will indeed require AI systems sold in the European Union to have proper statistical properties with regard to any potential discrimination they could engender. In particular, AI systems that exploit linguistic data for automatic job recommendation fall into the category of high risks systems in the AI act and will be tightly regulated. In this context, it will be necessary to define fairness metrics to quantify the level of fairness of prediction models. We see two main problems with this: (1) Each fairness metric measures the bias in a certain way and not all metrics are compatible with each other,

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

fanny.jourdan@irit.fr (F. Jourdan); ronan.pons@ut-capitole.fr (R. Pons); nicholas.asher@irit.fr (N. Asher); loubes@math.univ-toulouse.fr (J. M. Loubes); laurent.risser@math.univ-toulouse.fr (L. Risser)

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

which was already discussed in [1, 2, 3], among others. (2) The state of the art of fairness focuses on large datasets. However, the norm in many industrial applications, in particular in Natural Langage Processing (NLP), is to use small linguistic datasets [4].

2. Experimental protocol

To explore the second question of introduction, we used a new experimental protocol developed hereafter: We fine-tuned the DistilBERT [5] neural-network model for automatic job predictions on the biographies of the *Bios* dataset [6]. This dataset contains about 400K biographies (textual data). For each biography, *Bios* specifies the gender and the occupation (28 classes) of its author. Although our model was trained to predict the authors' occupations out of the 28 possible choices, we focus in our study, on the analysis of these biases on two specific occupations: *Surgeon* (many less (15%) females than males) or *Physician* (well balanced between males and females) versus the 27 remaining occupations. To measure the impact of the training set size, We randomly sampled 50 different training sets containing 10K, 20K, 50K, and 120K biographies. We trained a model on each of these 200 samples. Each of these models has the same architectures and the same hyper-parameters. To guarantee the representativeness of the sample, we ensured that each sample had the same percentage of each gender for each occupation as in the initial data set. For the split between the train and test sets, we respectively used 70% and 30% of the dataset.

Let $\hat{Y}$ and Y be the predicted and the true target occupations, respectively. Let G be a random variable representing the binary gender of the biography's subject. For each model, we quantified the gender bias by using Group Parity $GP_{g,y} = P(\hat{Y} = y|G = g)$, True Positive Rate $TPR_{g,y} = P(\hat{Y} = y|G = g, Y = y)$ and Predictive parity $PP_{g,y} = P(Y = y|\hat{Y} = y, G = g)$. To measure the gender gap with these metrics, we computed the difference between binary genders g and $\tilde{g}$ — for each occupation y: $M\_Gap_{g,y} = M_{g,y} - M_{\tilde{g},y}$, where M is GP, TPR or PP.

3. Results

All the models we trained reached a prediction accuracy ranging from 0.72 to 0.86, as in [6], which we consider as good since 28 different occupations are distinguished.

3.1. Results on small data samples

Although the model is trained to predict the occupations of bio authors from the 28 possible choices, we focus, in our study, on the analysis of the biases on two specific occupations: *Surgeon* versus the 27 remaining occupations, and *Physician* versus the other occupations. We chose these professions so that we could compare an occupation with an imbalanced gender distribution and one with balanced a gender distribution.

Our experiments clearly show that the lower the amount of observations in the training set, the more the fairness metrics vary on the test set. The samples with 10K and 20K observations present particularly unstable biases. For example, most TPR (resp. GP) Gender Gaps are negative

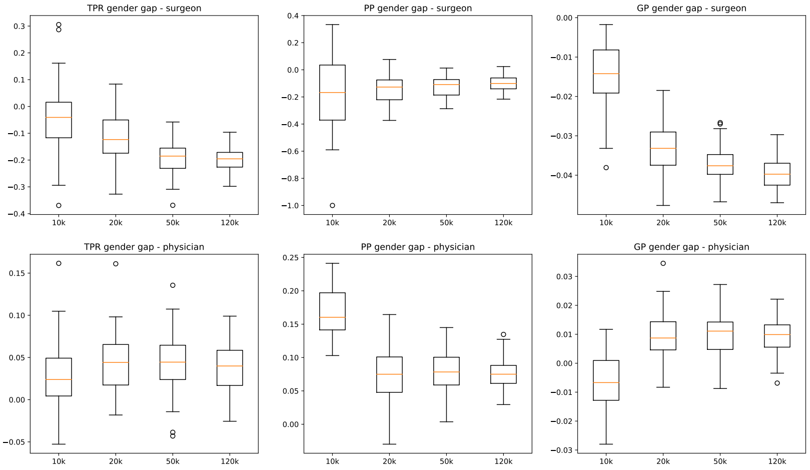

Figure 1: Boxplots of the gender gaps obtained using 10K, 20K, 50K, and 120K randomly sampled observations (50). **(from Left to Right)** True Positive Rate (TPR), Predictive Parity (PP) and Group Parity (GP) gender gaps for **(Top)** surgeons and **(Bottom)** physicians.

(resp. positive) for *surgeon* (resp. *physician*) but some samples yield positive TPR (resp. negative GP) Gender Gaps. This is problematic since we cannot not deduce a priori that a particular sample should produce a discrimination one way or the other.

In addition, the average biases also depend on the sample size. Again, we obtained unstable average biases for small samples (10K, 20K). The bias indicators are estimated on the minority class: an amount of 41, 115, 334 and 903 predicted surgeons were obtained in the test set for the 10K, 20K, 50K and 120K sampling sizes. Hence, their estimation is unstable for small samples.

However, GP appears as more stable than the other metrics in our experiments, in particular when there was little observations. Its variance was indeed close to 0.01, which is much lower than the variances of 0.1 and 0.2 for GP and PP, respectively. We explain this because on our dataset, for TPR and PP, they do not use all predicted surgeons (unlike GP), but only the predicted surgeons who are also real surgeons (in 10k sampling, there are 41 predicted surgeons vs. 30 real surgeons and predicted surgeons, which is an information loss of 26,8%).

3.2. Bias analysis with different metrics

Even for large samples with 120K observations, biases sometimes differed from what we expected. For the occupation *surgeon* (15% of females) the gender gap was negative for all metrics, which was expected. For *physician* (49,5% of females), we also expected to have negative or zero gender gap (see [7]). However the gender gaps were positive for all metrics, which means that the models discriminated against males. This example shows that intuitions of model-builders about biases are not always correct and this awareness should influence model construction and testing.

3.3. Results for all classes

In this section, we confirm our analysis on the specific occupations of *Surgeon* and *Physician* from a global point of view on all the classes of the model.

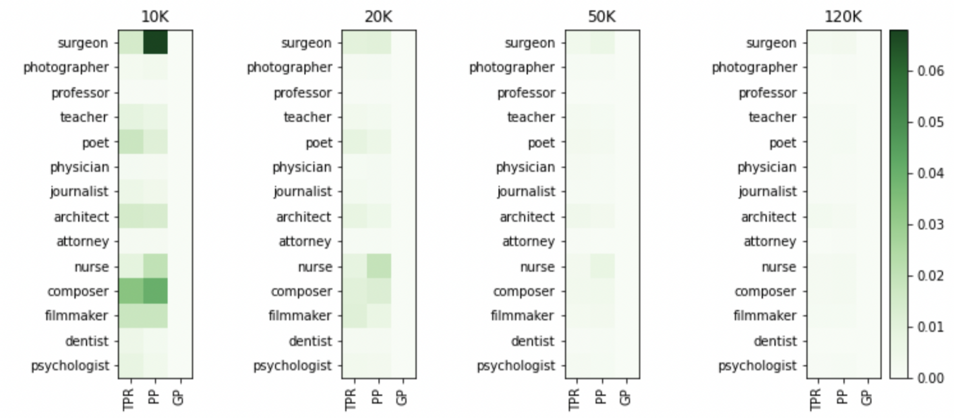

Figure 2: Variance of TPR/PP/GP gender gap for all occupations for model training on classic dataset for all sample sizes. The higher the variance, the darker the green. We have 50 sampling for each sample size. We kept only professions that have at least one prediction per gender for all samplings.

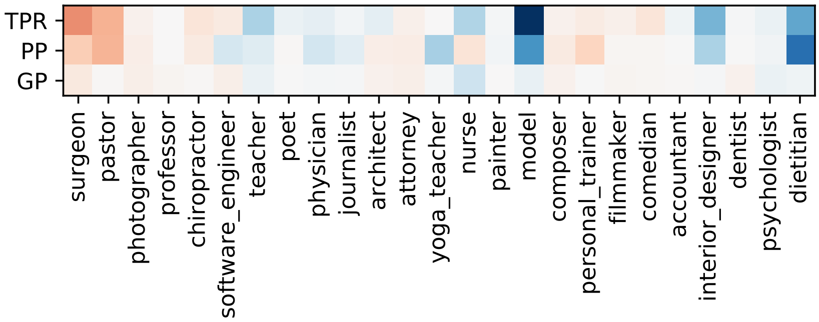

Figure 3: Mean of TPR/PP/GP gender gap for all occupation for model trained on 120K samplings. The more it is red, the more it is biased in favor of males, the more it is blue, the more it is biased in favor of females. We kept only professions with more than 10 predictions per gender.

The general results on all occupations confirm the analysis we made on the two occupations previously:

1. In the Figure 2, we have more and more important deviations on the variance of the metrics as the size of the data set decreases. And that on most trades. As explained before, the GP gender gap is more stable, because it has more data.
2. In the first table of Figure 3, the metrics give inconsistent results for several occupations: depending on the metric bias in favor of men or women for the same profession and the same model. This is particularly visible for the occupations: *software engineer*, *poet*, *architect*, *attorney*, and *nurse*.

These results give us guarantees on the generalization of our analysis carried out on the two classes previously. We find the same problems on the metrics and the size of the sample, regardless of the occupation being looked at.

4. Conclusion

Our paper used the *Bios* dataset to study the influence of the training set size on discriminatory biases. Our results shed light on new phenomena: (1) fairness metrics did not converge to stable results for small sample sizes, which precluded any conclusions about the nature of the biases; (2) even on large training samples, the biases discovered were not always those expected and varied according to the metrics for several occupations. These results give two clear messages to data scientists who must design NLP applications with a potential social impact. They should first be particularly careful, as the decision rules they train may have unexpected discriminatory biases. In addition, a bias metric not only returns a score but has a strong practical meaning and may be unreliable, in particular when working with small training sets. So multiple metrics should be considered. Also, statistical methods to obtain the variance of the observed metrics are necessary to guarantee the fairness of a model.

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.

References

[1] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, arXiv preprint arXiv:1609.05807 (2016).

[2] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big data 5 (2017) 153–163.

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

[4] A. Ezen-Can, A comparison of lstm and bert for small corpus, arXiv preprint arXiv:2009.05451 (2020).

[5] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[6] 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.

[7] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016) 4349–4357.