Are fairness metric scores enough to assess discrimination biases in machine learning?

Fanny Jourdan
IRIT, Université Paul-Sabatier
Toulouse, France
`fanny.jourdan@irit.fr`

Laurent Risser
IMT, Université Paul-Sabatier
Toulouse, France

Jean-Michel Loubes
IMT, Université Paul-Sabatier
Toulouse, France

Nicholas Asher
IRIT, Université Paul-Sabatier
Toulouse, France

Abstract

This paper presents novel experiments shedding light on the shortcomings of current metrics for assessing biases of gender discrimination made by machine learning algorithms on textual data. We focus on the *Bios* dataset, and our learning task is to predict the occupation of individuals, based on their biography. Such prediction tasks are common in commercial Natural Language Processing (NLP) applications such as automatic job recommendations. We address an important limitation of theoretical discussions dealing with group-wise fairness metrics: they focus on large datasets, although the norm in many industrial NLP applications is to use small to reasonably large linguistic datasets for which the main practical constraint is to get a good prediction accuracy. We then question how reliable are different popular measures of bias when the size of the training set is simply sufficient to learn reasonably accurate predictions. Our experiments sample the *Bios* dataset and learn more than 200 models on different sample sizes. This allows us to statistically study our results and to confirm that common gender bias indices provide diverging and sometimes unreliable results when applied to relatively small training and test samples. This highlights the crucial importance of variance calculations for providing sound results in this field.

1 Introduction

Potential biases introduced by Artificial Intelligence (AI) systems are now both an academic concern, but also a critical problem for industry, as countries plan to regulate AI systems that could adversely affect individual users. The so-called *AI act*[1] will require AI systems sold in the European Union to have good statistical properties with regard to any potential discrimination they could engender. In particular, under the AI Act, AI systems that exploit linguistic data like those for reviewing job candidates from text-based candidacies fall into the category of tightly regulated AI systems, as they are intended to be used for the recruitment or selection of natural persons (see Annex III of the AI act). Such AI systems will require frequent and rigorous statistical testing for unwanted biases.[2]

These regulatory advances have made it a pressing issue to define which metrics are appropriate for evaluating whether machine learning models can be considered fair algorithms in various industrial settings. In this context, we believe that these articles open at least two issues: (1) Each fairness metric quantifies the fairness of a model in a different way and not all metrics are compatible with each other, as already discussed in (Kleinberg et al., 2016; Chouldechova, 2017; Pleiss et al., 2017). It is therefore easy to optimize its algorithm according to a single metric to claim fairness while overlooking all the other aspects of fairness measured by other metrics. (2) Given that contemporary, theoretical discussions of fairness focus on large datasets but that the norm in many industrial

---

[1] `https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN`

[2] Such AI systems are considered high-risk. The *AI act* (Article 9.7) states: "*The testing of the high-risk AI systems shall be performed, as appropriate, at any point in time throughout the development process, and, in any event, prior to the placing on the market or the putting into service. Testing shall be made against preliminarily defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system*". Article 10.2 specifies that "*Training, validation, and testing data sets shall be subject to appropriate data governance and management practices. Those practices shall concern in particular, examination in view of possible biases*" (among others). Article 71 states that "*non-compliance of the AI system with the requirements laid down in Article 10 ... shall be subject to administrative fines of up to 30 000 000 EUR or, if the offender is a company, up to 6 % of its total worldwide annual turnover for the preceding financial year*".

NLP applications is to use small linguistic datasets (Ezen-Can, 2020), one can wonder how reliable different popular measures of bias when the size of the training and validation sets is simply sufficient to learn reasonably accurate predictions. In general, this leads us to pose two questions, which are central to this paper: Are fairness metrics always reliable on small samples, which are common in industrial contexts? How do they behave when applying standard debiasing techniques?

To answer these questions, we propose a new experimental protocol to expose gender biases in NLP strategies, using variously sized subsamples of the *Bios* dataset (De-Arteaga et al., 2019). We create 50 samples for each sample size (10k, 20k, 50k, and 120k) and train a model on each of the 200 samples. This gives us a mean and a variance on our results for all sample sizes to be able to compare them from a statistical point of view. We study the biases in these samples using three metrics; each sheds light on specific properties of gender bias.

Our study shows how bias is related to the training set size on a standard NLP dataset by revealing three points: First, commonly accepted bias indices appear unreliable when computed on ML models trained on relatively small training sets. Moreover, our experiments reveal that the group parity gender gap metric (3.5) appears to be more reliable than other metrics on small samples. Second, in the tested standard and large training sets, results are not homogeneous across professions and across the measures: sometimes there is gender bias against males, and sometimes against females in professions where one would expect something different. Finally, the most traditional de-biasing methods, which consist in removing gender-sensitive words or replacing them with neutral variants, makes different metrics yield surprising and sometimes seemingly incompatible bias effects. We explain this phenomenon by the definitions of the metrics. In light of these findings, we think that one should use the main fairness metrics jointly to look for biases in smaller datasets and run enough models to have a variance. Such bootstrapping procedures appear essential to robustly analyze how fair a prediction model is.

Our paper is structured as follows. Section 2 surveys related work. Section 3 introduces our experimental setup. Section 4 discusses our results, with conclusions coming in Section 5. Section 6

discusses some of the limitations of our work.

2 Related Work

Gender bias is pervasive in NLP applications: in machine translation (Vanmassenhove et al., 2019; Stanovsky et al., 2019; Savoldi et al., 2021; Wisniewski et al., 2021), in hate speech detection (Park et al., 2018; Dixon et al., 2018), sentiment analysis (Kiritchenko and Mohammad, 2018; Zellers et al., 2019), and in coreference resolution (Rudinger et al., 2018; Zhao et al., 2018). Gender bias with respect to classification has already been examined in (De-Arteaga et al., 2019; Gonen and Goldberg, 2019; Bolukbasi et al., 2016a; Lu et al., 2020; Bordia and Bowman, 2019), and reduced in (Pruksachatkun et al., 2021; Zhao et al., 2019, 2017). In particular, for the BERT model, Bhardwaj et al. (2021) investigated gender bias. More generally, Bender et al. (2021) has studied the impact of increasingly large language models and has highlighted the sexist or racist biases and prejudices that result from them.

However, the above-mentioned works only focused on single, large datasets. Recently, a growing literature has started to propose to leverage statistical properties of fairness metrics, thus providing both sophisticated analysis and practically useful algorithms (Lum et al., 2022; DiCiccio et al., 2020). In particular, a more rigorous statistical approach for BERT models was introduced in (Sellam et al., 2021).

In this paper, we investigate the pertinence of different fairness metrics on samplings of different sizes out of a large dataset. We apply our principled statistical procedure and we present the results of these measures, along with their standard deviation and properties coming from Student's t-tests. In addition to our scientific contribution, we have paid particular attention to the clarity of our explanations and the simplicity of our proposed protocol to allow small players to easily employ them for their real-world use cases. Finally, our results attest to the importance of applying techniques of statistical analysis to Fairness problems, and we hope that the guarantees gained through them provide a convincing argument for its more generalized application in the field.

3 Experimental protocol

In this section, we detail the various components of our experimental setup. Section 3.1 describes the

dataset and Section 3.2 the general type of model used to train the 200 models. Section 3.3 introduces our debiasing technique used to illustrate our protocol. Section 3.4 explains the sampling procedure and gives guarantees on the representativeness of the samples. Finally, Section 3.5 describes the different fairness metrics that we will compare and we justify these choices.

3.1 The *Bios* data set

The *Bios* dataset (De-Arteaga et al., 2019) contains about 400K biographies (textual data). For each biography, *Bios* specifies the gender (M or F) and the occupation (among 28 occupations, categorical data) of its author. Figure 7 (Appendix) shows the distribution of each occupation by gender.

3.2 DistilBERT model

Our task is to predict the occupation using only the textual data of the biography. This task is relevant in the case of our study because job prediction from LinkedIn biographies is used for job recommendation. It is therefore easy to imagine the consequences of gender discrimination in this context.

For this task, we will use the DistilBERT architecture. DistilBERT (Sanh et al., 2019) is a transformer architecture derivative from but smaller and faster than the original BERT (Devlin et al., 2018). This model is commonly used to do text classification. DistilBERT is trained on BookCorpus (Zhu et al., 2015) (like BERT), a dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables and headers), using the BERT base model as a teacher.

We have fine-tuned DistilBERT to adapt it to our text classification task. In our protocol, only the datasets were intervened on while keeping other factors the same in each model. We used 5 epochs, a batch size of 16 observations, an AdamW optimizer with a learning rate of 2e-5, and a cross-entropy loss when training the model.

3.3 De-biasing methodology

In this part, we state the debiasing technique used for the illustration of our protocol. Note that this technique is very basic and is only used to explain our experimental protocol. This protocol could be applied with any more elaborate debiasing technique.

A classic method for debiasing consists of removing explicit gender indicators (*i.e.'he', 'she', 'her', 'his', 'him', 'hers', 'himself', 'herself', 'mr',*

'mrs', 'ms', 'miss' and first names). For a model like DistilBERT, however, we could not just remove words because the model is sensitive to sentence structure, not just lexical information. We, therefore, adjusted the method by replacing all the first names with a neutral first name[3] (*Camille*) and by choosing only one gender for all datasets (e.g., for all individuals of gender g, we did nothing; for the others, we replaced explicit gender indicators with those of g). We then created two datasets with only female or male gender indicators, and the only first name *Camille*.

3.4 Sampled training and test sets

We tested the robustness of our model with respect to the various bias measures on training sets of different sizes. We randomly sampled 50 different training sets containing 10K, 20K, 50K, and 120K biographies out of the 400K of (De-Arteaga et al., 2019). We trained a model on each of these 200 samples. Each of these models has the same architectures and the same hyper-parameters stated previously. 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 example, given 2002 female surgeons out of 388862 persons in the initial dataset (0.51 %), we randomly picked 51 women surgeons for a sample with 10000 individuals (0.51 %). For the split between the train and test sets, we respectively used 70% and 30% of the dataset.

Creating these 200 different models and observations makes it possible to quantify the variability of the results obtained using each size of subsampled training sets. This will additionally allow us to ensure that all differences discussed in our results are statistically significant using Student's t-tests. Our experimental protocol, therefore, gives us more guarantees than traditional protocols based on a single model.

3.5 Gender bias metrics

Let $\hat{Y}$ and Y be the predicted and the true target labels (*i.e.*, the 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 the following metrics: Group Parity (GP), True Positive Rate

---

[3]We can take any first name because, since we change all the first names of the dataset by this one, it will necessarily be neutral.

(TPR), and Predictive parity (PP). They are defined as:

$$GP_{g,y} = P(\hat{Y} = y | G = g) \,, \quad (1)$$

$$TPR_{g,y} = P(\hat{Y} = y | G = g, Y = y) \,, \quad (2)$$

$$PP_{g,y} = P(Y = y | \hat{Y} = y, G = g) \,. \quad (3)$$

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. We now discuss each measure in turn.

Statistical/Group Parity (GP) The condition GP verifies gender balance (ie. $GP\_Gap_{g,y} = 0$) if the males and females groups have equal probability of being assigned to the predicted class $\hat{y} \in \{noty, y\}$. It is the most famous and intuitive metric of fairness.

True Positive Rate (TPR) $TPR\_Gap_{g,y} = 0$ if among all individuals in the y class, the probability of being predicted $\hat{y}$ is the same for males and females. This metric is widely used in the field of NLP in particular.

Predictive Parity (PP) $PP\_Gap_{g,y} = 0$ if among all individuals predicted to belong to the y class, the probability of $Y = y$ is the same for males and females. PP is similar to Calibration (within groups), and widely used in fairness to compare with other metrics. We use PP here because it uses the same parameters Y and $\hat{Y}$ as our other metrics.

4 Results and discussion

As shown in Figure 1, all the models we trained reached a prediction accuracy ranging from 0.72 to 0.86, as in (De-Arteaga et al., 2019), which we consider as good since the classification problem involved 28 different occupations.

All comparisons in this part were considered as significant by using Student's t-tests (p-value of 0.05).

We created two datasets without gender information: one version with all female gender indicators and the other with all male gender indicators. Gender, therefore, has no impact on the finetuning part of our model. However, since we are starting from a pre-trained DistilBERT model (without a gender-neutral dataset), we had to check that the pre-training had no impact on the prediction. We therefore also made a Student's test between the predictions of one model trained on the dataset with all the female gender indicators, and of another trained on the dataset with all the male gender indicators. The difference was not statistically significant; using one model or the other makes no difference.

The analysis of the results of our protocol is made in two steps: a specific part and a general part. Below in Section 4.1, we analyze biases on two specific occupations, *Surgeon* and *Physician*. These two occupations are socially very interesting and their male/female distribution is very different, which is something we wanted to study. In Section 4.2, we also observe the biases across the gamut of occupations in *bios*. All the results found in this preliminary study remain valid in a generalized case where we look at all the classes of the model. Dividing our study like this allows us to discuss various details which support our key message without weighing down the article in the specific part while guaranteeing that our analysis is global and applies to the other classes of the model in the general part.

4.1 Results and discussion for the classes Surgeon and Physician

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. The occupation *Physician* is well balanced in the training set between males and females (49,5% female), while the training set for *Surgeon* contains many more males than females (15% female).

We computed F1-scores in Figure 2, which are good to reasonable, except for the 10K samples for surgeons, which appear as too small for our predictive task. Quantitative results related to the fairness metrics are shown in Figures 3 and 4. Each box-plot contains the TPR, GP, PP Gender Gaps obtained on the test set for *surgeons* and *physician*. Negative (positive) gender gaps mean that there is discrimination against females (males).
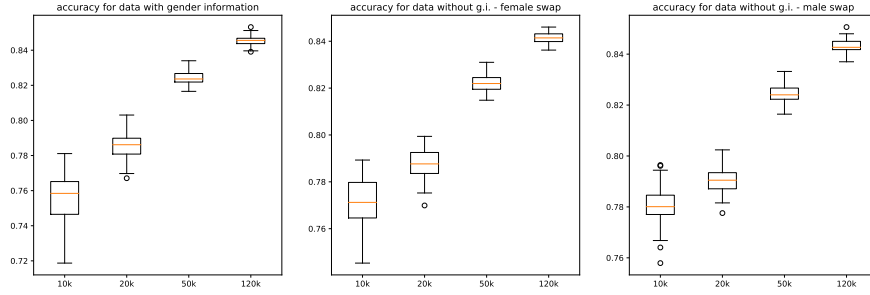
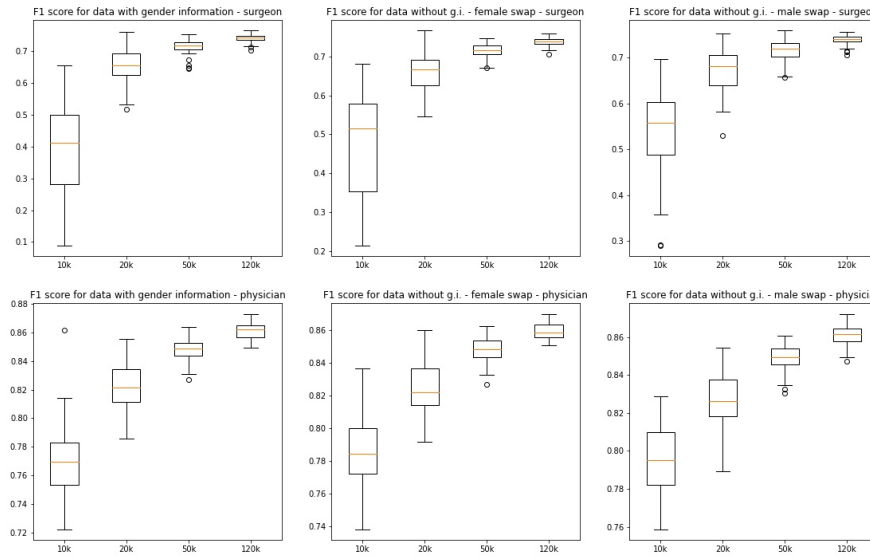Figure 1: Boxplots representing the variations of prediction accuracy for all sampling sizes

Figure 2: Boxplots representing the variations of prediction F1-scores for all sampling sizes for surgeon (top) and physician (bottom)

4.1.1 Results on small data samples

Our experiments clearly show that the lower the amount of observations in the training set, the more the fairness metrics vary in the test set. The samples with 10K and 20K observations present particularly unstable biases. For example, most TPR (resp. GP) Gender Gaps are negative (resp. positive) for *surgeon* (resp. *physician*) but some samples yield positive TPR (resp. negative GP) Gender Gaps. This is problematic since we cannot deduce a priory that a particular sample should produce 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 more stable than the other metrics in our experiments, in particular when there were few 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 $26,8\%$). We explain this intuition mathematically in detail in appendix A.
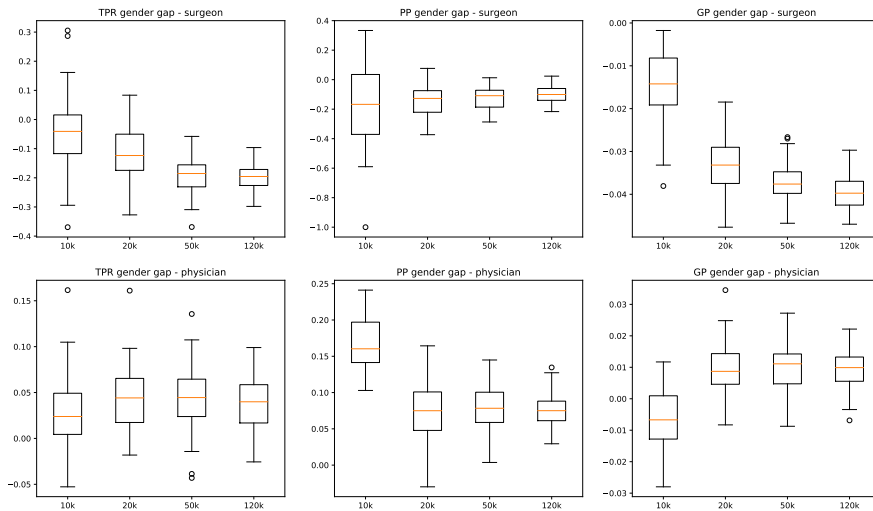
Figure 3: Boxplots of the gender gaps obtained using 10K, 20K, 50K, and 120K randomly sampled observations (50). **(Left)** True Positive Rate (TPR) gender gaps for surgeons and physicians; **(Middle)** Predictive Parity (PP) gender gaps for surgeons and physicians; **(Right)** Group Parity (GP) gender gaps for surgeons and physicians.

Considering this result, we recommend using a simpler indicator like GP gender gap for small-size sets.

4.1.2 Bias analysis with different metrics

General results 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 a negative or zero gender gap (see (Bolukbasi et al., 2016b)). 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.

Results with debiasing Intuitively, removing explicit gender indicators should reduce the bias (De-Arteaga et al., 2019). As shown Figure 4, however, our experiments show that this is not necessarily the case. Using TPR and GP Gender Gaps, we see a bias initially in favor of women (resp. men) and increases (resp. decreases) for the *physician* (resp. *surgeon*) class after debiasing. Removing gender indicators thus favored women in these two occupations.

PP Gender Gap shows different effects for de-

biasing: For *physician* (resp. *surgeon*), the initial bias in favor of women (resp. men) decreases (resp. increases) after debiasing. Removing gender indicators thus favored men in these two occupations.

To explain this phenomenon, we can remark that removing gender indicators allowed us to predict more women than before in the two professions. The metrics interpret this differently. By definition, $PP_{f,y} = P\left(Y = y | \hat{Y} = y, G = f\right)$ decreases when the number of $\hat{Y}$ increases. In addition, $TPR_{f,y} = P\left(\hat{Y} = y | Y = y, G = f\right)$ and $GP_{f,y} = P\left(\hat{Y} = y | G = f\right)$ increases when the number of $\hat{Y}$ increases.

Using either GP/TPR gender gap or PP gender gap amounts to choosing between focusing on the number of people predicted in the discriminated group (parity) or focusing on the people in the discriminated group who are well predicted (truth). This explains the different interpretations of these indicators.

4.2 Results and discussion for all classes

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

The general results on all occupations confirm the analysis we made on the two occupations previously:
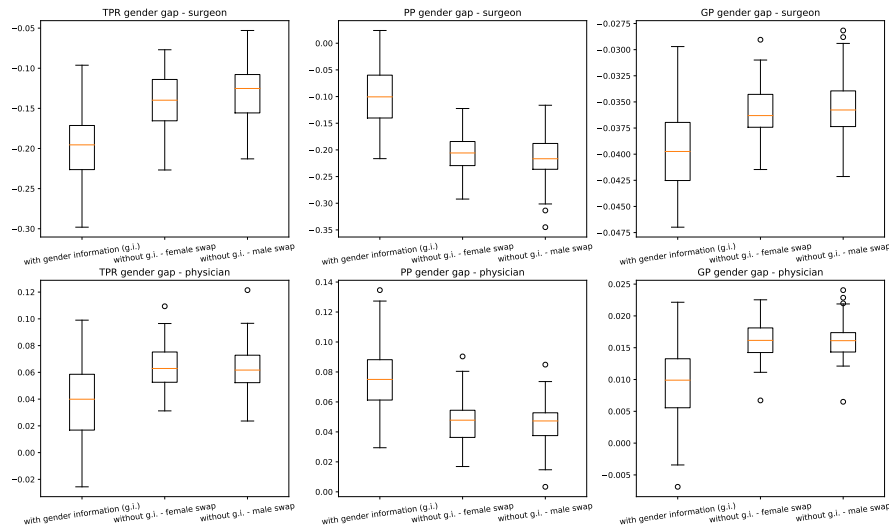
Figure 4: Boxplots representing the fairness metrics for surgeon (top) and physician (bottom) for 120k samplings for the base model, model with only female indicators, and model with only male indicators
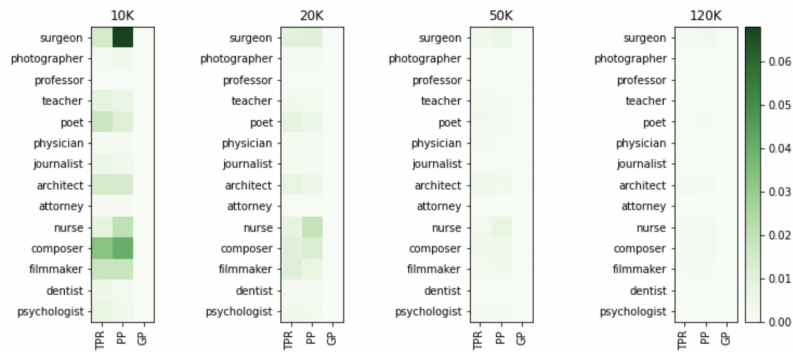

Figure 5: **Variance of TPR/PP/GP gender gap for all occupations** for model training on the classic dataset for all sample sizes. The higher the variance, the darker the green. We have 50 sampling for each sample size. We kept only professions that have at least one prediction per gender for all samplings. So we had to remove *paralegal*, *dj*, *rapper*, *pastor*, *chiropractor*, *software engineer*, *attorney*, *yoga teacher*, *painter*, *model*, *personal trainer*, *comedian*, *accountant*, *interior designer*, and *dietitian*

1. In Figure 5, 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 6 (for the classic model), 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*.

3. By comparing the two tables in Figure 6, we confirm that depending on the metric we are looking at, the basic debiasing technique used will not have the same effects on the bias. In several professions, we see that the bias on the TPR gender gap in favor/against women increases when on the bias on the PP gender gap decreases and vice versa. This is evident in the professions: *surgeon*, *pastor*, *photographer*, *chiropractor*, *teacher*, *journalist*, *architect*, *attorney*, *nurse*, *composer*, *personal trainer*, *comedian*, *interior designer*, and *dietetitan*.

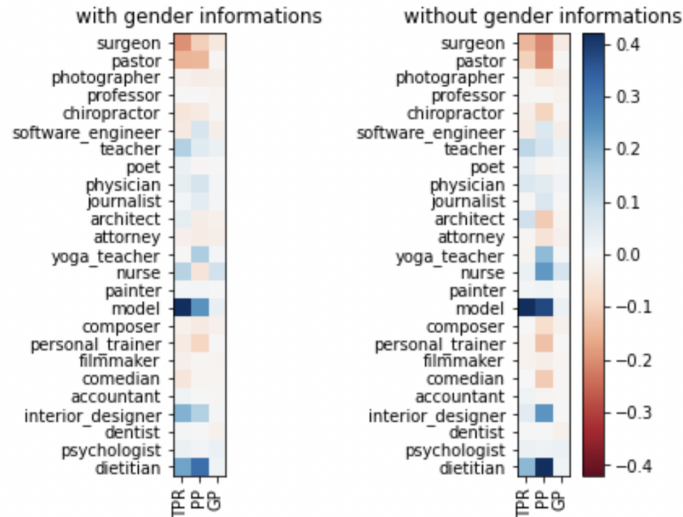These results give us guarantees on the general-

Figure 6: **Mean of TPR/PP/GP gender gap for all occupations** for model trained on 120K samplings. On the right, the model was trained on the classic dataset, and on the left, the model was trained on the dataset without gender indicators. 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. So we had to remove *paralegal*, *dj* and *rapper*.

ization of our analysis carried out on the two classes previously. We find the same problems with the metrics and the size of the sample, regardless of the occupation being looked at.

5   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; (3) a simple debiasing method, which consists in removing explicit gender indicators, had an unstable impact in our results depending on the metrics, though our analysis of the metrics can explain the instability. 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 and statistical methods to obtain the variance of the observed metrics are necessary to guarantee the fairness of a

model.

6   Limitations

A limitation of our conclusions is that although it is necessary to use several fairness metrics to be able to properly quantify the bias, this is not enough. These metrics must be well chosen according to the context and the task being looked at. The expertise of a person working in the field is therefore always necessary to have the most complete possible interpretation of the bias. More specifically, the different fairness metrics measure distinct properties, and the fact that they are often incompatible has been a core part of the fair ML conversation from the beginning (Barocas et al., 2017). Thus, suggesting to choose a different metric depending on the sample size may sometimes be inappropriate, since this choice may depend on the meaning of the metric in a given application. We must therefore be very careful and see the notion of robustness as additional necessary information and not as a replacement for the metric's meaning.

We also did not reduce the bias using advanced strategies because this paper focuses more on the analysis intended for a population closer to the law than to machine learning. In this vein, it is interesting to note that more and more tools are available to reduce bias. In particular, (Sikdar et al., 2022) makes it possible to reduce the bias according to several fairness metrics, therefore remaining in our

logic of taking several metrics.

The main problem raised by our article comes from the fact that fairness indices are not stable when they are calculated. We should consider them as random variables and look at their law. The first step is to look at the mean and the variance as done in this paper but having the full distribution would be more interesting. Works that compute the asymptotic law can be taken as an example like (Ji et al., 2020; Besse et al., 2022).

Ethics Statement

Natural Language Processing is gaining a considerable amount of attention these days and it is extremely important to evaluate how NLP datasets will impact the gender bias when used to train models that will be used in the real world. This work uses different experiments and fairness metrics to shed light on the shortcomings of these metrics with respect to gender bias made by ML algorithms on textual data. We believe that transmitting knowledge from research to industry on a subject like fairness is essential to make the field of ML more ethical. Hence, this work focuses on issues that most affect the industrial ML landscape and contains a clear message to them on how they should change their current practices.

Acknowledgements

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *Nips tutorial*, 1:2017.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. 2022. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, 76(2):188–198.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, pages 1–11.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

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

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*, pages 120–128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi, and Deepak Agarwal. 2020. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1467–1477.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Aysu Ezen-Can. 2020. A comparison of lstm and bert for small corpus. *arXiv preprint arXiv:2009.05451*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. *Advances in Neural Information Processing Systems*, 33:18600–18612.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-biasing "bias" measurement. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 379–389.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. In *ACL-Finding*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *arXiv preprint arXiv:2104.06001*.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, et al. 2021. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.

Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2022. Getfair: Generalized fairness tuning of classification models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 289–299.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*.

Guillaume Wisniewski, Lichao Zhou, Nicolas Ballier, and François Yvon. 2021. Biais de genre dans un système de traduction automatiqueneuronale: une étude préliminaire (gender bias in neural translation: a preliminary study). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1: conférence principale*, pages 11–25.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *NAACL (short)*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.

A   Mathematical intuition

Intuition   Let $\hat{Y}$ and Y be the predicted and the true target labels, respectively. Let G be a random variable representing the binary gender and let n be the number of all individuals. We can write the estimators of Group Parity, True Positive Rate, and Predictive Parity metrics like this:

$$\hat{GP}_{g,y} = \frac{\sum_{i=1}^{n} 1_{\{\hat{Y}_i = y \,\cap\, G_i = g\}}}{\sum_{i=1}^{n} 1_{\{G_i = g\}}}$$

$$\hat{TPR}_{g,y} = \frac{\sum_{i=1}^{n} 1_{\{\hat{Y}_i = y \,\cap\, Y_i = y \,\cap\, G_i = g\}}}{\sum_{i=1}^{n} 1_{\{Y_i = y \,\cap\, G_i = g\}}}$$

$$\hat{PP}_{g,y} = \frac{\sum_{i=1}^{n} 1_{\{\hat{Y}_i = y \,\cap\, Y_i = y \,\cap\, G_i = g\}}}{\sum_{i=1}^{n} 1_{\{\hat{Y}_i = y \,\cap\, G_i = g\}}}$$

We set $A = \{\hat{Y}_i = y \,\cap\, G_i = g\}$ and $B = \{Y_i = y\}$. By definition, $\#(A \cap B) \leq \#A$ where $\#$ is the cardinal of the set. So we have $\#\{\hat{Y}_i = y \,\cap\, Y_i = y \,\cap\, G_i = g\} \leq \#\{\hat{Y}_i = y \,\cap\, G_i = g\}, \forall i = 1, ..., n$.

We can define n_{GP}, n_{TPR}, n_{PP} the number of individuals respectively looked by the estimator of Group Parity, True Positive Rate, and Predictive Parity metrics and we have:

$$n_{GP} = \sum_{i=1}^{n} \#(\{\hat{Y}_i = y \cap G_i = g\} \cap \{G_i = g\})$$

$$= \sum_{i=1}^{n} \#\{\hat{Y}_i = y \cap G_i = g\},$$

$$n_{TPR} = \sum_{i=1}^{n} \#(\{\hat{Y}_i = y \cap Y_i = y \cap G_i = g\}$$

$$\cap \{Y_i = y \cap G_i = g\})$$

$$= \sum_{i=1}^{n} \#\{\hat{Y}_i = y \cap Y_i = y \cap G_i = g\},$$

$$n_{TPR} = \sum_{i=1}^{n} \#(\{\hat{Y}_i = y \cap Y_i = y \cap G_i = g\}$$

$$\cap \{\hat{Y}_i = y \cap G_i = g\})$$

$$= \sum_{i=1}^{n} \#\{\hat{Y}_i = y \cap Y_i = y \cap G_i = g\}.$$

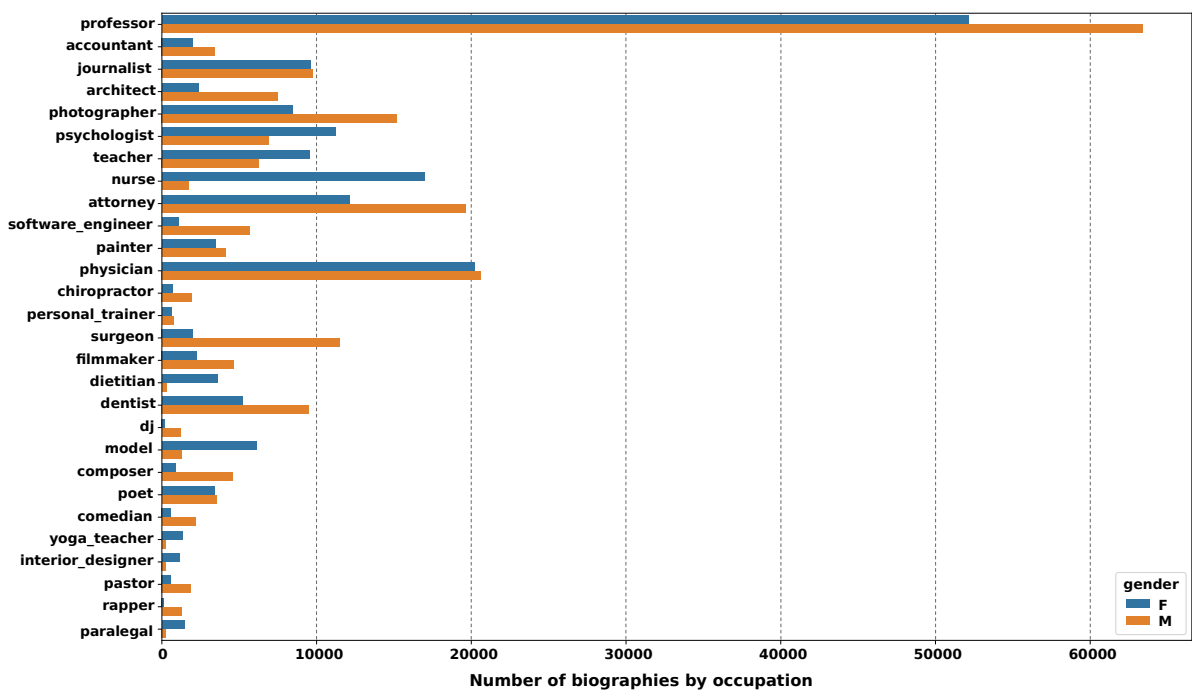Then : $n_{TPR} = n_{PP} \leq n_{GP}$.

B Additional figure

Figure 7: Number of biographies for each occupation by gender on the total dataset