Bias Identification in Language Models is Biased
Fanny Ducel, Aurélie Névéol, Karên Fort

To cite this version:
Fanny Ducel, Aurélie Névéol, Karên Fort. Bias Identification in Language Models is Biased. Workshop on Algorithmic Injustice 2023, Jun 2023, Amsterdam, Netherlands. hal-04171198

HAL Id: hal-04171198
https://inria.hal.science/hal-04171198
Submitted on 26 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Bias Identification in Language Models is Biased

Fanny Ducel¹, Aurélie Néveol², Karën Fort¹

¹) Sorbonne Université/LORIA, France
²) Université Paris-Saclay, CNRS, LISN, France

The issue of biases in language models (LMs) has recently gained significant attention within the Natural Language Processing (NLP) community. Addressing this problem is crucial due to its potential sociological, ethical, and political implications, especially considering the widespread use of language models by the general public. There exists now a significant amount of literature aimed at evaluating and mitigating these biases. In this work, we perform a non-exhaustive survey of this literature and find that these efforts are not immune to biases themselves. We then outline a meta-data based analysis of these works to partly explain this phenomenon.

To conduct this study, we manually annotated 66 NLP research articles written in English between 2018 and 2023 that focus on bias evaluation or bias mitigation in language models. We will present in further details our methodology, and the overlap with the papers studied in [Blodgett et al., 2020]. Amongst them, 53 articles propose debiasing technique or metrics, while 13 articles are surveys and position papers.

As a starting point, we note some common limitations mentioned in multiple approaches and surveys we studied, namely: English is the target language, the perspective is US-centric, and the only type of studied bias is gender, and more precisely, binary gender.

There are 23 different languages targeted amongst the 53 experimental articles we studied. However, we find that 96% (51/53) of articles focus on English and that 83% (44/53), exclusively so (see Figure 1). We also note that 33% (17/51) of papers do not explicitly state that they work on English. As pointed out in [Ducel et al., 2022], it is important to do so, as English is not a “default language” and approaches proposed for it cannot be trivially adapted for other languages. However, we want to acknowledge the recent efforts that are being made to work on more diverse languages. Some of the articles (9/53) even propose multilingual solutions [Lauscher et al., 2021, Nozza et al., 2021, Arora et al., 2023].

Further, we claim that the perspective of a vast majority of these articles is US-centric. Our corpus contain 237 different authors that are based in 21 different countries. Nonetheless, similarly to the distribution of studied languages, we can see on Figure 2 that 56% of papers (37/66) contain at least one author affiliated to the United States. This rises to 72% (37+11 out of 66) when extrapolating country of residence from affiliated companies, in cases where countries are not specified. This can be problematic in that biases are cultural, so the biases that are taken into account by US authors are specific to their country.
Therefore, it is likely that a language model that has been supposedly "de-biased" using an American interpretation of bias would contain lots of other biases, which we could neither detect, nor mitigate [Malik et al., 2022].

We also study the proportion of industry affiliations present in the papers. We find that 42% (28/66) of them have at least one author affiliated to a company (see Figure 3). In total, 13 companies are represented, and we find the most famous BigTech amongst them : Microsoft, Google, Facebook and Amazon. We will explain in more details the potential consequences of this, relying in particular on [Abdalla and Abdalla, 2020].

Finally, we turn our attention to the type of bias that is being studied. For this part, we exclude again the 13 survey and position papers. We find that 88% (47/53) of the articles focus on gender biases (see Figure 4), and 95% of them (45/47) on binary gender more specifically. Nonetheless, 58% (31/53) of the total of papers deal with several biases in parallel, and 11% (6/53) are intersectional, studying different types of biases simultaneously. Efforts towards intersectionality are very valuable as biases emerge from different sources, take different forms and individuals can suffer from different kinds of biases at once [Cao et al., 2022, Crenshaw, 1989].

To conclude, we would like to emphasize that our goal is not to belittle the efforts that have been produced, but to shed light on the biases inherent to the research. We would like to make simple recommendations, especially to encourage researchers to write a short paragraph stating where they write from, as it is usually done in social sciences.

We also want to highlight that the current research on biases in language models does not mirror the reality of biases. Biases are not universal, they depend heavily on language and culture. Moreover, gender is not the only source of bias, and it is well known that gender is not only binary, assuming that can even be harmful [Larson, 2017, Dev et al., 2021]. We advocate for more diverse resources to better profile these other types of biases in language models, so that we can try and tackle the notion of bias in all its complexity.
Figure 1: Distribution of studied languages among papers

Figure 2: Distribution of authors’ countries affiliations among papers

Figure 3: Distribution of authors’ companies affiliations among papers

Figure 4: Distribution of studied biases among papers
References


Belgium. Association for Computational Linguistics.


