A call for a research agenda on fair NLP for Portuguese

Luiz Fernando F. P. de Lima[1], **Renata Mendes de Araujo**[2,3,4]

[1] Centro de Estudos e Sistemas Avançados do Recife (CESAR) - Recife, PE, Brasil

[2]Universidade Presbiteriana Mackenzie - São Paulo, SP, Brasil

[3]Universidade de São Paulo (USP) - São Paulo, SP, Brasil

[4]Escola Nacional de Administração Pública (ENAP) - Brasília, DF, Brasil

lffpl@cesar.org.br, renata.araujo@mackenzie.br

Abstract. *Diverse areas widely apply artificial intelligence and natural language processing (NLP) tools to their contexts. However, these algorithms present ethical issues, such as biased and discriminatory decisions. For example, representation biases in NLP can result in discriminatory behavior towards race and gender. Works have been addressing this issue and seeking to build fair NLP solutions, however they mainly focus on Anglo-Saxon languages. This work aims to challenge the scientific community in order to stimulate and motivate further research in the fair NLP specifically for the Portuguese language. To achieve this, a literature review was conducted to identify existing research efforts and indicate future directions.*

1. Introduction

Different biases, such as historical biases, representation biases, evaluation biases, and human interpretation biases, can be embedded into a machine learning (ML) model during its training [Ruback et al. 2021]. The incorporation of these biases by the model can result in unfair and discriminatory outcomes.

We can also observe cases of algorithmic discrimination in NLP tools. For example, Amazon's algorithm for recruiting and selecting employees that penalized resumes belonging to women and identified the applicant's gender, even if this information was omitted [Dastin 2018]. On the other hand, recent literature presents attempts to achieve fair NLP algorithms for different tasks through various techniques such as data augmentation, gender marking and learning gender-neutral embeddings [Mehrabi et al. 2021, Bolukbasi et al. 2016, Leavy 2018].

However, these fair NLP researches mostly use English databases, not worrying about whether their approaches work when concerning other languages. In this scenario, we can raise concerns about the democratization of these fair NLP solutions for different languages, especially for the Portuguese language. Therefore, it is essential to stimulate and advance research in the aspect of fair NLP for Portuguese, making this technology fairer for a larger segment of society [Camões - Instituto da Cooperação e da Língua 2023].

In this sense, the main goal of this work is to identify research gaps, and outline a research agenda in order to motivate scholarly peers to foster a broader

democratization of fair NLP solutions for Portuguese. As a first step to achieve this goal we focused on mapping, through a literature review, fair NLP solutions that are concerned with solving problems of algorithmic discrimination in Portuguese.

2. Background

In this work, we aim to outline a research agenda for fair NLP concerning with Portuguese. We understand fair NLP as techniques that attempts to mitigate representation biases which can lead algorithms to discriminatory behavior, such as, denigration, stereotypes, recognition, and under-representation [Sun et al. 2019].

Fair NLP works also points out and discuss the societal and ethics implications of these problems and introducing metrics, evaluations, and fair architectures/models for diverse NLP tasks. For instance, [Bolukbasi et al. 2016] investigated the gender bias inherent in word embeddings and proposed a framework to mitigate such behavior. Moreover, [Bender et al. 2021] presented some of the social impacts, limitations, and potential harms associated with language models.

The literature present some surveys and systematic reviews that map research from the perspective of fair NLP. For example, [Mehrabi et al. 2021] bring a broad survey on algorithmic fairness, including mapping fair NLP solutions. However, the authors do not present their research methodology, leaving aside the steps for reproducing the performed review. Given the many references raised, this work can help identify and aggregate research not found by our own methodology.

Among the research presented in the survey, we highlight the approach proposed by [Font and Costa-Jussa 2019], which focuses on mitigating gender biases in translating sentences between English and Spanish. In addition, from the perspective of mitigating gender biases in machine translation (MT) tasks, the work by [Vanmassenhove et al. 2018] uses a multilingual dataset with examples of texts in all languages present in the European Union, including Portuguese.

In their literature review, [Sun et al. 2019] categorize the research under four perspectives of biases: denigration, stereotypes, recognition, and under-representation. Although this work does not present any methodological aspect to map the presented papers, its final section reinforce our concerns on building a research agenda focusing on the mitigation of biases in languages other than English.

Finally, [Blodgett et al. 2020] present a survey with a more critical perspective on NLP research that address issues of bias, mainly due to vagueness or inconsistency in the motivations of the reviewed papers.

3. Method

The methodology applied to this literature review is based on the procedure presented by [Blodgett et al. 2020], however adding guarantees that we are returning research papers that consider Portuguese. The steps and results are detailed below.

We seek to answer the following research questions: (1) What papers are concerned with examining NLP techniques just for Portuguese? (2) What biases do the literature approaches aim to mitigate? And for which NLP tasks? (3) Are there

database resources in Portuguese that can be used in the context of fair NLP? (4) What is the maturity of reproducibility and openness of information of these works?

To select the articles, we used research strings with terms commonly present in studies on NLP, adding constraints to proper select works concerning Portuguese: ((“bias” AND (“fairness” OR “equity”)) AND (“natural language processing” OR “nlp”) AND “portuguese”) OR ((“viés” AND (“justiça” OR “equidade”)) AND (“processamento de linguagem natural” OR “NLP”) AND “português”).

The IEEE Xplore, ScienceDirect, and Periódicos CAPES were used as research sources, collecting research works in the five-year interval (2018 - 2022). In addition, as a guarantee of not excluding relevant work, we manually searched for papers in some relevant conferences in the areas of AI/ML, NLP, and ethics in AI: ICML, NeurIPS, AIES, FAccT, WWW, BRACIS, ENIAC, PROPOR, and STIL.

Using the search string in research sources, no papers were returned. With the manual search in the conferences' proceedings, we could identify two papers that fit this review. Finally, we add to this set the work of [Vanmassenhove et al. 2018], pointed in Section 2 as a potential reference for the analysis. As defined by [Blodgett et al. 2020], the scope of the papers was analyzed as an inclusion criterion. Those articles that talk about NLP only with text are included, thus omitting other works that use speech. In this last step, we kept the raised three articles.

4. Results

Table 1 presents an overview of the analysis, pointing out the answers to the research questions. All papers follow the positivist research paradigm, bringing algorithmic proposals and experiments to analyze and mitigate gender biases in NLP. [Vanmassenhove et al. 2018, Cho et al. 2021] focus on MT tasks, while [Santana et al. 2018] work on de-biasing word embeddings for analogy tasks. In addition, we observed how the quality of the works in relation to issues of reproducibility of their experiments vary due the presentation of reproduciability details.

Table 1. Overview of the analyzed papers concerning the research questions

Reference	Getting gender right in neural machine translation [Vanmassenhove et al. 2019]	Is there gender bias and stereotype in portuguese word embeddings? [Santana et al. 2018]	Towards cross-lingual generalization of translation gender bias [Cho et al. 2021]
Source	Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018	International Conference on the Computational Processing of Portuguese (PROPOR) 2018	ACM Conference on Fairness, Accountability, and Transparency 2021
NLP Task	Machine translation	Word embeddings analogies	Machine translation
Studied Bias	Gender bias	Gender bias	Gender bias
Dataset	Corpus built from Europarl	Corpus proposed by [Hartmann et al. 2017]	Corpus built from the proposed by [Cho et al. 2019] + modifications with systematic process for translation
Reproduciability	Moderated	High (with limitations)	High

Considering the differences in language structures in terms of gender and observing the loss of information in automated translations, [Vanmassenhove et al. 2018] seek to mitigate gender biases in MT that result in morphologically incorrect translations. Their propose focuses on the use of gender information to help MT algorithms to perform translations with better quality. The assessments for translations into FR and PT suggest that the approach has the potential to improve gender agreement in translations between EN-PT.

[Santana et al. 2018] aim to analyze and remove gender biases in Portuguese word embeddings for the analogy task. The authors propose an evaluation pipeline in three stages: 1) use of the word2vec model with strategy proposed

by [Hartmann et al. 2017]; 2) application of the debiasing algorithm proposed by [Bolukbasi et al. 2016] and; 3) model accuracy assessment. The authors focus on mitigating bias in Portuguese word embeddings and investigating its effects on the accuracy of the model. However, they do not propose any adjustments to the applied debiasing algorithm to account for the specificities of Portuguese gender structures.

[Cho et al. 2021] also present a approach based on algorithmic advances to evaluate gender biases in machine translation tools. However, they aim to conceptualize and consider several linguistic aspects, such as the presence of gender-neutral pronouns, agreement of articles concerning gender, and derivation of the noun according to its gender, on the languages presented, to build a more generic and agnostic model to the language pairs for this assessment. Those concerns also regard to the gender structures present in the Portuguese language.

5. Outlining a Research Agenda

In this current work, we could not find any work in the chosen search databases, only two papers were identified in conference proceedings, and one more research was added to this review because it was referenced in a previously analyzed paper.

These results indicates how many perspectives still open to be explored in research on fair NLP in Portuguese. For example, future works can continue to examine the interchangeability of existing fair NLP solutions that mitigate gender bias for Portuguese, e.g., one could evaluate the use of other techniques such as those presented by [Mehrabi et al. 2021].

Other opportunities are related to the concern with the regional linguistic variations of Portuguese. Similarly, concerning in mitigating harms to specific and marginalized groups, additional research efforts may focus on exploring linguistic variations within a particular country, such as Brazil [Drager et al. 2021, Guy 1981]. Furthermore, works could extend to variations and dialects specific to marginalized groups, such as the LGBTQIA+ community.

Beyond gender biases, future work needs to bring efforts to the mitigation of racial biases in NLP in Portuguese. The racial perspective has been explored in the context of fair NLP by the international community [Blodgett and O'Connor 2017] and needs to be amplified for the Portuguese context.

For all these perspectives, it is essential to engage in the endeavor of seeking more diverse and representative datasets. Research efforts should focus on examining the presence of biases and sources of harm within already published datasets. Additionally, it is crucial to make a concerted effort to evaluate datasets when a group or entity intends to construct and publish a new dataset. In this regard, one can adopt the methodology proposed by [Gebru et al. 2021].

Finally, we point out opportunities for carrying out research using other paradigms. As suggested by [Araujo et al. 2017] for the IS area, research in NLP, can use interpretive approaches in order to present contextualized works to society, analyzing the impacts of proposed solutions to the world. Likewise, critical research is welcome in this area and may refer to, for example, critical works on algorithmic racism in facial recognition tools [Buolamwini and Gebru 2018, Silva 2020].

References

Araujo, R., Fornazin, M., and Pimentel, M. (2017). Uma análise sobre a produção de conhecimento científico nas pesquisas publicadas nos primeiros 10 anos da isys (2008-2017). *iSys-Brazilian Journal of Information Systems*, 10(4):45–65.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Blodgett, S. L. and O'Connor, B. (2017). Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*.

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

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Camões - Instituto da Cooperação e da Língua (2023). Dados sobre a língua portuguesa. `https://www.instituto-camoes.pt/images/img_agenda2023/Dados_sobre_a_l%C3%ADngua_portuguesa_2023.pdf`.

Cho, W. I., Kim, J., Yang, J., and Kim, N. S. (2021). Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 449–457.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`.

Drager, K., Rilliard, A. O. B., Vieira, M. d. S. M., and Wiedemer, M. L. (2021). Linguistic varieties in brazil and beyond. *Revista Diadorim*, 23(1):24–33.

Font, J. E. and Costa-Jussa, M. R. (2019). Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Guy, G. R. (1981). Linguistic variation in brazilian portuguese: Aspects of the phonology, syntax, and language history.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pages 14–16.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Ruback, L., Avila, S., and Cantero, L. (2021). Vieses no aprendizado de máquina e suas implicações sociais: Um estudo de caso no reconhecimento facial. In *Anais do II Workshop sobre as Implicações da Computação na Sociedade*, pages 90–101, Porto Alegre, RS, Brasil. SBC.

Santana, B. S., Woloszyn, V., and Wives, L. K. (2018). Is there gender bias and stereotype in portuguese word embeddings? *arXiv preprint arXiv:1810.04528*.

Silva, T. (2020). Visão computacional e racismo algorítmico: branquitude e opacidade no aprendizado de máquina. *Revista ABPN*, 12:428–448.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.