Gender-Neutral English to Portuguese Machine Translator: Promoting Inclusive Language

Ricardo Trainotti Rabonato[1], Evangelos Milios[2], and Lilian Berton[1]

[1] Universidade Federal de Sao Paulo, Avenida Cesare Mansueto Giulio Lattes, nř 1201 - Eugênio de Mello, SP - Brazil
[2] Dalhousie University, 6050 University Avenue, Halifax, NS - Canada
trainotti.ricardo@unifesp.br,lberton@unifesp.br

Abstract. Machine translation (MT) plays a crucial role in globalization, making access to information more inclusive, although challenges persist for less popular languages, like Portuguese. One of the most complex challenges in the automatic translation into languages such as Portuguese is the precise preservation of masculine and feminine grammatical gender. There are still situations where translation does not adequately reflect gender equality, often reinforcing societal stereotypes. We aim to explore approaches to ensure fairness in English to Portuguese MT through post-processing techniques, which aim to apply some transformation to the model's output. To this end, we used the MarianMT model as our foundation, then we fine-tuned it using a dataset of English-Portuguese sentences that was generated and carefully crafted to mitigate gender bias within the sentences. The results on gender disparities metrics, based on the WinoMT test set for MT such as ΔG, ΔS, and the overall accuracy (preserving the gender of the entity from the original) significantly improved with some drop in BLEU (Bilingual Evaluation Understudy) score. Our study focuses on addressing gender bias in the Portuguese language. However, it can also be adapted to other languages, since it is crucial to ensure truly fair and inclusive global communication.

Keywords: NLP · Machine Translator · Fairness · Portuguese.

1 Introduction

Natural language processing (NLP) plays a fundamental role in our current society, being an essential technology in several areas. It enables machines to understand, interpret, and generate human text in a similar way to humans [21]. This capability is critically important in an era of massive data, where most information is unstructured and textual. NLP is essential for improving the efficiency and accuracy of search engines, enabling function as smarter chatbots and virtual assistants [1, 23], advanced sentiment analysis on social media [17, 26], automatic translation [20] and text classification [12, 18, 13]. Additionally, it plays a crucial role in accessibility, making technology more inclusive for people with communication disabilities. In short, NLP is a technology that drives innovation in many sectors, facilitating interaction between humans and machines more naturally and effectively.

Based on advances in machine learning (ML) and NLP, machine translation (MT) systems have constantly evolved, providing increasingly accurate and contextual translations [20, 8]. They are widely used in various areas, from business document translations to online international communication, helping to overcome language barriers and promoting global understanding. As the demand for global communication increases, the ability to translate efficiently between less widely used languages becomes increasingly important. Although machine translation technologies have made notable advances in widely spoken languages, translating into less common languages involves several obstacles, such as a lack of sufficient training data and linguistic diversity.

In this work, we aim to explore automatic translation into Portuguese, which is the mother tongue of more than 230 million people around the world[1]. Its importance goes far beyond the borders of Portugal and Brazil and is also spoken in several other countries such as Mozambique, Angola, Cape Verde and Timor-Leste. Furthermore, it is one of the official languages of international organizations, such as the Community of Portuguese-Speaking Countries (CPLP)[2] and the European Union[3], which makes it a crucial vehicle for diplomacy and global trade.

Grammatical gender rules present in some languages are a significant challenge for machine translation. It is possible to identify three language grouping [27]. The first are genderless languages (for example, Finnish and Turkish), in which gender distinctions are minimal and often limited to essential lexical pairings such as kinship or address terms. Notional gender languages (e.g., Danish, English) include languages that distinguish between lexical gender (e.g., mom/dad) and pronominal gender (e.g., she/he, her/him). Finally, grammatical gender languages (for example, French, Greek, German and Spanish) have a classification system that assigns masculine, feminine, and sometimes neuter gender to nouns, frequently with meaning ties to human referents. Gender assignment may be formal for inanimate objects, but it is frequently dependent on meaning for human referents. These languages use a morphosyntactic agreement system, with gender inflections extending to different components of speech such as verbs, determiners, and adjectives [27]. This represents an additional challenge for the task of automatic translation, especially when translating from languages lacking grammatical gender and grammatical gender languages.

Thus, translating professions and gender-related terms can be challenging in many languages, including translating from English to Portuguese [10]. Many gender issues can arise in this context:

– Generic Masculine: Often, in English, terms such as “engineer” or “doctor” are used in the generic masculine, without distinguishing gender. However, when translating into Portuguese, it is necessary to decide whether the term

[1] https://www.babbel.com/en/magazine/how-many-people-speak-portuguese-and-where-is-it-spoken
[2] https://www.cplp.org/
[3] https://european-union.europa.eu/principles-countries-history/languages_en

will be translated into masculine or feminine, which can perpetuate gender stereotypes.
- Gender Ambiguity: Some terms in English, such as “actor” or “waitress”, do not make a gender distinction in the original language. However, when translating them into Portuguese, it is necessary to decide whether they will be translated neutrally or with a corresponding feminine form, such as “actress” or “waiter”.
- Terms of Address and Gender: In some professions, terms of address or titles may vary based on a person’s gender. For example, in English, we use “Mr.” and “Mrs.” (Sir and Madam) to make this distinction. However, in Portuguese, this distinction is more complex, with variations such as “Sr.” and “Sra.” or “Dr.” and “Dra.” depending on the context, the translation needs to be carefully chosen.
- Traditionally Single-Gender Professions: Some professions have historically been associated with a single gender. For example, “nurse” used to be predominantly female, while “pilot” was considered a male profession. When translating these terms, it is important to consider how cultural and gender norms are evolving and adjust to these changes.

Table 1 shows two examples of translations containing gender bias. In the first sentence, the occupation “the mechanic” should have been translated as “a mecânica”, which is the feminine form in Portuguese, because it is related to the pronoun “she” (ela). The same occurs in the second sentence, where “the farmer” should also be translated in the feminine form as “a fazendeira” for the same reason.

Source sentence	**Portuguese translation**
The mechanic gave the clerk a present because she won the lottery.	O mecânico deu um presente ao funcionário porque ela ganhou na loteria.
The CEO helped the nurse because he needed help.	O CEO ajudou a enfermeira porque precisava de ajuda

Table 1. Example of gender bias in machine translation using MarianMT pre-trained model. Entities’ grammatical genders are distinguished by colors: blue represents male entities and pronouns, red signifies female ones, and orange denotes neutral ones.

In this work, we aim to reduce this type of gender bias existing in English-Portuguese automatic translation with a focus on fairness. Fairness is the property that algorithms and systems do not perpetuate prejudices or unfair discrimination [3, 14, 5, 25]. This involves creating and implementing metrics and strategies to mitigate bias and ensure that automated decisions are impartial and equitable for all people, regardless of their ethnic origin, gender, age, or other protected characteristics. Fairness is essential for promoting ethics and

equality in the digital age. Previous work explored fairness concerns in translation such as [24, 30, 16]. However, none delved into post-processing methods or explored the Portuguese language. Therefore, developing machine translation systems capable of handling challenging languages is a promising area of research, with significant implications for promoting cultural diversity and inclusive global communication. The main contributions of our work are summarized as follows:

- We proposed a post-processing fine-tuning approach based on the MarianMT model for gender-neutral English to Portuguese Machine Translators. Post-processing provides the opportunity to rectify errors or biases that may have arisen during data collection or modeling.
- We achieved improvements on gender disparities metrics used in MT such as ΔG, ΔS and the overall accuracy, in comparison to the baseline model, although some drop in BLEU (Bilingual Evaluation Understudy) occurred, which is expected given the fairness $\times$ accuracy trade-off.
- We have created a dataset comprising 10,400 sentences, inspired by those utilized in the WinoMT test but enhanced with custom modifications. This dataset will be made publicly accessible.
- We highlight the importance of studying fairness in NLP systems, especially for languages with fewer resources such as Portuguese. It is crucial to ensure equitable access and representation in AI technologies for underrepresented linguistic communities.

This work is organized as follows. Section 2 mentions other works that previously explored MT. Section 3 provides important concepts used in the work related to Fairness in Machine Learning. Section 4 presents the methodology adopted in the work, the datasets, and the algorithms used. Section 5 shows the results achieved in MT translation. Section 6 presents the final remarks.

2 Related Work

Some works posit that MT tools can be harnessed through the use of gender-neutral languages to offer insights into the issue of gender bias in AI. An exhaustive list of occupational titles sourced from the U.S. Bureau of Labor Statistics (BLS) was used to construct sentences such as "They are an Engineer" (with "Engineer" replaced by the specific job title of interest) in twelve distinct gender-neutral languages, including Hungarian, Chinese, Yoruba, and others [24]. These sentences are subsequently translated into English utilizing the Google Translate API, and they gather data on the prevalence of female, male, and gender-neutral pronouns in the resulting translations. Their findings reveal a pronounced bias towards male defaults within Google Translate, particularly in fields characterized by imbalanced gender representation or stereotypes, such as Science, Technology, Engineering, and Mathematics (STEM) professions. We juxtapose these findings with BLS data on the actual gender distribution within each job

title, illustrating that Google Translate fails to replicate real-world gender demographics.

Authors enhance neural machine translation (NMT) systems by introducing gender information in their work [30]. They created extensive datasets containing speaker information for 20 language pairs and performed experiments that integrated gender data into NMT for multiple language pairs. They employed the OpenNMT-py toolkit which is structured as sequence-to-sequence encoder-decoders utilizing LSTM recurrent units. They demonstrate that the inclusion of a gender feature in an NMT system yields a significant enhancement in translation quality for select language pairs.

Other works proposed a gender-debiased approach for MT. An adversarial learning approach was used to reduce gender bias in a seq2seq machine translation model in [15]. In this approach, a prediction model M with weights W learns to predict an output Y from the input X, while remaining neutral with respect to the protected variable Z. The adversary A tries to predict Z from the model's output predictions $\hat{Y}$.

The study's approach to reducing gender bias hinged on a nuanced adaptation of the Transformer model, targeting the word embeddings within both the encoder and decoder components [16]. To bolster gender neutrality, pre-trained word embeddings were introduced, yielding a diverse array of models, each utilizing distinct pre-trained word embeddings sourced primarily from GloVe . The study explored multiple experimental scenarios, including training models without any pre-trained word embeddings to allow autonomous learning, and incorporating pre-trained embeddings, such as standard GloVe, HardDebiased GloVe, and Gender Neutral Glove (GN-GloVe), derived from the same corpus. Additionally, the investigation delved into the specific utilization of pre-trained embeddings, examining three cases: encoder-only, decoder-only, and both encoder and decoder, to comprehensively evaluate their impact on mitigating gender bias during translation. The authors evaluate the proposed system on the Workshop on Machine Translation[4] (WMT) English Spanish benchmark task. This holistic framework facilitated a comprehensive assessment of embedding strategies, shedding light on effective means of enhancing translation fairness.

In their work on mitigating sensitivity to protected attributes such as gender and age in sentiment classification, authors evaluated round-trip translation as a technique [9]. They demonstrate, in particular, that translating Danish product reviews into English and back minimizes group disparity across three distinct classification structures. They used two different pre-trained language models, namely the multilingual LASER model and a monolingual BERT trained for Danish. On top of these, some classifiers were employed, including nearest neighbor, logistic regression, and (Gaussian kernel) support vector machines (SVMs). The authors discover that round-trip translation at test time reduces the fairness gap (by up to 47%), but that the effect disappears for the best models (SVMs stacked on BERT representations) when both training and test data are translated into a foreign language and back.

[4] http://www.statmt.org/wmt13/

3 Background

3.1 Fairness

Broadly speaking, in ML and NLP studies, fairness is focused on ensuring that systems and algorithms treat individuals and groups fairly without introducing biases or discrimination based on characteristics such as gender, race, or ethnicity. Two important aspects of fairness must be considered: its definition and its metrics [25].

Fairness definition As pointed out by the authors, addressing the bias subject, researchers in AI, Software Engineering, and Law communities have proposed more than twenty different notions of fairness in the last few years [31]. However, there is no consensus regarding the appropriate definition for specific situations. Furthermore, comprehending the intricate distinctions among numerous definitions poses a considerable challenge.

One of the fundamental aspects of fairness definition is the concepts of group fairness and individual fairness. Group fairness focuses on ensuring that outcomes are equitable for different predefined groups (e.g., men and women). On the other hand, individual fairness seeks to treat similar individuals similarly, regardless of their group membership. Despite appearing to be in conflict, individual and group fairness measures do not inherently represent distinct normative principles [2].

Also, fairness definitions often revolve around the concepts of anti-discrimination and equal opportunity. Anti-discrimination aims to prevent biased treatment against any specific group. Equal opportunity, on the other hand, aims to provide individuals from different groups with an equal chance of a positive outcome.

Disparate Impact and Disparate Treatment are two concepts in the realm of discrimination and bias that helps to address and identify different forms of discrimination and play a crucial role in ensuring fairness and equal opportunities. Disparate impact deals with situations where a system's outcomes disproportionately affect one group, even if no explicit bias exists. Disparate treatment relates to situations where individuals from different groups are treated differently due to bias or discrimination.

Fairness metrics Many fairness metrics are based on the concepts from a confusion matrix (show in Table 2), a tabular representation used to assess the effectiveness of a classification algorithm in terms of positive and negative results correctly or incorrectly predicted by the model. Some common fairness metrics include:

- Demographic Parity: Measures whether the positive prediction rates ($TP + FP/TP + TN + FP + FN$) are equal across different groups.
- Equal Opportunity: Ensures that the True Positive Rate ($TP/TP + FN$) is the same across groups.

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Table 2. Example of a confusion matrix.

- Equalized Odds: Requires both True Positive Rate ($TP/TP+FN$) and False Positive Rate ($FP/FP+TN$) to be equal across groups.
- Predictive Parity: Checks if the Precision ($FP/FP+TN$) is the same across groups.
- False Positive Rate Balance: Ensures that the False Positive Rate ($FN/FN+TP$) is equal across groups.
- False Negative Rate Balance: Ensures that the False Negative Rate ($FN/FN+TP$) is equal across groups.

These metrics help assess whether a model is fair across different demographic groups or categories.

Approaches Ensuring that models treat all demographic groups equitably requires careful consideration across the entire machine learning pipeline. Usually three approaches have been explored: pre-processing, in-processing, and post-processing [6].

1. Pre-Processing: In fairness-aware machine learning, pre-processing involves modifying the training data to mitigate biases before the learning process begins. Techniques include re-weighting samples, altering features to remove sensitive information, and generating synthetic data to balance classes across different groups.
2. In-Processing: This stage incorporates fairness constraints directly into the learning algorithm. Modifications can include altering the objective function to penalize unfair predictions or adding constraints that ensure equitable treatment across different demographic groups.
3. Post-Processing: After the model has been trained, post-processing techniques adjust the model's predictions to reduce unfair outcomes. These methods can involve re-calibrating prediction thresholds or applying transformations to the output probabilities to ensure fairness across groups.

When it comes to combating biases in NLP tasks, the primary emphasis is on addressing these issues during the pre-processing phase, as noted in the work by [6]. This entails actions such as removing or substituting specific words, adjusting dictionaries, and applying unsupervised techniques to balance the training dataset.

Bias and fairness in MT Bias evaluation in translations is frequently difficult to do because there isn't a clear ground truth. However, studies such as that of

[4], entitled "Man is to Computer Programmer as Woman is to Homemaker?" show that word embeddings exhibit biases that reproduce the gender stereotypes prevalent in society.

In their study, the researchers began by selecting occupations closely associated with the terms "she" and "he" within the word embeddings they were analyzing. To assess whether these selected occupations indeed reflected gender stereotypes, the researchers engaged individuals to evaluate each occupation and to determine whether it conveyed female stereotypes or male stereotypes or if it was gender-neutral. To quantify the degree of stereotypicality, they employed a rating scale ranging from 0 to 10, where higher scores indicated a stronger association with gender stereotypes. Upon analysis, the study found the positions of these selected occupation words along the "she-he" axis within the word embeddings exhibited a substantial correlation with the ratings assigned. In other words, the geometric properties of the embeddings closely reflected the human judgments of gender stereotypes.

WinoMT, a challenge set for evaluating gender bias in machine translation using a concatenation of Winogender and WinoBias, was presented in [29]. The evaluation dataset comprises $3,888$ sentences designed to probe potential gender bias within the translations. In each sentence, a primary entity, coreferent with a pronoun, interacts with a secondary entity and it seeks to reveal gender bias tendencies in the translation system. The set has an equal balance of both male and female genders, as it does with stereotypical (e.g., a male doctor) and nonstereotypical (e.g. a female engineer) gender-role assignments. Their methodology for evaluating machine translation systems begins by calculating the overall accuracy, which is determined by assessing the percentage of instances in which the translation successfully maintains the gender of the entity from the original English sentence. Their findings reveal that across eight different languages, most of the tested machine translation systems perform poorly in preserving gender accuracy. Even the best-performing model for each language typically exhibits performance that is not significantly better than random guessing when correctly inflecting gender in the translations.

ΔG represents the disparity in F1 scores between sentences containing masculine entities and those featuring feminine entities. They point out that all evaluated systems, except one, perform better on male roles, which may indicate these are more frequent in the set used for training. Finally, ΔS quantifies the discrepancy in accuracy when translating the antecedent in sentences with pro-stereotypical and anti-stereotypical role assignments. According to the authors, the results of this metric show that all tested systems have a meaningful better performance when presented with pro-stereotypical assignments (e.g., a female housekeeper), as their performance worsens when translating anti-stereotypical roles (e.g., a male librarian).

Despite the inclusion of a diverse set of eight languages (Spanish, French, Italian, Russian, Ukrainian, Hebrew, Arabic, and German) in the WinoMT test, it is worth noting that Portuguese was not part of the selection.

4 Methodology

This section will detail the methodology adopted. Thus, Section 4.1 details the post-processing fairness techniques used, Section 4.2 details the datasets used, and Section 4.3 details the steps for executing the experiments.

4.1 Post-Processing Fairness Techniques

Post-processing techniques are a well-known method for addressing potential biases in model outputs related to protected variables or subgroups in the field of machine learning fairness [6]. These techniques stand out for their remarkable flexibility because they don't need access to the internal models or algorithms; instead, they only depend on the model's predictions and sensitive attribute data [6]. These techniques are especially well suited for "black-box" scenarios, which are those in which the entire ML pipeline is not completely transparent.

As a post-processing approach, fine-tuning holds significant promise in reducing gender bias in machine translation. In the context of this study, fine-tuning involves adapting the MarianMT pre-trained machine translation model using a gender-balanced dataset. By incorporating a diverse set of gender-specific examples into the fine-tuning process, the model becomes more attuned to gender nuances without undermining translation fluency or accuracy.

4.2 Datasets

In our endeavor to reduce gender bias in machine translation, we adopted a strategic approach that involved the creation of a specialized dataset for fine-tuning the MarianMT model. This dataset, consisting of 100,400 parallel English-Portuguese sentences, was crafted to balance gender bias reduction while maintaining high translation quality. To achieve this, we strategically combined 90,000 sentences from the CAPES TDC corpus [28] and 10,400 artificial sentences, generated specifically for this task. CAPES corpus is a trusted source that had initially been employed in the original training of the MarianMT model. This corpus was compiled from the abstracts of all theses and dissertations produced in Brazil between 2013 and 2016. This choice was driven by the intention to preserve the translation quality and prevent overfitting that might occur when relying solely on artificial data.

These artificial sentences were thoughtfully generated, drawing inspiration from the sentences used in the WinoMT test but with tailored modifications. By blending authentic sentences from the CAPES corpus with designed artificial examples, we aimed to balance maintaining translation quality and enhancing the model's ability to address gender bias. This approach acknowledges the importance of real-world translation dynamics, ensuring that our fine-tuning process remains effective in reducing gender bias while upholding the standards of translation quality established by the original model.

4.3 Experimental Setup

The experimental setup for this research was conducted on the "Béluga" cluster, a high-performance computing resource provided by Digital Research Alliance of Canada, equipped with substantial hardware resources. The hardware configuration included four NVidia V100SXM2 GPUs, each equipped with 16GB of RAM.

In terms of software, the experiment was conducted using a well-established stack of tools and libraries. Python 3.10.1 served as the primary programming language. PyTorch 2.0.1, a widely recognized deep learning framework, was employed for model development and training. Natural Language Toolkit (NLTK) 3.8.1 facilitated text preprocessing and linguistic analysis, while Pandas 2.0.3 offered efficient data manipulation capabilities. The Transformers' library, version 4.31.0, played a pivotal role in facilitating the fine-tuning process of the MarianMT model, streamlining the integration of transformer-based architectures into the research workflow.

The parameters were configured to tailor the fine-tuning process. The model's decoder layer dropout was set to 0.2, contributing to the regularization of the model during training. The fine-tuning parameters were also defined, encompassing a learning rate of 1e-5, a batch size of 8, and a total of 10 training epochs. The optimizer selected for this task was AdamW.

4.4 Translation and Gender Bias Evaluation

For the evaluation process, we have developed custom scripts inspired by the original methodology WinoMT [29], tailored to our specific English-Portuguese translation context. These scripts automatically extract the grammatical gender assigned to the primary entity within each translation. Following this extraction, a comparison is made between the gender of the translated primary entity and the gender annotated in the gold standard data. Our objective is to gauge the extent to which our translation models align with the gender of the primary entity, as per the gold annotations. The key performance metrics for evaluating the WinoMT dataset include ΔG, ΔS, and the overall accuracy of preserving the gender of entities during translation, referred to as **"acc."**.

The study compared the quality of translations before and after fine-tuning using the BLEU (Bilingual Evaluation Understudy) [22] metric to determine whether it had a negative effect on translation quality. BLEU is a well-established metric for machine translation evaluation that measures the overlap between the model machine translation and reference translation. It is calculated based on accuracy, which measures the percentage of n-grams in the machine translation that also appear in the reference translations. The BLEU score ranges from 0 to 1, with 1 being a perfect match with the reference translations [22].

Utilizing the MarianMT model as our foundation, we conducted fine-tuning using a dataset comprising English-Portuguese sentences. This dataset was self-generated to include sentences that mitigate gender bias. Ultimately, upon evaluating the model, we observed a decrease in translation accuracy concurrent with a reduction in gender bias.

5 Results and Discussion

The experimental outcomes underscore success in achieving the primary goal of gender bias reduction in machine translation through fine-tuning. The three fairness metrics, ΔG, ΔS, and "acc." from the WinoMT test, exhibited improved results post-fine-tuning, as shown in Table 3. These metrics, designed to assess the model's capacity to produce fair and equitable translations, demonstrated clear progress in mitigating gender bias, aligning with the objective of the research. These positive results indicate that the fine-tuning process significantly improved the model's ability to generate translations that respect gender neutrality and balance, a pivotal step toward creating more inclusive and unbiased machine translation systems.

However, it's worth noting that these promising improvements in fairness metrics were accompanied by a trade-off, as anticipated, in terms of the BLEU score. The reduction in the BLEU score underscores the well-established fairness-accuracy trade-off[5] that often accompanies efforts to reduce bias in machine translation. Nevertheless, the 0.38 score of fine-tuned model tells that it still produces good translations[6]. While the primary focus of the research was to enhance fairness and mitigate gender bias, this trade-off in translation quality serves as a reminder of the intricate balance that exists between these two objectives and indicates the need to continue the research process in order to achieve better results.

	Pre-trained	Fine-tuned
ΔG	0.1367	**0.0148**
ΔS	0.1319	**0.0871**
Acc.	0.6021	**0.7086**
BLEU	**0.54**	0.38

Table 3. Results for bias metrics and translation quality on pre-trained MarianMT model and fine-tuned MarianMT model.

5.1 Discussion

The fine-tuned MarianMT correctly translates the occupations "the mechanic" and "the farmer" to the feminine forms in Portuguese, see Table 4. This suggests that the model is capable of producing gender-neutral or gender-sensitive translations when dealing with gendered occupational terms.

[5] Most bias reduction algorithms harm the performance of the prediction model in the process of making it fair, which is commonly referred to as the fairness-accuracy trade-off [7].

[6] https://cloud.google.com/translate/automl/docs/evaluate

Source sentence	Portuguese translation
The mechanic gave the clerk a present because she won the lottery.	A mecânica deu à funcionária um presente porque ela venceu a lotaria.
The CEO helped the nurse because he needed help.	O CEO ajudou o enfermeiro porque ele necessitava de ajuda.

Table 4. Example of translations generated using fine-tuned MarianMT. Entities' grammatical genders are distinguished by colors: blue represents male entities and pronouns, red signifies female ones, and orange denotes neutral ones.

However, despite reducing the ΔS metric compared to the original model, the fine-tuned model, like all systems tested in [29], also makes more translation errors in sentences with anti-stereotyped gender assignment (e.g. a female farmer) than in stereotyped roles (e.g. a male lawyer), although to a lesser degree than before the fine-tuning.

Even though the test's focus is related to the stereotyped gender assignment in occupations, the test includes a small sample (240 sentences) in which the main agent has a gender-neutral role. In these cases, the ability of the fine-tuned model to preserve the gender (acc.) is only 0.2292, falling even below the model before fine-tuning, which is 0.3125. This limitation is primarily found in sentences like "*The technician told the customer that* ***they*** *could pay with cash.*" or "*The employee gave the supervisor feedback on* ***their*** *managing style.*". In these cases, "they" and "their" are been employed to refer to an individual in a generic or gender-neutral context. According to [11] this use particularly provides a general reference to that individual without specifying their gender (like *she* or *he*). Although, for cases like this, there are proposals for the use of gender-neutral pronouns in Portuguese, such as "elu"[7], the MarianMT translator model, and consequently the fine-tuned version, do not include this variation.

5.2 Limitations and future work

While the results of this research have demonstrated the effectiveness of fine-tuning as a technique for reducing gender bias in MT, it is important to acknowledge some limitations in this work. The dataset used in this study was specifically crafted based on the sentences of the WinoMT test. While it was designed to be well-suited for this test, it may not fully represent the complexities of gender bias in "real-world" translation scenarios, which may require further investigation. Also, the test primarily focuses on gender bias related to occupation, which is undoubtedly significant. However, gender bias in MT extends beyond occupational contexts, which can be explored in future works.

[7] According to [19], in Portuguese, the most commonly used gender neuter pronoun is "elu".

Building upon the outcomes of this research, there are several avenues for future work on reducing gender bias in machine translation. It is essential to continue research into improving fairness metrics translations while ensuring these enhancements do not significantly reduce translation quality. Exploring alternative datasets or adjusting fine-tuning parameters can be a promising avenue. Future research can focus on enhancing MT systems to include translations incorporating neutral Portuguese language. This would help bridge the gap in providing translations that include gender-diverse language choices and align with evolving societal norms.

6 Conclusions

Machine Translation tools have significantly contributed to communication, enabling seamless global interactions and bridging cultural gaps through the Internet. However, as the field of fairness in Machine Learning has grown, concerns about equity have extended to MT tools, especially regarding their potential to propagate gender bias present in training data.

In this work, we explored the theme of gender bias in English-Portuguese machine translation and proposed a bias reduction method using fine-tuning on the pre-trained model MarianMT. The significance of studying fairness in English-to-Portuguese MT is primarily linked to the limited research focus on the Portuguese language. Therefore, research of this nature tends to contribute to the advancement of fairness research and the development of translation tools with a reduced degree of bias, benefiting a community of more than 230 million Portuguese speakers worldwide.

Language is a powerful tool for shaping perceptions, reinforcing stereotypes, and potentially influencing attitudes. Gender biases present in the texts used as training data can perpetuate harmful representations of individuals and communities, leading to potential discrimination and distorted perceptions. By understanding and addressing these biases, we can work towards creating inclusive and equitable communication platforms that respect the dignity and diversity of all users. In this regard, our research aimed to contribute to this goal by investigating gender bias in MT and developing techniques to mitigate its effects.

The experiments conducted in this study demonstrated the effectiveness of fine-tuning as a technique for reducing gender bias in machine translation models. During this fine-tuning process, the model adapts its existing knowledge to the new dataset (in this case a gender-balanced parallel corpus), adjusting its parameters to make less biased translations, while aiming to preserve the translation quality of the original model.

One of the main concerns when applying bias reduction techniques is the fairness-accuracy trade-off. Our experiments have shown that fine-tuning can reduce bias without significantly sacrificing translation accuracy. Comparisons of BLEU scores before and after fine-tuning indicate that, despite some loss, the generated translations remain of good quality.

Finally, it is worth noting that the work presented here represents only a small contribution towards equity in Machine Translation. It is necessary to expand research and explore new avenues to enhance bias detection, understand the implications of various types of bias, and refine mitigation techniques, ensuring that translations are fair and high-quality.

Acknowledgements Authors thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

1. Adamopoulou, E., Moussiades, L.: An overview of chatbot technology. In: IFIP international conference on artificial intelligence applications and innovations. pp. 373–383. Springer (2020)
2. Binns, R.: On the apparent conflict between individual and group fairness (2019)
3. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050 (2020)
4. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings (2016)
5. Broder, R.S., Berton, L.: Performance analysis of machine learning algorithms trained on biased data. In: Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. pp. 548–558. SBC (2021)
6. Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Computing Surveys (2020)
7. Chakraborty, J., Majumder, S., Yu, Z., Menzies, T.: Fairway: a way to build fair ML software. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 654–665. ACM, Virtual Event USA (Nov 2020)
8. Chauhan, S., Daniel, P.: A comprehensive survey on various fully automatic machine translation evaluation metrics. Neural Processing Letters pp. 1–55 (2022)
9. Christiansen, J.G., Gammelgaard, M., Søgaard, A.: The effect of round-trip translation on fairness in sentiment analysis. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4423–4428 (2021)
10. Cunha, C., Cintra, L.: Nova gramática do português contemporâneo. LEXIKON Editora Digital ltda (2016)
11. Dictionary, O.E.: they, pron., sense i.2.b (9 2023), https://doi.org/10.1093/OED/9782781428
12. Duarte, J.M., Berton, L.: A review of semi-supervised learning for text classification. Artificial intelligence review **56**(9), 9401–9469 (2023)
13. Duarte, J.M., Sousa, S., Milios, E., Berton, L.: Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations. Information Sciences **570**, 278–297 (2021)
14. Field, A., Blodgett, S.L., Waseem, Z., Tsvetkov, Y.: A survey of race, racism, and anti-racism in nlp. arXiv preprint arXiv:2106.11410 (2021)
15. Fleisig, E., Fellbaum, C.: Mitigating gender bias in machine translation through adversarial learning (2022)

16. Font, J.E., Costa-Jussa, M.R.: Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv preprint arXiv:1901.03116 (2019)
17. Garcia, K., Berton, L.: Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. Applied soft computing **101**, 107057 (2021)
18. Garcia, K., Shiguihara, P., Berton, L.: Breaking news: Unveiling a new dataset for portuguese news classification and comparative analysis of approaches. Plos one **19**(1), e0296929 (2024)
19. Lau, H.D.: O uso da linguagem neutra como visibilidade e inclusão para pessoas trans não-binárias na língua portuguesa: a voz del@ s ou delxs? não! a voz delus. V Simpósio Internacional em Educação Sexual: saberes/trans/versais currículos identitários e pluralidades de gênero. Anais do V Simpósio Internacional em Educação Sexual: saberes/trans/versais currículos identitários e pluralidades de gênero. Masringá (2017)
20. Maruf, S., Saleh, F., Haffari, G.: A survey on document-level neural machine translation: Methods and evaluation. ACM Computing Surveys (CSUR) **54**(2), 1–36 (2021)
21. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. Journal of the American Medical Informatics Association **18**(5), 544–551 (2011)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. p. 311318. ACL '02, Association for Computational Linguistics, USA (2002)
23. Prado, C., Netto, A.V., Berton, L., Takahara, A.K.: Aplicação de healthbots em língua portuguesa: revisão narrativa. Journal of Health Informatics **13**(4) (2021)
24. Prates, M.O., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: a case study with google translate. Neural Computing and Applications **32**, 6363–6381 (2020)
25. Rabonato, R.T., Berton, L.: A systematic review of fairness in machine learning. AI and Ethics pp. 1–12 (2024)
26. Santos, D.K.S., Berton, L.: Analysis of twitter users' sentiments about the first round 2022 presidential election in brazil. In: Anais do XX Encontro Nacional de Inteligência Artificial e Computacional. pp. 880–893. SBC (2023)
27. Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., Turchi, M.: Gender Bias in Machine Translation. Transactions of the Association for Computational Linguistics **9**, 845–874 (08 2021)
28. Soares, F., Yamashita, G.H., Anzanello, M.J.: A parallel corpus of theses and dissertations abstracts. In: Lecture Notes in Computer Science, pp. 345–352. Springer International Publishing (2018)
29. Stanovsky, G., Smith, N.A., Zettlemoyer, L.: Evaluating gender bias in machine translation. In: ACL. Association for Computational Linguistics, Florence, Italy (6 2019)
30. Vanmassenhove, E., Hardmeier, C., Way, A.: Getting gender right in neural machine translation. arXiv preprint arXiv:1909.05088 (2019)
31. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. p. 17. FairWare '18, Association for Computing Machinery, New York, NY, USA (2018)