Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 31, Num. 2 (2024) 74-90

RESEARCH ARTICLE

UnbIAs - Google Translate Gender Bias Mitigation and Addition of Genderless Translation

UnbIAs - Google Tradutor Com Mitigação de Viés de Gênero e Tradução Neutra

Vanessa Schenkel[1], Blanda Mello[1], Sandro José Rigo[1], Gabriel de Oliveira Ramos[1]*

Abstract: Machine Learning, increasingly present in everyday life, is subject to bias. These biases not only reflect social inequalities but also reinforce them. The present study seeks to mitigate gender bias in Google Translate, the most used translation system in the world. For this, we created a translation model with high gender accuracy and performed a linguistic analysis with the spaCy tool and entity identification with roBERTa. The Constrained Beam Search technique is used to maintain the sentence structure of the business model, but with the replacement for the correct genre indicated by the created model. The final sentence is the result of an alignment done with the SimAlign tool. In addition, the present study also produces an algorithm so that sentences without gender indication in English present translations with inflection for feminine, masculine, and neutral gender. Our approach yields a BLEU score of 48.39. In relation to Google Translate, the model increased gender accuracy from 68.75 to 70.09, enhanced in 15.7% the score that measures the difference in accuracy between male and female entities, and improved stereotyped translations in 43%.
Keywords: Language Models — Gender bias — Google Translator — Constrained Beam Search — Non-binary gender

Resumo: A aprendizagem de máquina, cada vez mais presente no cotidiano, está sujeita a vieses. Estes vieses não apenas refletem desigualdades sociais como as reforçam. O presente estudo busca mitigar o viés de gênero no Google Tradutor, o sistema de tradução mais utilizado no mundo. Para isto, foi criado um modelo de tradução com uma alta acurácia de gênero. Para tanto, realizou-se uma análise linguística com a ferramenta spaCy e identificação de entidade com roBERTa. É utilizada a técnica de Constrained Beam Search para se manter a estrutura da frase do modelo comercial, porém com a substituição para o gênero correto apontado pelo modelo criado. A frase final é o resultado do alinhamento feito com a ferramenta SimAlign. Além disso, o presente estudo também produziu um algoritmo para que frases sem indicação de gênero em inglês apresentem as traduções com flexão para feminino, masculino e neutro. A abordagem aqui proposta apresentou pontuação BLEU de 48.39. Em relação ao Google Tradutor, o modelo melhorou a acurácia de gênero de 68.75 para 70.09, elevou em 15.7% a pontuação que mede a diferença de acerto entre entidades masculinas e femininas, e melhorou traduções estereotipadas em 43%.
Palavras-Chave: Modelos de linguagem — Viés de gênero — Google Tradutor — Constrained Beam Search — Gênero não-binário

[1] *Graduate Program in Applied Computing, Universidade do Vale do Rio dos Sinos, São Leopoldo, Rio Grande do Sul, Brazil*
***Corresponding author**: gdoramos@unisinos.br
DOI: https://doi.org/10.22456/2175-2745.139902 • **Received:** 29/04/2024 • **Accepted:** 01/07/2024

1. Introduction

The presence of artificial intelligence (AI) is evident in everyday life. We use it to perform simple tasks, such as completing a word by typing [1], to more complex tasks such as diagnosing a disease [2, 3]. Artificial intelligence solves problems that have a high social impact and has the possibility to make decisions without prejudice or social stigma. However, machine learning systems has already concluded that darker skin is unattractive [4], women are less qualified than men [5], black people are more likely to re-offend criminally [6], and created a neo-Nazi chatbot in a few hours [7].

As the use of artificial intelligence grows, so does the concern about algorithmic fairness. It is proven that it can be harmful against underrepresented groups [4], as the algorithm can be trained with biases or even can inherit social biases from those who program them.

Translation systems use machine learning, so they are

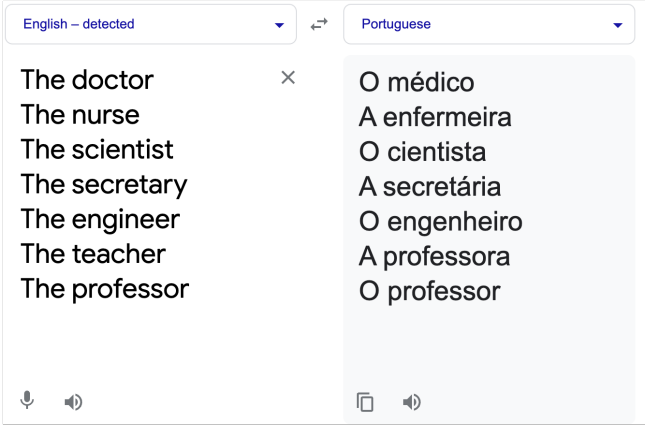

Figure 1. This figure illustrates the gender bias in Google Translate when translating gender-neutral professions from English to Portuguese. The professions *doctor*, *nurse*, *scientist*, *secretary*, *engineer*, *teacher*, and *professor* are translated with gender-specific articles in Portuguese. The translations reflect traditional gender roles, such as *O médico* (male doctor) and *A enfermeira* (female nurse), indicating a bias towards associating certain professions with specific genders. This highlights the need for tools like unbIAs to address and mitigate such biases in machine translation.

also subject to bias. The most used translator in the world is Google Translate[1]. There are more than 500 million daily users on the site, with Brazil being the country that uses it the most[2]. In 2021, its Android app surpassed 1 billion downloads, with support to 109 languages[3]. Google Translate uses Google Neural Machine Translation, which uses a large, deep artificial neural network. However, like any system that uses artificial intelligence, Google Translate is subject to bias. These biases may not only reflect social inequalities but also reinforce them [8, 9].

It is possible to verify the presence of gender bias when translating a word without gender indication in English. Figure 1 brings examples of resulting translations of neutral professions. These professions do not have gender inflection in English, but in Portuguese they do have gender. For example, *The doctor* is translated as *O médico* (male doctor) and *The nurse* as *A enfermeira* (female nurse). The work presented in this article seeks to mitigate bias not just in a specific type of phrase, such as those with professions, but in any phrase that is translated from English to Portuguese.

In addition to gender bias, there is the exclusion of non-binary people who currently represent 2% of the Brazilian population[4]. Despite the growth of social inclusion, such as records with more than two gender options, the same does not occur in translation models or in academic research.

The present study aims to identify and mitigate Google Translate's gender bias and introduce a new translation tool. For this, tests were carried out with different pre-trained and trained translation models. Each model was trained with different datasets. The fine-tuning of a model was also carried out with a balanced dataset created by hands specially created for this study. The results were analyzed and compared so that we could choose the model with the best gender accuracy result. After this process, an algorithm was created that identifies the entities of the sentence, whether or not there is an indication of gender and that performs the morphological analysis of the user's input sentence. The algorithm then uses the previously selected model and performs the constrained beam search technique. This combination of the model with high gender accuracy and the algorithm that allows the construction of the commercial translator to be maintained, allows it to mitigate gender bias and provide quality translations of any input provided by the user.

By ensuring translations that respect all gender nuances, including the specificities of non-binary individuals, the tool developed for the present study not only enhances linguistic accuracy but also promotes visibility and respect for identities that are frequently marginalized. This effort to reduce gender bias in AI systems marks a significant advancement towards gender equality, challenging entrenched stereotypes and prejudices that are perpetuated by inaccurate or insensitive translations. By fostering inclusive language, we not only mitigate direct gender bias but we also aim to educate and raise awareness among users, developers and other academics about the importance of gender accuracy and sensitivity. This educational aspect is crucial for promoting broader awareness of gender diversity and for driving social changes that value and respect the experiences of all individuals, regardless of their gender identity.

Previous studies in the field of ethics in Artificial Intelligence, recommend to address biases in data sets before model training[10]. Approaches to mitigating gender bias in translations include adjustments to the data set[11], the use of embeddings[12], and the inclusion of gender characteristics during training and inference[13, 14]. Techniques such as Constrained Beam Search have been adapted to increase gender diversity in translations[15], although they do not necessarily improve the quality of the translation or the BLEU score. Our research aims to balance translation quality with bias mitigation[16].

To the best of our knowledge, this is the first work to mitigate gender bias in a business model by focusing on translation from English to Portuguese. It is also the first work that allows the entry of sentences to be translated by the user and that includes neutral translations, focusing on the inclusion of non-binary people.

The main contributions of this work can be enumerated as follows:

- We created a model with greater gender accuracy by per-

[1] ⟨https://preply.com/en/blog/how-accurate-is-google-translate/⟩
[2] ⟨https://blog.google/products/translate/ten-years-of-google-translate/⟩
[3] ⟨https://blog.google/products/translate/one-billion-installs/⟩
[4] ⟨https://ibdfam.org.br/noticias/9307/Cerca+de+2+em+cada+100+brasileiros+s%C3%A3o+transg%C3%AAneros+e+n%C3%A3o+bin%C3%Arios%2C+revela+search⟩

forming fine-tuning with a small, balanced dataset. The model has been published on Hugging Face, allowing easy access and application by developers worldwide [5];
- We developed an algorithm to mitigate bias with the implementation of the constrained beam search generated from Google Translate translations to maintain the quality of the translations;
- We created a tool called unbIAs that combines our model and the algorithm to offer translations with greater accuracy of gender and neutral translation. The entire codebase has been made publicly available, ensuring transparency and encouraging community contributions [6];
- We created dataset with over 50,000 words annotated with labels for word type, part-of-speech, and all possible forms of the word such as singular, plural, feminine, and masculine, which enriches the linguistic nuances captured in gender bias evaluation [7].
- We improved and extended WinoMT to include Portuguese [8];

In this article, we explore the complex theme of gender-neutral machine translation. The research is framed within the context and motivation that drives our work, aimed at addressing specific gaps identified in existing studies. Our approach is comprehensive, systematically unfolding through a series of well-defined methodologies and evaluative metrics, ensuring a thorough investigation into the subtleties of gender neutrality in translation.

The structure of the article is divided into five main sections for clarity and depth. Section 1, "Introduction," introduces the research theme and establishes the context and motivation behind our work. Section 2, "Background and Related Work," summarizes previous studies, identifying the existing gaps that our research addresses. Section 3, "Methodology," details our specific approaches, breaking down into sub-sections on the translation model, algorithm, and evaluation metrics. Section 4, "Results and Discussion," presents our findings and discusses the performance and implications of our methods. Finally, Section 5, "Conclusion and Future Work," summarizes our study's contributions and explores potential future research directions. This organized layout allows readers to easily understand and follow our research progression and findings.

2. Background and related work

There are 23 possible origins of bias in artificial intelligence [4]. In the area of Natural Language Processing (NLP), bias can arise through the dataset on which the model is trained, through the resources used, in the pre-trained model, or in the [17] algorithms themselves.

The identification of gender bias in Natural Language Processing can be performed through psychological tests. In [18, 1], volunteers are used to analyze whether the bias demonstrated by machine learning reflects social bias and evaluation of results. Another technique involves space analysis between words in a vector, where gender bias is defined as the correlation between the magnitude of the projection in the gender subspace in a word embedding representing a neutral word [18].

Measuring the difference in performance between different genders, races, or any categories is a common way to check for [17] bias. According to study [17], if the model does not make decisions based on gender, it should perform equally for both phrases. Otherwise, the difference between assessment scores reflects the extent of gender bias found in the system.

Finally, it is possible to identify gender bias when analyzing the distribution of neutral sentences translated into feminine, masculine, or neutral. In the article [9], a comparison is made of what was translated from professions to masculine or feminine and how much this reflects real data. The authors state that the translation of professions is much more linked to a social stigma than to reality.

2.1 Techniques for Mitigating Gender Bias

Recent recommendations in the area of ethics in Artificial Intelligence suggest that biases or social imbalances in a dataset can be addressed before training the [10] model. There are different techniques for mitigating gender bias in translation tasks.

One approach is the use of gender tagging, as training datasets can be dominated by data points of male origin, so models learn skewed statistical relationships. Thus, they are more likely to predict that the speaker is male when the source is ambiguous [13]. Gender tagging mitigates this by adding a *tag* indicating the gender of the data source, pointing to the beginning of the sentence. This approach helps preserve gender context and allows the model to create more accurate translations. The main advantage is its simplicity and direct impact on preserving gender information, although it requires careful tagging and can be challenging for large datasets.

Another technique is training models with balanced datasets. For this, the number of sentences must be increased or reduced, aiming to have the same number of sentences of all genres for training. Studies that mitigate translation bias focused on the dataset to face two challenges: they are not easy to apply in real problems, and they are not practical [16]. This approach is straightforward and effective but can be impractical in real-world applications due to the difficulty of obtaining balanced datasets.

Another possibility is to use word embeddings to perform the mitigation [18]. The gender subspace is identified, and the total space is reduced. There is a neutralization that moves the gendered subspace away from neutral words and ensures that all neutral words are removed and zeroed out of the gendered

[5] ⟨https://huggingface.co/VanessaSchenkel/pt-unicamp-handcrafted⟩
[6] ⟨https://github.com/ramos-ai/unbIAs⟩
[7] ⟨https://huggingface.co/datasets/VanessaSchenkel/pt-inflections⟩
[8] ⟨https://github.com/ramos-ai/winopt⟩

subspace. Subsequently, equalization is performed to ensure that the neutral words are equidistant from the set of gendered words. This method helps maintain the integrity of the original embedding while reducing bias, although it requires detailed knowledge of embedding spaces and may not completely eliminate bias.

Some mitigation methods adjust the models' predictions, known as algorithm tuning methods. Translation models exacerbate the biases present in the training set [17]. For example, if 80% of references to *secretary* are female in a training dataset, and a model trained on that set predicts 90% coreference on a test to be female, then that means the model broadens the trend. One of the proposed techniques [19] is Reduction of Bias Amplification (RBA) which applies an optimization function of an existing model and constrains this function to ensure that its predictions fit the defined conditions. This approach can fine-tune models to reduce bias amplification, but it can involve complex optimization processes.

Another innovative approach involves the use of Generative Adversarial Networks (GANs). Sun et al. [17] used GANs to prevent the discriminator from identifying gender in a task. The generator tries to prevent the discriminator from identifying gender in a given task, such as completing an analogy. This method can effectively obscure gender information but requires significant computational resources and expertise in GANs.

Integrating external information into the translation process can also mitigate gender bias. Miculicich-Werlen and Popescu-Belis [20] improved pronoun translation by integrating cross-sentence coreference links. Vanmassenhove et al. [13] incorporated gender features both in training data and during inference. Moreover, the implementation of gender tags at the word level, as discussed by [14], serves to further reduce biases in translation outputs. These approaches enrich the model's understanding and improve translation accuracy but add complexity to model training and inference.

The use of constrained beam search has also been explored to generate better translations [21, 22]. However, the [15] study applies a technique to increase the gender diversity of the translations and thus mitigate the bias. The study performs traditional Beam Search to obtain the best `N` translations. The best one is chosen, and the *constraints* are identified. After that, Constrained Beam Search is performed to generate `N` translations. The pronoun or entity is identified, and the entities are aligned. Finally, a classification is made using points according to the gender agreement of the sentences generated by the previous step.

2.2 Comparison and Discussion

Comparing these approaches, some key differences and similarities emerge. Data-centric approaches such as balanced datasets and gender tagging focus on improving training data, while model-centric approaches such as word embeddings and algorithm tuning modify the model's behavior. Data-centric approaches are generally easier to implement, but can be limited by data availability. Model-centric approaches offer more flexibility, but requires more technical knowledge.

Each approach has its advantages and disadvantages. Gender tagging and balanced datasets are simple to implement and directly address gender bias. Nevertheless, they require extensive data preparation and may be impractical for large datasets. Word embeddings and algorithm tuning are flexible and can be applied to existing models. Although, they require detailed technical knowledge and may not completely eliminate bias. Integrating external information enriches context understanding and improves translation accuracy, but it also increases model complexity. Constrained beam search balances translation quality and gender bias mitigation, but it can be computationally intensive and requires detailed linguistic resources.

As seen, related work uses models trained by the authors and not commercial systems. This creates a usability gap, since commercial systems have more users and thus generate a greater social impact. Furthermore, the mentioned techniques generate better gender accuracy, but in a determined set of sentences.

The present study builds on these approaches by proposing a methodology that integrates constrained beam search with external information to mitigate gender bias in Google Translate. This approach aims to balance translation quality and gender bias mitigation, providing inclusive translations for non-binary individuals, for any input received by the user. Another important gap to be highlighted is the limited presence of studies that include translation from English to Portuguese, the focus of this work. With the intention of filling those gaps, in the next sections we detail our approach that aims to balance translation quality and bias mitigation.

3. Methodology

The present study seeks to identify and mitigate gender bias in Google Translate. To that end, we create a pipeline to process the translation considering the bias scenario. One of the steps is fine-tuning the model with the highest score, called Lite with a set of balanced data, detailed in Section 3.1. After this step, we created an algorithm that receives a sentence in English as input and provides the correct translations in Portuguese, as explained in Section 3.2. For this, our algorithm performs a linguistic analysis of the sentence and creates *constraints* with the translation performed by Google Translate, aligning the translations provided by the model created by the study and performing gender inflection. The extension of the WinoMT tool for Portuguese is also created, as detailed in Section 3.3. Finally, to assist other researchers in replicating our experiments precisely, we provide detailed step-by-step guidelines and supplementary materials in Section 3.4.

A detailed pipeline of the proposed method can be seen in Figure 2. The process begins with the input sentence *The mechanic charged the auditor one hundred dollars although she had done no work.* The pipeline includes several key steps:

- **Gender Indication Check:** The sentence is checked for gender indications.
- **Entity Identification:** Entities like *mechanic* and *auditor* are identified using SpaCy.
- **Model Translation and Beam Search:** The model translates the sentence, and constrained beam search is used to generate gender-neutral translations.
- **Alignment and Gender Inflection:** Translations are aligned with the original sentence using SimAlign, and gender inflection adjustments are made using roBERTa and SpaCy.
- **Final Translations:** The final step produces translations that are gender-neutral or include appropriate gender indicators as needed.

We used Python to create the algorithm and train the model on the extension back-end. For the front-end, we used React. We performed the linguistic analysis with spaCy[9], and the identification of which entity the pronoun refers to was performed with roBERTa[10] [23], and the alignment between model translations was done with SimAlign[11] [24].

3.1 Translation model

In order to mitigate gender bias in Google Translate translations, we created a translation model. For better results, tests were performed with different pre-trained and trained models in Section 3.1.2 and different data sets in Section 3.1.1. The results were compared and the model with the best score was chosen for the implementation of the algorithm in the next step in Section 3.1.3.

3.1.1 Dataset

Following a previous study [16], three sets of data were selected and one was created for training a pre-trained model and for fine-tuning it. For model training, the dataset Opus Books[12], News Commentary[13] and TED 2013 [14] were selected. The datasets were selected because they have translations from English to Portuguese and because they are hosted in the HuggingFace[15] website, without the need to download the files. For the fine-tuning step, we created our own dataset, which we call *handcrafted* henceforth.

We have selected Opus Books as the first dataset to model training. The set is formed of 1404 pairs of sentences and 51,650 words, formed by sentences taken from books without copyright, such as Alice in Wonderland. The News Commentary was also selected, which is a parallel corpus obtained by WMT19[16]. This dataset supports 15 languages. The English-Portuguese set, which has 53,135 pairs of sentences and 2.49 million words, was used. Finally, TED2013, with 154,848 and 5.12 million words, is made up of English-Portuguese phrases taken from TED Talks in 2013.

In advancing the work of [14], we have enriched the research landscape by translating an existing dataset into Portuguese, a step not undertaken in the original study. This dataset, meticulously constructed based on occupational categories from US labor statistics as detailed by [9], simplifies specific job titles into broader categories to mitigate bias. For example, rather than using diverse titles like *industrial engineer*, *locomotive engineer*, *civil engineer* and *electrical engineer*, the dataset employs the generalized term *engineer*. Each sentence is structured to follow a standardized format:

```
The [PROFESSION] finished [his/her] work.
```

From this framework, we translated 194 English sentences, effectively doubling the dataset to 388 sentences to accommodate gender variations. Our translation work provides a critical resource for further research and development in bias mitigation within Portuguese language machine translation contexts. This handcrafted dataset, a testament to our rigorous efforts, is publicly available to the research community[17].

3.1.2 Model

To choose a model, we performed tests between three models and the selected data sets. The first one selected is mBART [25], a pre-trained multilingual *sequence-to-sequence* model. mBART is widely used and provides translation into 50 languages. To test with pre-trained models focused on English-Portuguese, models called Zero[18] [26] were selected, which uses mBART as a base but focuses on back-translation to prevent translation into the wrong language, as with other multi-lingual models. This model was trained using News Commentary and Opus Books dataset.

We also tested the Lite model PT-EN-Translator[19] [27], created for translations only from English to Portuguese. It uses the T5 encoder-decoder (Text-to-Text Transfer Transformer) [28] and it was pre-trained with more than 5 million sentences obtained by ParaCrawl.

3.1.3 Fine-tuning

Fine-tuning was performed with the different datasets selected to compare the BLEU score and the accuracy obtained with WinoMT. We also created a small and balanced dataset, following a previous study [16], in order to mitigate the gender bias of the models.

We created an algorithm that was used to combine datasets and models. The algorithm checks if the model is mBART or if they have T5 as checkpoint. Each template has different prerequisites. If it is mBART, two variables are defined: the source language variable as `en-XX` and the target language variable as `pt-BR`. Otherwise, it is T5, and the prefix *translate English to Portuguese* is added.

[9] ⟨https://spacy.io/⟩
[10] ⟨https://github.com/facebookresearch/fairseq/blob/main/examples/roberta/README.md⟩
[11] ⟨https://github.com/cisnlp/simalign⟩
[12] ⟨https://huggingface.co/datasets/opus_books⟩
[13] ⟨https://huggingface.co/datasets/news_commentary⟩
[14] ⟨https://huggingface.co/datasets/ted_iwlst2013⟩
[15] ⟨https://huggingface.co/⟩
[16] ⟨https://www.statmt.org/wmt19/⟩
[17] ⟨https://huggingface.co/datasets/VanessaSchenkel/handmade-dataset⟩
[18] ⟨https://github.com/bzhangGo/zero⟩
[19] ⟨https://github.com/unicamp-dl/Lite-T5-Translation⟩

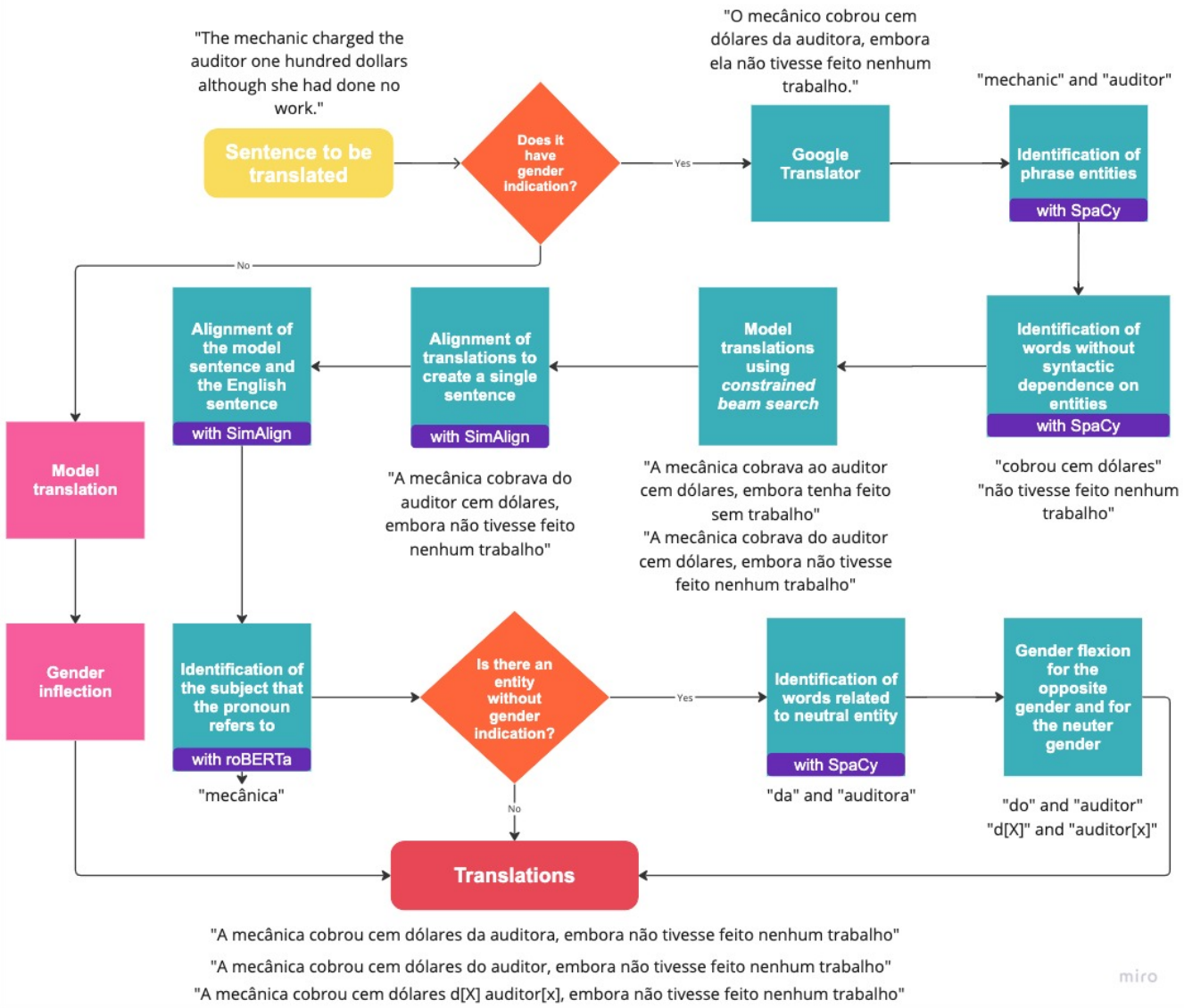

Figure 2. Proposed pipeline for translating a sentence from English to Portuguese with gender bias mitigation. The process begins with the input sentence and checks for gender indication. If gender is indicated, Google Translate performs the initial translation, and entities are identified using spaCy. The model translation utilizes constrained beam search, aligning translations to form a single sentence. Gender inflection is applied, followed by identification of subject pronouns using roBERTa. The final output includes gender-mitigated translations for feminine, masculine, and neutral forms, demonstrating the algorithm's capability to maintain context while ensuring gender neutrality.

Table 1. Comparative performance of different translation models across various datasets. The performance metrics include BLEU score and WinoMT Accuracy (ACC). The Lite model, fine-tuned with the handcrafted dataset, achieved the highest scores in both metrics, demonstrating its effectiveness in mitigating gender bias while maintaining translation quality.

		BLEU	WinoMT ACC
Zero	News commentary	33.74	54.35
	News commentary + handcrafted	33.70	54.30
	Opus books	32.94	52.34
Lite	Opus books	35.07	63.76
	TED	20.09	44.60
	Baseline	35.83	68.62
	Handcrafted	**36.01**	**72.99**
	News commentary	35.74	54.35

All selected models use the Transformer architecture. The dataset was split into 80% for training, 15% for testing, and 5% for validation. The training data is used to train the model. Validation data is used to compare different models and hyperparameters. The datasets were randomly split using the `train_test_split` method from the Datasets[20] library, with the value of `seed` in 42. The models were trained with a maximum size of 128 and the truncation option set to true.

We maintained the batch pattern size at 16, the learning rate at 2e-5, and regularization with a value of 0.01. Models return tensors in TensorFlow format. Finally, we used AdamWeightDecay optimizer from the Transformers[21] package, which allows regularization in gradients.

The experiments were performed on a MacBook Pro 2021 with an 8-Core Intel Core i9 processor and 16 GB of RAM. Tests were performed with different pre-trained models and datasets to create the study translation model. The results can be seen in Table 1.

The mBART model was discarded from the other tests due to its high computational demand. While others take an average of 15 minutes to translate 1,000 sentences, mBART took more than 4 hours. BLEU scores were obtained using the sacreBLEU[22] library with the TEDx Talks corpus called *mtedx_p_en*[23].

As demonstrated in [16], it is possible to obtain better results when using a smaller, balanced dataset than using a larger dataset. In addition, they have the advantage of requiring less computational time than retraining the model on a smaller dataset. Thus, the present study selected the Lite model with the fine-tuning done with the *handcrafted* dataset.

[20] ⟨https://pypi.org/project/datasets/⟩
[21] ⟨https://pypi.org/project/transformers/⟩
[22] ⟨https://github.com/mjpost/sacrebleu⟩
[23] ⟨http://openslr.org/100⟩

3.2 Algorithm

After creating the translation model, we develop an algorithm for using the model in real time, as demonstrated in the pseudocode 1. The algorithm receives as input the phrase to be translated, then it queries the translation from Google Translate, creates the *constraints* with linguistic analysis, does the translation with the model presented in previous section, aligns the translations generate by the model creating a final translation, does the gender inflections, and returns as output the gender-neutral translations. Each step is explained in detail in the following sections.

Algorithm 1 TranslateSentences

```
procedure TRANSLATESENTENCES(input)
    sentences ← splitInput(input)
    translations ← []
    for sentence in sentences do
        translationGoogle ← translateGoogle(sentence)
        entities ← extractEntities(sentence)
        genders ← extractGenders(sentence)
        if hasEntities(entities) then
            constraints ← identifyConstraints(sentence, entities)
            translationsConstraints ← []
            for constraint in constraints do
                translationConstraint ← translateConstraint(constraint)
                translationsConstraints.append(translationConstraint)
            end for
            finalTranslation ← alignTranslations(translationsConstraints)
            translations.append(finalTranslation)
        else
            translations.append(translationGoogle)
        end if
        if moreEntitiesThanGenders(entities, genders) then
            genderlessEntity ← identifyGenderlessEntity(entities)
            translationGenderless ← translateGenderless(genderlessEntity)
            translationOppositeGender ← applyOppositeGender(translationGenderless)
            translationNeutral ← applyNeutralGender(translationGenderless)
            translations.append(translationOppositeGender)
            translations.append(translationNeutral)
        end if
    end for
    return translations
end procedure
```

3.2.1 Linguistic analysis of the input

The first step of the algorithm consists of receiving the sentence and analyzing whether there is a gender indication in the sentence in English. This information is used so that the algorithm chooses one of its different paths. If the sentence does not have any gender indication, the output to the user must have the three possible translations: feminine, masculine and neutral. If the sentence has an indication of gender, it is necessary to identify whether there is more than one entity and to whom the indicated gender refers. If there is more than one entity, the three translations will also be generated, but the entity with the gender indication will only have its translation for one gender and the neutral translation. For example, Figure 3 represents an input phrase taken from the WinoMT challenge set.

When this sentence is received as input, spaCy's linguistic analysis is used to identify if there are pronouns. Morphological analysis of the same library is also used to verify whether there is a gender. In the example, it is identified that there is the pronoun *she*, represented in pink in the image.

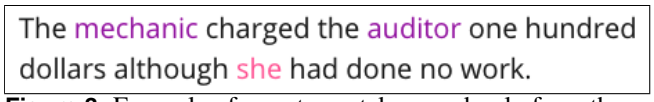

Figure 3. Example of a sentence taken randomly from the WinoMT challenge set to be translated into Portuguese. The words in purple represent the entities, and in pink, the indication of gender.

Through the tags it is also possible to identify the entities of the sentence. In this case, *mechanic* and *auditor* are identified, represented in purple.

3.2.2 Constrained Beam Search

The Constrained Beam Search technique is an approach used in machine translation to improve the quality of translations generated by language models, such as neural translation models. The goal of Constrained Beam Search is to incorporate additional constraints during the translation generation process to ensure that outputs meet specific criteria.

Typically, during machine translation, a language model generates several possible translations for a source sentence and ranks them based on their probability. The final outcome is chosen based on these probabilities. However, our objective is to maintain the structure of the commercial model, thus maintaining its translation quality. To achieve this, the study created an algorithm that uses spiCy's linguistic analysis to identify which words are not linked to any gender indication. With this technique, we inform our translation model which specific words it needs to generate.

For this step, first the translation is generated in Google Translate, represented by Figure 4. The translation *O mecânico cobrou cem dólares do auditor, embora ela não tivesse feito nenhum trabalho* exhibits gender bias by assigning masculine gender to *mechanic* and *auditor*, and feminine gender to the pronoun *ela*.

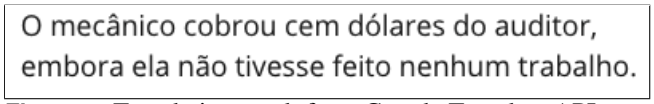

Figure 4. Translation result from Google Translate API highlights the limitations of current translation systems in handling gender-neutral language and underscores the necessity for our proposed gender-neutral translation approach.

We used the SimAlign library to know the corresponding translation of the entities *mechanic* and *auditor* in the commercial model, while maintaining the context. This tool does word-level alignment. Therefore, it is possible to verify that the translations, in this case, are *mecânico* and *auditor* as represented in Figure 5.

The next step is to generate the *constraints*, which will be sent to the model. The intention is to use the structure of the commercial model, but the genre of the translation model created by the present study. The final part is the post-processing based on the translation obtained by the model.

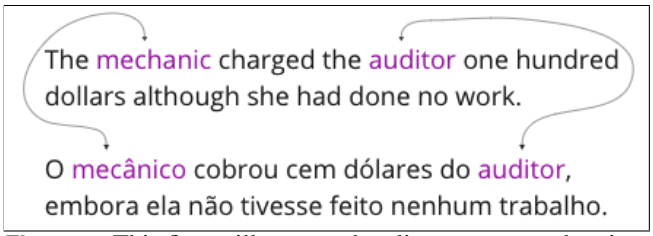

Figure 5. This figure illustrates the alignment step, showing how the entities in the English sentence are translated into Portuguese. The words *mechanic* and *auditor*, highlighted in purple, represent the entities being aligned. The arrows indicate the alignments between the source and target languages, highlighting the direct translations of the entities.

For this, a syntactic dependency analysis is performed with spaCy. All sets of words that have no syntactic dependency on the entities are added to a list of constraints.

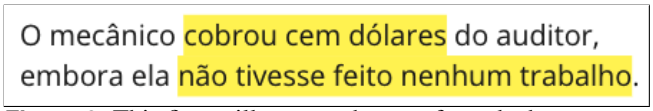

Figure 6. This figure illustrates the set of words that constitute the list of constraints in the translation process. The sentence has the constrained words highlighted in yellow. These constraints are used to ensure that certain parts of the translation remain consistent and accurate according to the model's requirements.

In Figure 6, the set of words identified as *constraints* are in yellow. They are identified *charged one hundred dollars* and *had not done any work*. The algorithm identifies constraints as words that are not linked to the entity that has an indication of gender. This is done with the spaCy library and its language dependency analysis tool.

3.2.3 Alignment

At each iteration of the *constraints* list, the model generates a different translation. So if *n constraints* are generated by the algorithm, *n* translations will be generated by the model.

Each constraint generates a different translation. To reach the final translation made by the model, it is necessary to iterate each of the translations. With the SimAlign tool, each translation is compared to the next. With alignment, the algorithm maintains the constraint and gender of the entity. The algorithm selects what should be kept and generates a translation. This translation is aligned and compared with the next translation, until reaching the end of the list of translations generated by the model. After iteration, the final translation will be formed. This way, it is possible to maintain the genre indicated by the model and the constraints identified previously, maintaining the quality of the commercial translator.

The words that are the translation of the model entity, that are a syntactic dependency of the same or that are part of the *constraints* are concatenated. The final sentence is the result of comparing the alignments.

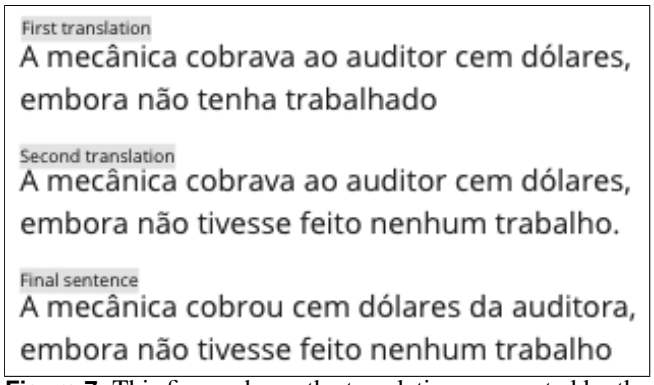

Figure 7. This figure shows the translations generated by the proposed model and algorithm. The process starts with the first translation, followed by the second translation and finally, the aligned and refined translation. These translations demonstrate the iterative refinement process to achieve a gender-neutral and accurate translation.

Thus, it is possible to maintain the genre generated by the model developed by this article and the structure generated by the commercial model, as shown in Figure 7. For example, when receiving the phrases *A mecânica cobrava ao auditor cem dólares, embora não tenha trabalhado* and *A mecânica cobrava do auditor cem dólares, embora não tivesse feito nenhum trabalho*, the first part is selected from the first sentence and the second from the second, as they are part of the *constraints* list. This is how the final translation of the model is generated *The mechanic charged the auditor a hundred dollars, even though she hadn't done any work*.

We obtained more than one translation because the number of sentences produced by the model will be the same number of constraints identified by the algorithm. Therefore, by carrying out the alignment step and keeping each part of the sentence that was identified as a constraint, we were able to generate the final translation.

3.2.4 Entity identification

The next step is to identify the entities of the text, so that we know which words are connected to that entity and whether there are gender indicators. In other words, through this identification we will be able to use the dictionary to generate a neutral translation for both the entity and the words that reference it.

In case there are more entities than gender indicators, we chose to use roBERTa. This library allows us to know which entity the indicator refers to. We need to use it in the case of having more than one entity, to know which entity should be kept neutral and which should have its gender translated. Considering the examples of previous sections, this step receives as input the sentence *The mechanic charged the auditor one hundred dollars although she [she] had done no work* and returns as output the sentence *The mechanic*. After identifying the main entity, alignment is performed again between the translation of the model and the sentence in English so that the translation of the main entity is identified while maintaining the context, as represented in Figure 8.

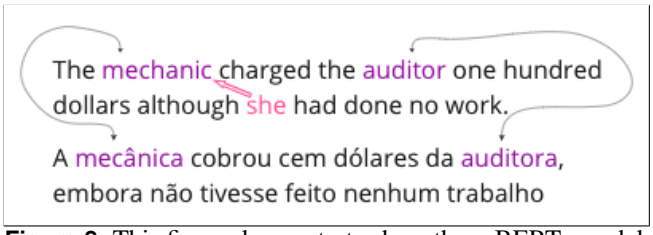

Figure 8. This figure demonstrates how the roBERTa model identifies the entity that the pronoun refers to in the sentence. In the sentence the entities *mechanic* and *auditor* are highlighted in purple, while the pronoun *she* is highlighted in pink. The arrow indicates that the pronoun *she* refers to *mechanic* allowing for accurate gender inflection in the translation.

As the roBERTa model returns that the entity *she* refers to is *The mechanic*, with the alignment, it is possible to identify that the translation of *The mechanic* is *A mecânica*. This allows the context of the sentence to be maintained, which makes it possible to correctly identify the gender. With the linguistic analysis performed by spaCy, it is identified that *the auditor* is also an entity. Therefore, it is possible to verify through the alignment that the translation should be *auditora*.

If other entities are not related and do not get gender indication, the syntactic dependence between the secondary entities and other tokens are analyzed. It is considered that these words do not have a defined gender, so they must go through the process of gender inflection. In the example, after identifying *auditor* as an entity that must be neutral, a list of tokens is created with *auditor* as their syntactic dependency. Then we created the list for the gender inflection step, which is composed of *da* and *auditora*.

3.2.5 Gender inflection

For gender inflection, a dataset was created from WikiMedia using the tool WiktExtract[24]. The selected parameters were: language in Portuguese, with inflections and etymologies. So the result was a text file with 366,447 words and their information. To facilitate handling, objects were created for each word, with their roots as a key and their inflections with gender, plural and singular version. Thus, 53,138 objects were generated, which were saved in a CSV file and made available[25]. Finally, objects with no gender inflection and only superlatives and comparatives were removed.

Each word identified as genderless is looked up in the inflection list. The algorithm maintains the constraints and performs a query on the remaining words in the dictionary created in the present study. If it is identified that the word is in the dictionary, it means that it has a gender variation. With the spaCy tool, its gender is then identified and whether it is plural or singular. As the dictionary, in addition to variations,

[24] ⟨https://github.com/tatuylonen/wiktextact⟩
[25] ⟨https://huggingface.co/datasets/VanessaSchenkel/pt-inflections⟩

has plural, singular, feminine and masculine labels, the grammatical number is maintained and the word of the opposite gender is returned. Thus, new sentences are generated with words that do not have gender in the source sentence, while maintaining the gender generation made by the model.

We only make neutral language inclusion if both results have a gender indication in the source sentence, and in those that do not, the suggestion made by [29] is to replace the ending *o* and *a* of the words by another letter. Therefore, words that are previously identified as neutral in the English sentence have their gender inflections identified as masculine and feminine. One of these sentences is duplicated, and the words with gender among those that should be neutral are identified. The end of these words is replaced with `[X]`.

In the example used to illustrate this flow, after identifying *da* and *auditora* as words that must undergo gender inflection, their morphology is identified with the spaCy tool. It is identified that the words are singular and feminine. Then the algorithm searches the list of inflections for their matches with the masculine and singular tags. After this identification, the phrase *A mecânica cobrou cem dólares do auditor, embora não tivesse feito nenhum trabalho*. The last step is to replace the ending of these words with `[x]`, which produces the neutral sentence *A mecânica cobrou cem dólares d[x] auditor[x], embora não tivesse feito nenhum trabalho*.

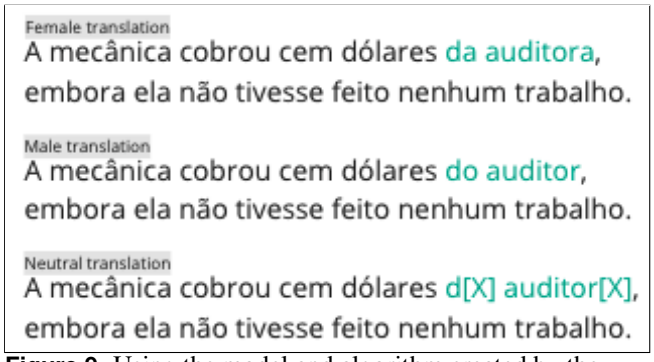

Figure 9. Using the model and algorithm created by the study, it is possible to keep the entity translation in the correct gender and translate the spare entities to feminine, masculine and neutral gender as highlighted in green.

In Figure 9, the entity without gender indication in English is represented in green. Translations are made for feminine, masculine and neutral gender.

3.2.6 Genderless sentence

If the phrase does not have a gender indication in English, first the entity of the phrase is identified. For example, in the sentence *I'm tired and I like to read a lot*, as shown in Figure 10. As seen, spaCy identifies that there is only one entity and no indication of gender. Therefore, the translation is carried out with Google Translate, the *constraints* are created to maintain the structure of the sentence, and the translation is carried out using the model. The translation is aligned until the final translation is obtained. After this step, gender inflection is performed.

Figure 10. Example sentence to be translated that is genderless in English but can be translated to feminine or masculine in Portuguese.

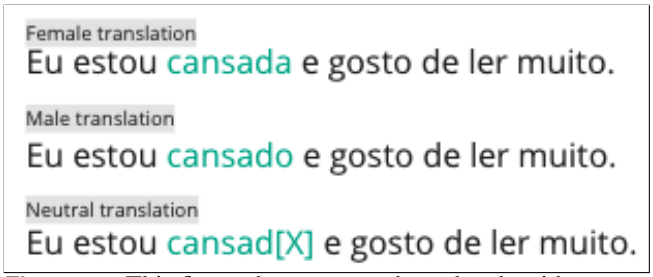

Figure 11. This figure demonstrates how the algorithm identifies gendered words in a sentence, consults a custom dictionary, and generates new sentences in feminine, masculine and gender-neutral forms.

We show the final result in Figure 11, as we can see, only the word *cansado* highlighted in green is inflected for feminine, masculine, and neutral gender.

3.2.7 Gendered sentence for all entities

In cases of sentences with gender indications for all entities, after spaCy's linguistic analysis, the sentence is translated by Google Translate and the *constraints* are created. The template creates the translation with the *constraints*, maintaining the sentence structure. Figure 12 shows an example sentence where each entity has a clear gender indication. The sentence *The doctor finished her work* uses the gender-specific pronoun *her* to refer to *doctor* which is highlighted in purple. This example is used to demonstrate the challenge of maintaining gender consistency in translations.

Figure 12. This figure presents an example of a sentence where all entities have a gender indication. In the sentence the entity *doctor* is highlighted in purple to indicate its gender-specific reference to *her*.

In this scenario, the step of indicating the entity using roBERTa is not carried out. Furthermore, the sentence is only translated into the gender indicated by the sentence in English and into neutral, as shown in Figure 13.

If there are more than one entity and more than one gender indication, but the same amount, the sentences are divided according to the syntactic dependency of each entity. After the list of phrases is created, they go through the same steps mentioned here.

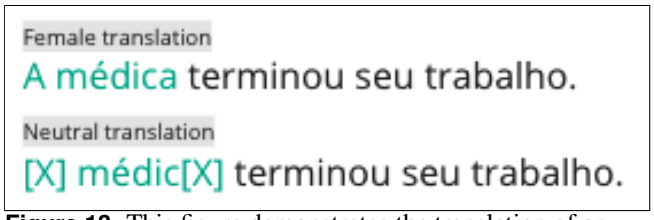

Figure 13. This figure demonstrates the translation of an entity into the gender indicated by the pronoun and into a neutral form. The first sentence shows the feminine translation, while the represents the neutral translation.

3.3 Evaluation metrics

WinoMT[26] [30] is a *challenge set* used to evaluate gender bias in translation models. It allows automatic bias assessment for translation from English to Arabic, Spanish, French, Hebrew, Italian, Polish, Russian, Czech, and Ukrainian.

It is composed of 3888 gender phrases from Winogender [31] and WinoBias [11]. Each sentence is composed of a primary entity that is co-referring to a pronoun and a secondary entity. For example:

```
The developer argued with the designer
  because she did not like the design.
```

In the example, the primary entity is *the developer* and is co-referent to the pronoun *she*. The translation, therefore, must be for the feminine. The secondary entity is *the designer* and must be gendered according to the grammar rules of the language for an ambiguous translation.

In the WinoMT assessment, grammatical gender is extracted from the primary entity of each translation hypothesis through automatic word alignment followed by morphological analysis. WinoMT then compares the translated primary entity with the gold gender, in order to verify that the translation is gender correct.

The *challenge set* has approximately the same number of sentences marked as masculine and feminine, and a few sentences with the neutral label, as can be seen in Table 2.

Table 2. Gender division on WinoMT challenge set

	Complete	**Pro-stereotype**	**Anti-stereotype**
Male	1826	792	794
Female	1822	792	790
Neutral	240	0	0

WinoMT offers the following metrics:

- **Accuracy:** percentage of hypotheses with the primary entity having the correct gender. The desire is to have the highest possible value in this metric.
- **△G:** difference in performance score (F1 score) between the set of sentences with male entities and female entities. △G = 0 means that the system makes the same number of errors between female and male professions. Therefore, the closer to the value zero, the better the translation model.
- **△S:** primary entity gender hit difference between pro-stereotyped and anti-stereotyped phrase sets. The closer to zero, the smaller the indication of bias.

There are also two other sets of phrases: pro-stereotypical and anti-stereotypical. They are formed using labor statistics from the United States [11]. For example, the pro-stereotyped set contains doctors and nurses, while the anti-stereotyped set contains female doctors and nurses. The objective of the models is to have high accuracy and scores △G and △S close to zero. A high score on △G means that the model translates male entities better, whereas a high △S score means that the model stereotypes male and female entities. Negative values are also undesirable, as they indicate a bias towards female or anti-stereotypic translation.

In addition, models need to be able to reduce gender bias without reducing translation performance. Commonly, the BLEU *(Bilingual Evaluation Understudy)* [32] score is also used, which is an algorithm for evaluating the quality of text between one language and another. The higher the score, the higher quality the translation.

3.3.1 WinoMT Extension for Portuguese

WinoMT, despite being widely used by the academic community to measure gender bias in translation models, did not have support for Portuguese. As it is *open-source*, we have modified it to measure translation models from English to Portuguese[27].

For this purpose, its packages were first updated. We also adjusted the spaCy Version from 2.3v to 3.4v, as the previous version did not support Portuguese. Furthermore, the *get_gender* method of the *util.py*[28] file was modified because in the new version of the package there is a new way to use the Morphologizer class. For the same reason, we modified *spaCyPredictor* class to update the language loading too.

We created *evaluate_bias_complete* method in order to generate *logs* with more information. In addition, Portuguese translations were added for the AWS, Bing, Google Translate, and Systran models.

In the latest version of the main *branch* there are some reported problems that prevent the tool from working. To solve these problems, the files containing the *challenge set* were modified for the previous version, more specifically modifying the line with the phrase *The baker yelled at the mover because he broke the stove* .

The *eval_vs_human.sh* file was modified, removing the iteration in the languages because it was broken. Finally, the names of the files in the *human_annotations* folder were modified to follow a pattern and the file *italian.csv* was removed because it had no annotations.

[26] ⟨https://github.com/gabrielStanovsky/mt_gender⟩

[27] ⟨https://github.com/ramos-ai/winopt⟩

[28] ⟨https://github.com/ramos-ai/winopt/blob/main/src/languages/util.py⟩

3.4 Reproducibility

To ensure the reproducibility of our study, we provide a public source code with instructions for its installation and execution. You must have the following tools installed: Python 3.8, spaCy, roBERTa and SimAlign. You must have a machine with at least 16GB of RAM and a GPU is recommended to train and run the models efficiently.

3.4.1 Data Used for Training

The data used to train our model are detailed in Section 3.1.1. We utilized publicly available datasets from Hugging Face:

- Opus Books: Contains pairs of sentences from books with translations from English to Portuguese.
- News Commentary: A parallel corpus obtained from WMT19, supporting multiple languages.
- TED 2013: Contains English-Portuguese sentence pairs taken from TED Talks in 2013.

For the fine-tuning step, we created a balanced dataset, available on Hugging Face:

- Handcrafted Dataset: Contains sentences specifically crafted for this study to ensure balanced gender representation.

3.4.2 Evaluation and Metrics

To evaluate our model, we used the WinoMT challenge set, which has been extended to support Portuguese. The code and evaluation data are available in our repository:

- WinoMT Extension

WinoMT provides several metrics for assessing gender bias in translation models, including accuracy, $\triangle$G, and $\triangle$S scores. These metrics help in evaluating the model's performance and its bias tendencies.

3.4.3 Source Code and Scripts

All the source code and scripts necessary to replicate the experiments are available on GitHub:

- unbIAs Repository

3.4.4 Detailed Steps

To replicate our experiments, follow these detailed steps for installing dependencies, setting up the environment, and running the provided scripts:

Cloning the Repository

Clone the repository from GitHub:

```
git clone https://github.com/ramos-ai/unbIAs.git
cd unbIAs
```

Setting Up the Environment

Ensure you have Python 3.8 installed. You can create a virtual environment to manage dependencies:

```
python3.8 -m venv env
source env/bin/activate
```

Installing Python Dependencies

Navigate to the `api` directory and install the required Python packages:

```
cd api
pip install -r requirements.txt
```

Running the API Server

Start the Flask server to provide the backend for the application:

```
flask run -p 8000
```

Setting Up the Front-End

Navigate to the `APP` directory and install Node.js dependencies:

```
cd ../APP
npm install
npm start
```

Training the Model You can follow the instructions in the Methodology section to train the model using the provided datasets. Ensure you have the necessary datasets available and correctly configured paths.

Evaluating the Model

Use the WinoMT extended toolkit to evaluate the trained model. The toolkit has been modified to support Portuguese and can be used as follows:

```
cd ../winopt
python evaluate.py --model_path /model --data_path /data
```

Generating Results

The evaluation script will generate logs and results, which can be analyzed to understand the model's performance and bias mitigation effectiveness.

By following these instructions, other researchers can reproduce our results and validate the methodologies used in this study.

4. Results and discussion

The present study fine-tuned a translation model with the construction of a *handcrafted* dataset balanced between female and male entities. The intent of the translation model is to have good gender accuracy to perform Google Translate mitigation. To maintain the quality of the commercial system, the constrained beam search technique was implemented. In addition to implementing the technique, we created an algorithm that uses spaCy to perform the linguistic analysis of the sentence received by the user to identify entities and pronouns. If there is an entity and at least one pronoun, we use roBERTa to identify the entity to which the pronoun refers. In addition to correct translation, the tool also offers neutral translation.

For a better comparison between translation systems, the commercial systems AWS, Bing, Google and Systran were

Table 3. The table compares the result of the BLEU and WinoMT scores in different commercial and study-created translators. The baseline model refers to the model choose by the present study. The fine-tuned represents the same model after being fine-tuned with the handcrafted dataset. Last, unbIAs represents the baseline model after being fine-tuned and using our algorithm to implement the constrained beam search technique. Bold scores represent the best results.

	BLEU	**Acc**	**△G**	**△S**
AWS	55.90	65.69	8.8	21.57
Bing	53.24	57.69	21.94	21.29
Google	56.90	68.75	6.23	16.09
Systran	**70.91**	60.19	18.08	18.15
Baseline	35.86	68.42	7.85	**-7.64**
Fine-tuned	36.01	**72.99**	**3.94**	-7.77
unbIAs	48.39	70.09	5.25	9.09

selected. Furthermore, the baseline model called Lite and the Lite model with the fine-tune made with the handcrafted dataset were scored. Finally, the system called unbIAs was scored, which uses the Lite model with fine-tuning and the algorithm created by the study. The complete results can be seen in Table 3.

The proposed model presents, in relation to Google Translate, an improvement in gender accuracy, an increase of 15.7% in the score in the correct answers between male and female entities and an improvement of 43% in relation to stereotyped translations.

The model's gender accuracy, as measured by WinoMT, was higher than all commercial models and translators created. In relation to Google Translate, it obtained an increase from 68.75% to 70.09% in the correctness of the sentence's gender.

The △G score that measures the difference in hitting male entities and female entities. A high positive score indicates that the model translates male entities better. Bing translator has the highest △G score with 21.94 points. The smallest is the fine-tuned model with the *handcrafted* dataset, an improvement from 6.23 to 3.94 over Google Translate.

The △S score represents the difference in accuracy between pro-stereotyped and anti-stereotyped sentences. A high positive score indicates that the model stereotypes female and male entities. The translator with the highest score is AWS, with 21.57 points. Baseline and fine-tuned translators obtained negative score, which represents a tendency towards anti-stereotyped translations. However, the values are closer to the desirable value, which is 0. The model that uses the constrained beam search technique obtained 9.09 points, while Google Translate obtained 16.09 points.

The model encounters difficulties translating certain professions, such as *mover*. For example, the sentence *The mover said thank you to the housekeeper because she is grateful* is incorrectly translated as *A mudança agradeceu à governanta porque ela está agradecida*. In cases where the gender is male, like *The mover said thank you to the housekeeper because he cleaned the truck*, the translation becomes *O homem mudança agradeceu domicílio, agradeceu à governanta porque limpou o caminhão*, which is clearly erroneous.

Additionally, our model struggles with phrases beginning with *"Did"*. For more complex sentences containing multiple entities and over 100 characters, the translation process can take approximately 25 seconds.

As pointed out by the cited study [14], translation models tend to translate all entities into the same genre. Both the model created by the present study and Google Translate have this tendency, but in our model it is translated into the correct genre. For instance, in the sentence *The physician prescribed the drugs to the designer because she thought the disease could be cured* our model translates it as *A médica receitou os remédios para a designer, porque ela achava que a doença poderia ser curada* Here, despite the second entity, *designer*, lacking explicit gender indication, it is translated into the feminine form, consistent with the gender indicated by the pronoun *she*. Additional examples illustrating this phenomenon are available in our repository[29].

The present study includes the generation of neutral translations for both genderless and gendered sentences, but not for all entities. Currently, Google Translate provides both male and female translations for one-word translation only. The model and algorithm created by the present study do this for all sentences.

The techniques increased the BLEU score from 35.86 in the initial model to 48.39 with the implementation of constrained beam search. Although the fine-tuned model had better results in the WinoMT score, the constrained beam search model had a 34% better BLEU score, indicating overall better translations. Translations above 40 points are considered high quality[30].

The objective of the study is not only to mitigate gender bias, but to have a quality translator that also provides the inclusion of non-binary people. For this reason, unbIAs can be considered better since in addition to having a considerable improvement in relation to Google Translate's gender bias, it has a translation considered to be of high quality.

4.1 unbIAs

For the implementation of the created algorithm, an interface was developed called unbIAs. The code is publicly available [31]. The tool intelligently identifies words without gender indications and translates these into masculine, feminine, and neutral forms. This selective approach ensures that only gender-neutral words receive the multiple translation forms, maintaining the intended meaning while offering gender-specific and gender-neutral options. This feature is crucial for promoting inclusivity, allowing users to choose translations that align with their gender preferences and supporting non-binary language use.

[29] ⟨https://github.com/ramos-ai/unbIAs/blob/main/api/model/src/data_score/wino/model-en-pt-wino.txt⟩

[30] ⟨https://cloud.google.com/translate/automl/docs/evaluate#bleu⟩

[31] ⟨https://github.com/ramos-ai/unbIAs⟩

Figure 14. This figure shows the user interface of the unbIAs translation tool, where users can input a sentence and receive translations in feminine, masculine, and neutral forms keeping the correct gender for the entitity with the gender indication. The example sentence *The lawyer yelled at the hairdresser because she was mad* is translated to *A advogada gritou com a cabeleireira porque ela estava brava* (feminine), *A advogada gritou com o cabeleireiro porque ela estava brava* (masculine), and *A advogada gritou com u cabeleireiru porque ela estava brava* (neutral).

The tool allows users to input sentences and choose how neutral translations should end, as shown in Figure 14. The example sentence *The lawyer yelled at the hairdresser because she was mad* is translated into feminine, masculine, and neutral forms, demonstrating the tool's capability to handle gender-specific and gender-neutral translations.

Figure 15 illustrates the dropdown menu in the unbIAs translation tool where users can select the preferred ending for neutral translations. This feature allows for customization of gender-neutral language, providing options such as `[X]`, `@`, `e`, and `u` to accommodate various preferences for inclusive language. If the user does not select an option, the default is `u`. The ability to choose the ending for neutral translations enhances user control and inclusivity, ensuring that the translations are respectful and accurate according to the user's needs.

In the algorithm, the neutral translation was chosen to have the `[X]` inflection, as it would be less frequent as user input as a sentence to be translated. The flexion `u` was used as a standard inflection in unbIAs, following the Manual for the use of neutral language in Portuguese [33] since `X` and `@` make access difficult for blind, deaf, people with disabilities, autism spectrum disorder (ASD) and dyslexia.

The interface adapts itself depending on the translations. If there is a gender indicator for all entities, two translations are returned: one with gender mitigation and the neutral option. Figure 16 illustrates how the unbIAs tool provides two translation outputs for a sentence where all entities have gender indications. The example sentence *The doctor finished her work* is translated into both a gender-specific form *'A médica terminou seu trabalho* and a gender-neutral form *[X] médic[X] terminou seu trabalho*. This demonstrates the tool's capability to offer both gender-specific and gender-neutral translations based on user preferences.

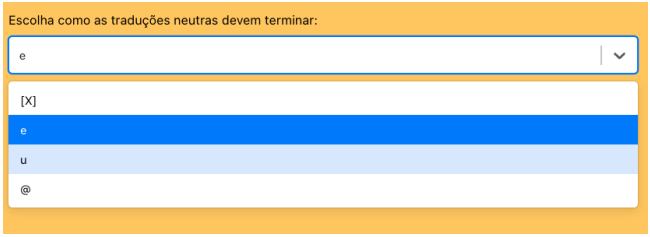

Figure 15. This figure shows the selection options for neutral translation endings in the unbIAs tool. Users can choose how gender-neutral translations should end, with options including *[X]*, *e*, *u*, and *@*.

On the other hand, if there is no gender indication for all entities, that is, a neutral sentence in English, three translations are shown: with feminine, masculine, and neutral gender, as shown in Figure 17. The example sentence *I'm tired but I like to read a lot* is translated into masculine *Eu estou cansado, mas gosto muito de ler*, feminine *Eu estou cansada, mas gosto muito de ler*, and gender-neutral *Eu estou cansadu, mas gosto muito de ler* forms. This demonstrates the tool's ability to offer multiple gender-specific and gender-neutral translations based on user preferences.

By implementing these features, the unbIAs tool ensures that translations are not only accurate but also inclusive, reflecting the diverse gender identities of users and promoting the use of non-binary language.

5. Conclusion and future work

The bias in artificial intelligence can be more damaging to society than human bias, as it manages to reach a greater number of people at the same time. This study seeks to provide an overview of gender bias in translation models,= in addition to

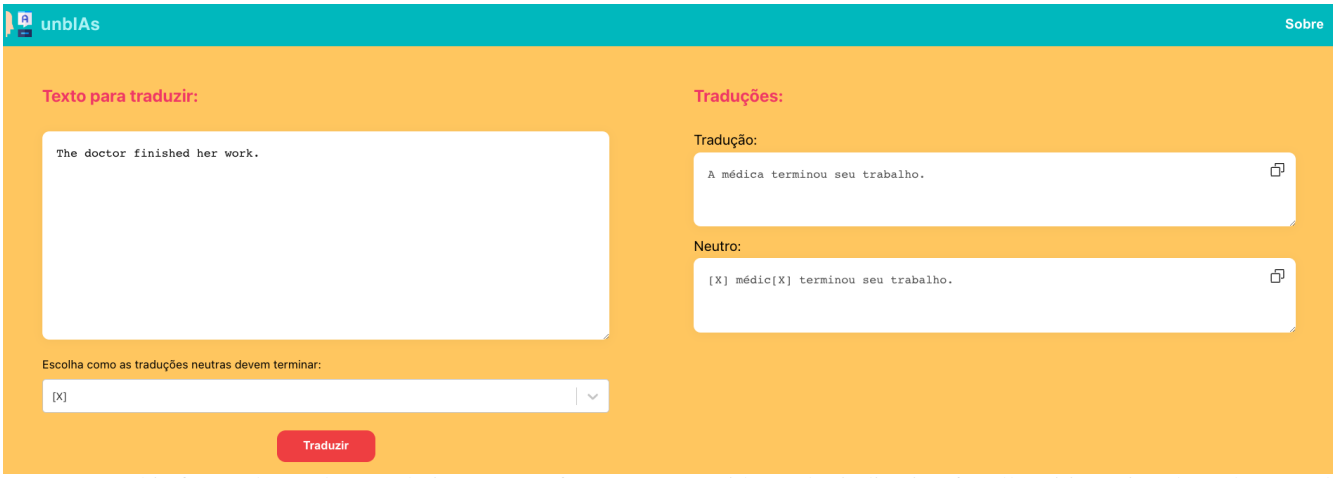

Figure 16. This figure shows the translation outputs for a sentence with gender indication for all entities using the unbIAs tool. The sentence *The doctor finished her work* is translated into *A médica terminou seu trabalho* (feminine) and *[X] médic[X] terminou seu trabalho* (neutral).

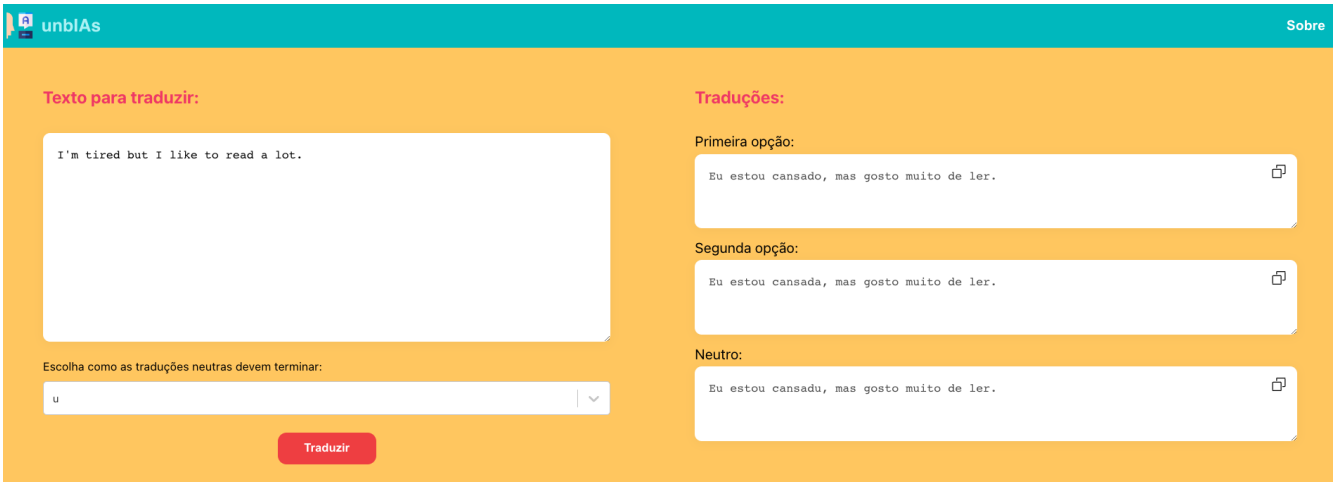

Figure 17. This figure shows the translation outputs for a gender-neutral sentence using the unbIAs tool. The sentence *I'm tired but I like to read a lot* is translated into *Eu estou cansado, mas gosto muito de ler* (masculine), *Eu estou cansada, mas gosto muito de ler* (feminine), and *Eu estou cansadu, mas gosto muito de ler* (neutral).

mitigating bias in Google Translate.

This article proposes the creation of a system that receives a sentence in English from the user, identifies if there is a gender indication, and if there are entities in the sentence. The sentence is translated by Google Translate so that the *constraints* can be identified. Each constraint is passed to the model using the constrained beam search technique. Thus, it is possible to maintain the gender accuracy of the model produced by the study, but also the translation quality of the commercial model. The phrases generated by the model are aligned until the final translation is generated.

The introduction of unbIAs, a translator that effectively mitigates gender bias in translations produced by popular tools like Google Translate, marks a significant milestone not only for the machine learning community but also for society at large. Gender bias in machine translation has been a pervasive issue, perpetuating stereotypes and contributing to unequal representations of genders. UnbIAs addresses this problem by providing a solution that promotes gender-neutral translations, fostering inclusivity and fairness in language processing. In a world where digital communication plays a central role, this innovation is crucial for ensuring that technology serves as a vehicle for positive change, rather than reinforcing harmful biases.

Machine translation plays a crucial role in global communication, influencing how people perceive and interact with different cultures and identities. When these tools perpetuate gender stereotypes they can reinforce existing prejudices and contribute to inequality.

We believe that tools that mitigate gender bias help reduce the perpetuation of gender stereotypes. For example, by ensuring that professions traditionally associated with a specific gender are translated neutrally or with feminine and masculine alternatives, the idea that certain jobs are inherently

masculine or feminine is avoided.

Another important achievement to highlight is the validation of non-binary people. The inclusion of gender-neutral translation options validates non-binary identities, recognizing and respecting gender diversity. Tools that offer neutral translations help create a more inclusive environment for all people by allowing their identities to be represented appropriately.

The creation of a dedicated dataset for fine-tuning and the development of a Portuguese extension for WinoMT not only contribute to the advancement of gender-neutral translation but also enrich the resources available for researchers exploring related topics. This work fosters collaboration and encourages further investigation into the nuances of bias in artificial intelligence systems. By actively engaging with and addressing the challenges of gender bias, the research community in machine learning continues to evolve towards more responsible and equitable AI, ensuring that technology benefits everyone in society, regardless of their gender.

The study faced computational limitations, which meant that fine-tuning was not possible for all the chosen datasets. Furthermore, all the chosen translation APIs were used in their free version, which could result in a worse translation. However, it is more similar to the translation that is received by the user.

In addition to creating the model and algorithm, WinoMT was also extended to Portuguese and a dictionary was created with more than 53,000 words in Portuguese with their gender and numerical inflections. While this study focuses exclusively on English-to-Portuguese translation, future work could expand the scope to include other languages that exhibit gendered nouns and pronouns more prominently. Such languages include Spanish, French, German, Italian, and Russian, which could benefit significantly from gender bias mitigation. Expanding to these languages would enhance the tool's applicability and effectiveness across a broader linguistic spectrum, allowing for a more comprehensive evaluation of the tool's performance in managing gender bias in various grammatical and cultural contexts.

It is important to emphasize that the study focused on gender bias. However, some racial and ethnic groups may be underrepresented in the training data used to build translation models, resulting in lower-quality translations for languages and dialects of these groups. This can further marginalize these groups and limit their visibility and adequate representation in translated content.

It is crucial to recognize that machine translation systems are subject to a wide range of other biases that also deserve attention and intervention. These biases include, but are not limited to, racial, ethnic, religious, and cultural biases. Future work may not only include mitigating these biases but also creating tools that help detect bias, such as WinoMT.

For the evaluation of the unbIAs tool to real-world scenarios, we conducted tests using 1,000 random sentences from the TEDx Talks corpus *mtedx_p_en*. This dataset was chosen to simulate realistic use cases and assess the tool's performance in a diverse range of contexts. However, future work could further validate the tool's effectiveness by deploying it on various online communication platforms and translation applications. This would provide a broader understanding of its performance and robustness in more dynamic and varied real-world environments.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (grants 313845/2023-9, 443184/2023-2, and 402679/2021-0), and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul - FAPERGS (grant 21/2551-0001944-1).

Author Contributions

Vanessa Schenkel: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation, Writing - Original Draft, Visualization. Blanda Mello: Methodology, Validation, Writing - Review and Editing. Sandro Rigo: Writing - Review and Editing, Funding acquisition. Gabriel Ramos: Conceptualization, Investigation, Writing - Review and Editing, Supervision, Funding acquisition.

References

[1] OLTEANU, A.; DIAZ, F.; KAZAI, G. When are search completion suggestions problematic? *Proceedings of the ACM on Human-Computer Interaction*, ACM New York, NY, USA, v. 4, n. CSCW2, p. 1–25, 2020.

[2] HUANG, S. et al. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer letters*, Elsevier, v. 471, p. 61–71, 2020.

[3] ZEISER, F. A. et al. Deepbatch: A hybrid deep learning model for interpretable diagnosis of breast cancer in whole-slide images. *Expert Systems with Applications*, Elsevier, v. 185, p. 115586, 2021.

[4] MEHRABI, N. et al. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 54, n. 6, p. 1–35, 2021.

[5] DASTIN, J. Amazon scraps secret ai recruiting tool that showed bias against women. In: *Ethics of Data and Analytics*. [S.l.]: Auerbach Publications, 2018. p. 296–299.

[6] SKEEM, J. L.; LOWENKAMP, C. T. Risk, race, & recidivism: Predictive bias and disparate impact.(2016). *Criminology*, v. 54, p. 680, 2016.

[7] METZ, R. Microsoft's neo-nazi sexbot was a great lesson for makers of ai assistants. *Artificial Intelligence*, 2018.

[8] HASSANI, B. K. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics*, Springer, v. 1, n. 3, p. 239–247, 2021.

[9] PRATES, M. O.; AVELAR, P. H.; LAMB, L. Assessing gender bias in machine translation–a case study with google translate. *National Communication Association (NCA)*, 2018.

[10] AI, H. *High-level expert group on artificial intelligence*. [S.l.]: European Commission. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019. 6 p.

[11] ZHAO, J. et al. Gender bias in coreference resolution: Evaluation and debiasing methods. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 15–20. Disponível em: ⟨https://aclanthology.org/N18-2003⟩.

[12] BOLUKBASI, T. et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, v. 29, 2016.

[13] VANMASSENHOVE, E.; HARDMEIER, C.; WAY, A. Getting gender right in neural machine translation. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2018. p. 3003–3008.

[14] SAUNDERS, D.; BYRNE, B. Reducing gender bias in neural machine translation as a domain adaptation problem. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2020. p. 7724–7736.

[15] SAUNDERS, D.; SALLIS, R.; BYRNE, B. First the worst: Finding better gender translations during beam search. In: *Findings of the Association for Computational Linguistics: ACL 2022*. [S.l.: s.n.], 2022. p. 3814–3823.

[16] SAUNDERS, D. *Domain adaptation for neural machine translation*. Tese (Doutorado) — University of Cambridge, 2021.

[17] SUN, T. et al. Mitigating gender bias in natural language processing: Literature review. *Association for Computational Linguistics (ACL 2019)*, 2019.

[18] BOLUKBASI, T. et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, v. 29, p. 4349–4357, 2016.

[19] ZHAO, J. et al. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Empirical Methods of Natural Language Processing*, 2017.

[20] WERLEN, L. M.; POPESCU-BELIS, A. Using coreference links to improve Spanish-to-English machine translation. In: *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*. Valencia, Spain: Association for Computational Linguistics, 2017. p. 30–40. Disponível em: ⟨https://aclanthology.org/W17-1505⟩.

[21] HOKAMP, C.; LIU, Q. Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*, 2017.

[22] CHOUSA, K.; MORISHITA, M. Input augmentation improves constrained beam search for neural machine translation: Ntt at wat 2021. In: *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. [S.l.: s.n.], 2021. p. 53–61.

[23] LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. 2019.

[24] SABET, M. J. et al. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. Online: Association for Computational Linguistics, 2020. p. 1627–1643. Disponível em: ⟨https://www.aclweb.org/anthology/2020.findings-emnlp.147⟩.

[25] LIU, Y. et al. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, v. 8, p. 726–742, 2020.

[26] ZHANG, B. et al. Simplifying neural machine translation with addition-subtraction twin-gated recurrent networks. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. p. 4273–4283. Disponível em: ⟨http://aclweb.org/anthology/D18-1459⟩.

[27] LOPES, A. et al. Lite training strategies for Portuguese-English and English-Portuguese translation. In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, 2020. p. 833–840. Disponível em: ⟨https://www.aclweb.org/anthology/2020.wmt-1.90⟩.

[28] RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, v. 21, n. 140, p. 1–67, 2020.

[29] AUXLAND, M. Para todes: A case study on portuguese and gender-neutrality. *Journal of Languages, Texts and Society*, v. 4, p. 1–23, 2020.

[30] STANOVSKY, G.; SMITH, N. A.; ZETTLEMOYER, L. Evaluating gender bias in machine translation. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2019. p. 1679–1684.

[31] RUDINGER, R. et al. Gender bias in coreference resolution. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. [S.l.: s.n.], 2018. p. 8–14.

[32] PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: . [S.l.: s.n.], 2002. p. 311–318.

[33] CAÊ, G. Manual para o uso da linguagem neutra em língua portuguesa. *https://drive. google. com/file/d/16BQ59w4ePbUqMAzrFwUiCsz3r9zJw9XL/view. Acesso em*, v. 25, n. 07, p. 2020, 2020.