Journal on Interactive Systems, 2025, 16:1, doi: 10.5753/jis.2025.5958

Analyzing Discourses in Portuguese Word Embeddings: A Case of Gender Bias Outside the English-Speaking World

Fernanda Tiemi de Souza Taso [**Federal University of Mato Grosso do Sul** | ***tiemi.taso@ufms.br***]
Valéria Quadros dos Reis [**Federal University of Mato Grosso do Sul, Leuphana University Lüneburg** | ***valeria.reis@ufms.br***]
Fábio Viduani Martinez [**Federal University of Mato Grosso do Sul** | ***fabio.martinez@ufms.br***]

✉ *Faculty of Computing, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil, ZIP Code 79070-900*

Received: 24 April 2025 • **Accepted:** 04 July 2025 • **Published:** 13 July 2025

Abstract: In this paper we meticulously examined a Word Embedding model in Portuguese, endeavoring to identify gender biases through diverse analytical perspectives, employing SC-WEAT and RIPA metrics that is widely used in the English realm. Our inquiry focused on three primary dimensions: (1) the frequency-based association of words with feminine and masculine terms; (2) the identification of disparities between grammatical classes pertaining to gender sets; and (3) the categorisation and grouping of feminine and masculine words, including their distinctive attributes. In regard to frequency groups, our investigation revealed a pervasive negative association of words with feminine terms in most subsets, indicative of a pronounced inclination of the model's vocabulary towards the masculine references. Notably, among the 100 most frequent words, 89 exhibited a stronger association with masculine terms. In the scrutiny of grammatical classes, our analysis demonstrated a predominant association of adjectives with feminine references, underscoring the imperative for supplementary description when referring to women. Furthermore, a conspicuous prevalence of participle verbs associated with feminine terms was observed, a phenomenon distinct from their male counterparts and one that requires further expert attention to be properly explained. The categorisation process underscored the existence of gender bias, as exemplified by the association of words with masculine terms within the domains of sport, finance, and science, while words related to feelings, home furniture, and entertainment were associated with feminine terms. These findings assume significance in fostering a discourse on gender analysis within non-English models, such as Portuguese models, thereby encouraging the Brazilian community to actively investigate biases in NLP models.

Keywords: Natural Language Processing, Computational Linguistics, Algorithmic Sexism, Ethics in AI, Non-English NLP

1 Introduction

Numerous quotidian activities are inherently interlaced with technologies employing Artificial Intelligence (AI) for the resolution of specific tasks, such as content recommendation algorithms, facial recognition, and customer service. The scrutiny of technological impact on societal dynamics assumes superlative importance for the identification and examination of ethical dilemmas arising from their interplay. Adverse repercussions have proliferated concomitant with the adoption of these tools within public security services across the United States, Europe, and more recently, in Brazil [Falcão, 2021], thereby exerting deleterious effects on specific demographic groups. Instances of the application of facial recognition in security cameras deployed in urban thoroughfares have been associated with the wrongful apprehension of innocent individuals from the black community [Werneck, 2019]. Reverberations on social media, precipitated by the discerned biases in the Twitter platform's image cropping algorithm that favours white countenances over black counterparts, instigated its subsequent modification [Yee *et al.*, 2021]. Diverse instances of algorithmic discrimination, including assistance services, are discernible within the existing scholarly literature [Silva, 2022].

Criticism directed towards technology monopolies and their algorithms, widely used by billions of individuals on a daily basis, contributes significantly to comprehending the potential ramifications they may transmit. Search tools, coupled with algorithms for website and image ranking exemplified by Google Search and Google Images, respectively, have been subject to accusations of algorithmic racism. This criticism emerged subsequent to numerous instances wherein the aforementioned tools were implicated in recommending explicit content when conducting searches containing the phrase "black girls." The company characterized this incident as an inadvertent lapse, subsequently rectifying the error. However, Google's response, marked by evasion, underscores a reluctance to acknowledge the imperative necessity of algorithmic adjustments to preclude discrimination against specific societal groups [Noble, 2018].

The consideration of algorithmic neutrality arises as a prospective remedy to mitigate human biases in decision-making processes wherein the cognitive processes of individuals are subject to undesirable influence. Nonetheless, computational models are not immune to biases, given that they are constructed and refined through human reasoning and decisions. These decisions may be individual, collective as determined by a team, or emanating from a higher authority.

Natural Language Processing (NLP) is concerned with the comprehension of extensive textual and speech data through computational systems. Within this domain, Suresh and Guttag [2021] expound upon various forms of biases that may manifest during data collection, model development, or model implementation. Historical bias becomes apparent when systems yield prejudicial and detrimental outcomes, notwithstanding the meticulousness of measurements and samples in data collection, thereby mirroring biases inherent in real-world data. Extensive examination has been carried out on Word Embeddings (WE) models, revealing their ability to encapsulate historical biases [Bolukbasi *et al.*, 2016; Caliskan *et al.*, 2017; Omrani Sabbaghi and Caliskan, 2022; Sogancioglu *et al.*, 2022]. WE, a widely employed NLP technique, involves the transformation of words from a corpus into vectors while preserving their semantic meanings.

Bolukbasi *et al.* [2016] were precursors to the proposal and scrutinization of a metric designed to quantify biases within Word Embeddings. Using models trained in news texts, the authors illuminate the pervasive presence of stereotypes, particularly in the context of professions, exemplified by the association "The man is to the computer programmer, as the woman is to the housewife." Subsequently, Caliskan *et al.* [2017, 2022] searched through the same thematic realm, focusing on the analysis of gender bias regarding frequency, syntax, and semantics within data sourced from the Web. The authors contend that the vocabulary employed in the scrutinized WE models exhibits a distinct inclination towards masculine terms, with verbs displaying a stronger association with the masculine gender and adjectives with the feminine gender. Furthermore, semantic clusters are observed, where word groups such as engineering and sports align closely with the masculine gender, whereas terms pertaining to appearance and cooking are more prominently associated with the feminine gender.

The exploration of gender bias within Word Embeddings has been extensively researched and documented, with a predominant focus on the English language [Sun *et al.*, 2019; Garg *et al.*, 2018; Park *et al.*, 2018; Chaloner and Maldonado, 2019]. However, this attention is not reflected in the research conducted on languages spoken in the Global South, including African, Indian, and Latin languages, where the amount of analysis remains notably insignificant.

Nevertheless, it is noteworthy to highlight recent research endeavours that aim to address sexism within Word Embedding models across various languages, encompassing Chinese [Chen *et al.*, 2022], Spanish [Torres Berrú *et al.*, 2023], German [Wagner and Zarrieß, 2022] and African languages [Wairagala *et al.*, 2022].

Within the realm of gender bias analysis in the Portuguese language, few works exist. Three of the most notable studies use different sorts of techniques to evidence gender bias in texts or embeddings composed of newspaper reports, literary books, and Wikipedia pages [Taso *et al.*, 2023a; Freitas and Martins, 2023; Salles and Pappa, 2021]. Nevertheless, none of them reveal gender discrepancies in WE according to a systematic frequency analysis of the association of words, such as the investigation conducted by Caliskan *et al.* [2022].

This research focuses on analyzing gender bias in Portuguese WE using a metric based on the Word Embedding Association Test (WEAT) [Caliskan *et al.*, 2017]. Due to the absence of consensus on bias identification measures, we opted to utilize WEAT due to its extensive adoption in the relevant literature, which facilitates comparisons and reproducibility of results. To validate our findings, we compared part of the results obtained using WEAT with those obtained using RIPA, another bias identification metric. Our experiments follow the methodology described by Caliskan *et al.* [2022] and produce information in terms of: (1) frequency-based correlation of words with feminine and masculine terms; (2) identification of disparities among grammatical categories related to gender constructs; and (3) classification and clustering of feminine and masculine words, along with their unique characteristics.

Our results are in agreement with the one obtained by Caliskan *et al.* [2022]. They also complement previous studies that identified differences in gender treatment in Brazil [Taso *et al.*, 2023a,b; Freitas and Martins, 2023; Salles and Pappa, 2021]. In terms of frequency groups, many of the most frequent words are associated with men relative to women. Notably, among the 100 most frequent words, 89 exhibited a stronger association with masculine terms. Focusing on the parts-of-speech used for men versus women, male-associated words are more likely to be verbs in the finite, gerund, and infinitive forms. In contrast, female-associated words are more likely to be adjectives and verbs in the participle form.

Finally, the categorisation process according to semantic content emphasized the domains of words according to gender. The masculine terms were associated with the domains of sport, finance, and science, while words related to feelings, home furniture, and entertainment were associated with the feminine terms.

The substantial disparities in gender inequalities noted herein are evident across multiple facets of life, including education, employment, healthcare, political participation, and social norms. Data from the Brazilian Institute of Geography and Statistics (IBGE) indicate that, on average, women dedicate three more hours per week than men to a combination of paid work, domestic chores, and caregiving responsibilities. Despite having a higher level of education, women earn an average of 76.5% of men's income [IBGE, 2024]. This discrepancy highlights a paradox in which women invest more in education and labour, yet experience lower financial returns compared to their male counterparts. Besides that, recent years in Brazil have witnessed concerning statistics, including a surge in cases of feminicide, gender-based violence, discrimination, and prejudice, as reported by the Violence Monitor LESFEM [2024].

Considering the arguments presented, it is essential to highlight the portrayal of female subjects in word embeddings used in Global South countries to effectively address gender equality and women's rights issues.

The main contributions of this work are summarized as follows:

- Our research lays the groundwork for further exploration into validating bias metrics beyond the English-speaking world while addressing the specificities of Portuguese.
- We address the relatively unexplored field of characterizing word embeddings for the Portuguese language.
- By focusing on Portuguese texts, our work makes essential resources available for linguistic and sociological

analysis, fostering further research on gender bias and language representation.
- We found that metrics such as SC-WEAT and RIPA, widely used for bias analysis in English, can also be utilized and adapted for Portuguese. However, we observed different results due to the morphological nature of Portuguese. This reinforces the need for careful consideration and new research frontiers with other types of techniques.
- Our research is directly aligned with the United Nations Sustainable Development Goals 5 and 10, advancing the topics of “Gender Equality” and “Reduction of Inequalities” [United Nations, 2024].

The subsequent sections of this manuscript delineate the pertinent literature concerning gender bias within Portuguese Word Embeddings, elucidate the experiments conducted in this study, present the findings, and provide conclusions.

2 Related Work

Our literature review was conducted through systematic searches in scientific databases using keywords such as *gender bias*, *NLP*, and *artificial inteligence*. The selected studies were analyzed with consideration of their linguistic contexts.

Freitas and Martins [2023] present results obtained from direct analysis of texts in literary books. They used NLP techniques (POS-tagging and dependency parsing) to look for words describing male and female characters. Their results emphasized the portrayal of women via vocabulary associated with physical appearance and domestic roles.

Salles and Pappa [2021] highlight the words most strongly associated with each gender in Wikipedia biographies using TF-IDF and the Pointwise Mutual Information metric. The results indicate that words most associated with men are related to sports, while those most associated with women are related to the artistic sphere. Frequently, descriptions of women are framed using predicates employed in the passive voice. Trainotti Rabonato *et al.* [2025] analyze the context of gender bias in machine translation from English to Portuguese. By creating a dataset of English-Portuguese sentences designed to mitigate bias and using post-processing techniques, they achieve significantly satisfactory results.

Caliskan *et al.* [2022], the authors expose very important findings about gender bias in English word embeddings trained on Internet corpora. They demonstrated that: i) The majority of the most frequent words are more associated with men than women; ii) The top male-associated words predominantly consist of verbs, whereas the top female-associated words primarily consist of adjectives and adverbs; iii) The top male-associated concepts encompass roles and domains in big tech, engineering, religion, sports, and violence. In contrast, the top female-associated concepts are less focused on roles and include female-specific slurs, sexual content, appearance, and kitchen-related terms; iv) Male-associated words are higher in arousal and dominance, while female-associated words are higher in valence. In this work, we hypothesize and confirm through experiments that word embeddings trained in Portuguese Internet corpora exhibit similar masculine protagonism.

Our results also align with other findings within the Portuguese language domain. In this realm, two works drawing upon WEAT merit attention. Taso *et al.* [2023a] examined gender bias in occupation-related words. They identified gender stereotypes, exemplified by occupations such as “Nurse” being most similar to the “she” pronoun, and “Soldier” being most similar to the “he” pronoun. Through the study of analogies, biases such as “vocalist-guitarist” analogized with the pronouns “she-he” and “volleyball-football” were also identified.

Taso *et al.* [2023b] demonstrated stereotypical associations reflective of Brazilian society, noting that family-related concepts exhibit stronger connections with female terms, while career-related concepts are more closely associated with male terms. They observed that the Word Embedding model successfully captured characteristics of the Brazilian job market, highlighting a correlation between occupations with a high percentage of women and their respective word associations with female terms.

Assi and Caseli [2024] uses the GPT-3.5 Turbo model as an evaluation object for the perception of “regard” (according to them, “regard” was considered a metric to calculate the model’s level of respect) for certain groups, in this case, genders (masculine, feminine, neutral) in English and Portuguese. They noted, contrary to their initial hypothesis, that the model has a slightly higher positive bias toward the feminine gender. However, they observed the model’s preference for the English language. Thus, they emphasize the importance of studying models in multiple languages since this can significantly impact their performance. Still within the Brazilian context, it is important to highlight the systematic review of de Lima and de Araujo [2023]. The authors present Brazilian works that address algorithmic fairness in some way and categorize the literature results and suggestions for advancement in the NLP field. They conclude that the area has a broad space of opportunities such as studies and mitigation from a racial perspective, in the context of linguistic variation within Portuguese, and evaluations of public datasets.

In other languages that also have morphologically marked gender, such as Spanish, and other languages with similar grammatical situations, such as French, German, and Polish, gender bias has been demonstrated in studies conducted by Raymakers [2020] and Kurpicz-Briki [2020]. Raymakers [2020] finds that Spanish closely aligns with the results presented in Portuguese.

3 Design of Experiments

Our experimental design aligns with the methodology outlined by Caliskan *et al.* [2022] for bias detection in English-written texts. Nevertheless, our data preparation methodology diverges to accommodate the available resources and address the specificities of the Portuguese language. The experiments encompass phases for word embedding selection, estimation of word similarity, identification of topics, and characterization of emotional states.

Word Embedding

Few Portuguese Word Embeddings are publicly available. Among them, we have selected the 300-dimensional GloVe proposed by Hartmann *et al.* [2017] due to its widespread adoption in syntax and semantic tasks. The model is composed of diverse corpora from Brazil and Portugal, comprising a total of 1.2 billion tokens. The sources of the 17 corpora used in the WE are diverse, coming from an internet-based encyclopedia, online newspapers, popular science magazines, news websites for children, literary works, and subtitles crawled from an internet movie database.

Due to the nature of the corpus, it is not possible to measure the influence that European Portuguese and Brazilian Portuguese have on the identified biases. However, the results are valuable for social studies to analyze them in light of gender discrimination statistics found in both countries.

Metrics

Implicit Association Test (IAT) assesses implicit bias in human subjects by measuring the speed and accuracy of their association between category (e.g. flowers) and attribute terms (e.g. love, peace, and happiness) [Greenwald *et al.*, 1998]. We measure gender bias through the use of Single-Category Word Embedding Association Test (SC-WEAT) – an adaptation of IAT to word embeddings [Caliskan *et al.*, 2017; Greenwald *et al.*, 1998]. SC-WEAT assumes that cosine similarity, a metric frequently used to measure semantic similarity between words represented in vector space, is analogous to the reaction time in the IAT. That is, the shorter the decision time, the greater the semantic proximity.

Therefore, SC-WEAT uses the cosine of the angle of two vectors to estimate the semantic similarity of two words represented by them in the multidimensional space. This theory is applied to the differential association of a single word with two groups of attribute words. Our work considers the words extracted from the corpus and their mean similarity with groups of masculine and feminine terms.

The formula for the SC-WEAT is given by Equation 1. Vector $\vec{w}$ is the target word, which belongs to the model's vocabulary (e.g. *cozinha*, *escritório*, *televisão*). Sets A e B are the sets of gender attributes, defined and validated by Caliskan *et al.* [2022]. Female attributes are *feminino, ela, dela/delas, mulher, menina, filha, irmã*[1], whereas male attributes encompass *masculino, ele, dele/deles, homem, menino, filho, irmão*[2]:

$$ES(\vec{w}, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std\_dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})} . \tag{1}$$

Effect size (ES) denotes the strength of association. In accordance with Cohen's d, absolute values of ES greater than 0.8, 0.5 and 0.2 indicate a strong, moderate and weak association, respectively. A positive effect size corresponds to an association with female attributes, while a negative effect size corresponds to an association with male attributes.

Effect size serves as the foundation for all analyses presented in this work.

Although metrics based on WEAT are widely used in gender bias analysis, there are other studies that explore their limitations. Notably, research by Ethayarajh *et al.* [2019] demonstrates that Caliskan's metric tends to overemphasize bias. To address this, the authors propose a new bias evaluation metric using word association values, called Relational Inner Product Association (RIPA). The authors show that their new metric does not amplify gender bias within the embedding space, while still highlighting words that are specifically gender-related. As the name suggests, RIPA calculates the inner product between each word vector and a *relational vector*, which is the first principal component of a gender embedding calculated by the difference between all specified gender pair vectors.

For comparison purposes, and in order to validate the results obtained by SC-WEAT, we considered it important to observe and discuss the outcomes of using SC-WEAT and RIPA. Thus, results with both metrics are presented in part of our analysis.

Most Frequent Words

For the analysis of gender association in the most frequent words, we created four subsets with the top 100, 1.000, 10.000, and 20.000 most frequent words in the model. Stopwords and words with no meaning were excluded from the lists using the MAC-MORPHO corpus – a manually annotated corpus with part-of-speech tags with over one million texts retrieved from a Brazilian daily newspaper [Fonseca *et al.*, 2015].

Subsets larger than 20000 words were disregarded due to the presence of many items with few occurrences.

Parts-of-Speech

We employed Floresta Treebank for POS-tagging [Freitas *et al.*, 2008]. Floresta is a treebank for Portuguese that contains syntactically annotated sentences, manually tagged by linguistic specialists, containing more than six million words collected from multiple sources such as journalistic, scientific, and cultural texts.

Only the most frequent words of our model, which were also contained in Floresta, were used to build two lists of gendered POS-words – one with words moderately related to the feminine attributes with a minimum ES of 0.2; and another with words moderately related to the masculine attributes with an $ES \leq -0.2$. We built a subset with 1.500 POS-words for each list and subdivided it into 3 more subsets of 100, 500 and 1.000 words. The same concept was applied to RIPA metric, considering the sign of the score.

Words belonging to some grammatical classes, such as nouns and verbs, were subdivided. The former were categorized into feminine and masculine nouns. The latter, into finite, gerund, infinitive, and participle.

The verb group showed interesting results, so we made them explicit in a separate table. The POS tagger already displays the nominal forms of the verbs properly classified, so no preprocessing was necessary.

[1] Translations of female, she, her, woman, girl, daughter, and sister.

[2] Translations of male, he, his, man, boy, son, and brother.

Clustering

We employed K-Means Clustering, with Elkan Method to partition words into non-overlapping groups. K-Means is well-suited for our problem due to its bottom-up technique, which reveals semantic categories in an organic manner.

Due to the limited vocabulary size of our model, using a moderate association effect size, such as $ES \geq 0.5$, did not yield enough points for clustering. Therefore, we systematically reduced this threshold until the partitions became visually meaningful. We chose to include only words with an absolute $ES \geq 0.3$ for female-associated terms and $ES \leq 0.3$ for male-associated terms. In total, the top $1,000$ words were considered for each gender subset.

4 Results

For the sake of a clear analysis, the results are presented considering three aspects: most frequent subsets, parts-of-speech tagging, and clustering.

Frequency of Words Associated with each Gender

Figure 1 presents the effect size distribution of words for subsets of length 100, 1.000, 10.000 and 20.000. The mean effect size for all the sets was between 0 and -0.3, denoting a tendency of association of the model with masculine terms. The great majority of the 100th most frequent words are associated with male attributes. More specifically, 89% of the words in the first group are correlated to men's context. The discrepancy between genders diminishes when the number of words increases. For the three other groups, the percentage of words related to male attributes is 77.9%, 63.5%, and 57.5%, as shown in Figure 2. We can also notice that as the size of the groups grows, the effect size increases, indicating that new extreme gender-related terms appear.

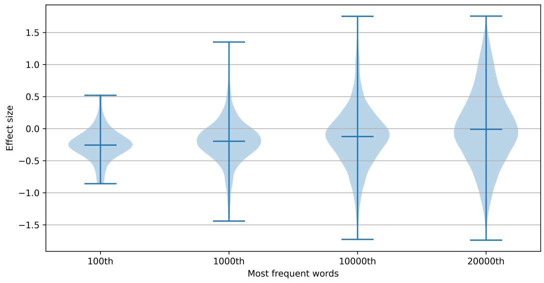

Figure 1. WEAT's Effect Size distribuition across the subsets.

Similarly, RIPA also showed mean values between 0 and -0.3. However, Figure 3 shows that these values are much more concentrated within this range across all frequency subsets, unlike SC-WEAT. Figure 4 shows in percentage terms the relationship between RIPA and feminine and masculine attributes. Across all subsets, the percentage was higher than the percentages presented by SC-WEAT. The first subset of the 100 most frequent words showed 93% associations with masculine attributes, 4% more than SC-WEAT. These results indicate that the subsets contain words with RIPA values more strongly associated with masculine attributes compared to SC-WEAT, despite its premise suggesting otherwise.

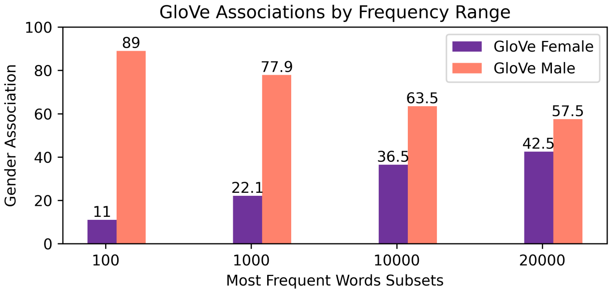

Figure 2. Gender Association Percentage in four groups of Frequency Range.

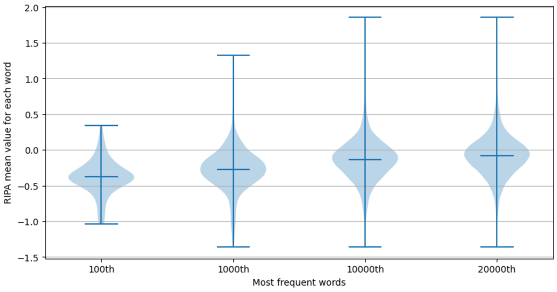

Figure 3. RIPA mean value across the subsets.

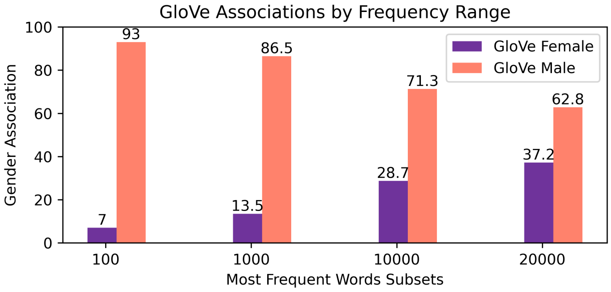

Figure 4. Gender Association Percentage in four groups of Frequency Range using RIPA values.

The results show evidence that topics intertwined with the men-world dominate the discourses in the texts used to train the model. Following the idea that frequency shapes our preferences, decisions based on this model would consider men's topics more important than the women's ones [Zajonc, 2001].

Parts-of-speech Associated with each Gender

The analysis of POS-tagging shed light on grammatical classes associated with gender. Table 1 summarizes the results of SC-WEAT and RIPA. Nouns constitute the most common class in all groups of male- and female-associated words, making up 50% of the words in the subsets in both metrics.

Substantives in Portuguese present female and male genders. Consequently, an approximation to attributes representing the same sex is expected in the GloVe model. Aiming to show that bias can be strong enough to go against this tendency, we compiled, for each list of nouns associated with a gender, those that are related to the opposite gender. For instance, *amor* (love) is a masculine noun but was associated with attributes related to women, with an effect size of 0.44. Other examples for the feminine association are: *câncer* (cancer), *cabelos* (hair), and *divórcio* (divorce), with effect sizes of 0.48, 0.7, and 0.4, respectively. For the male association, some prominent instances are: *guerra* (war), *liderança* (leadership), and *conquista* (achievement), with effect sizes of 0.51, 0.56, and 0.54. Each of the given examples falls outside the mean effect size observed in every frequency-words subset (Figure 1).

Adjectives represent 21% of the words related to women in the 100-words subset, compared to only 9% for words related to men on SC-WEAT metric. This trend persists in the other subsets, with an average of 19% compared to 15%. In RIPA, the same pattern is observed, with virtually no significant difference in proportionality across the classes. As Caliskan *et al*. [2022] emphasize in their work, this heightened association of adjectives with feminine terms underscores the necessity for additional information and description when referring to women.

When considering subsets greater than 500 words, we observe, in both metrics, a higher frequency of verb occurrences in groups related to the female gender. To further investigate hypotheses justifying such a difference, we categorized verbs according to their nominal forms – infinitive, gerund, participle, and finite (conjugated verbs expressing the notion of time). Table 2 presents the prevalence of each category. Participle forms overwhelmingly dominate all feminine subsets. For the 100-words subset using SC-WEAT metric, participle represents 75% of the verbs. For the 1500-subset, it represents 55%. Such a pattern is not noticed for the male subsets. For instance, no participle is present in the male 100-words subset using SC-WEAT. Verbs that correspond to male gender are primarily non-participles, expressing actions in simple present, past and future tense.

In Portuguese language syntax, a word ending in *ado* or *ido* (and their endings *-a* and *-s* for feminine and plural) can serve as a participle of a verb or an adjective. For instance, "*A notícia foi anunciada por ela.*" (The news was announced by her.) and "*O vento enfurecido açoitava a rancharia.*" (The furious wind lashed the ranch.). The classification is dependent on the context [Cunha and Cintra, 2016]. Since GloVe embeddings do not provide any context information, we consider these words as participles, irrespective of whether they function as one or the other.

If we acknowledge that a portion of the feminine participles can indeed function as adjectives, it reinforces that women are excessively described in texts. Conversely, if we consider them entirely as verbs, we could assume that the passive voice is frequently used in topics related to the female universe. This might suggest that women are frequently portrayed as recipients of actions rather than active agents in their own activities. In either case, it leads to the conclusion that our model tends to treat men as protagonists. This result aligns with Caliskan *et al*. [2022] observations, highlighting that men are depicted as active agents in the world and in the majority of textual data, in contrast to women who are often described differently. Just as in the first part of the POS-tagging experiment, we noticed that the results obtained with RIPA remained aligned with SC-WEAT. This helps us confirm the presence of gender bias across different metrics.

Clustering

Twelve clusters were obtained for each subset of gender association. Figures 5 and 6 present the distribution of these clusters for female and male data, respectively. Subsets of 1.000 words were used along with TSNE algorithm for dimensionality reduction. Sports and financial terms were noticed in the male subset. Partitions with words related to *feelings* and terms associated with *clothing* and *household furniture* were found in the female subset. A cluster of words related to *courses* and *science* was also identified in the group of male attributes, while clusters related to *nature* and *music* were identified in the group of female attributes, reinforcing these gender stereotypes.

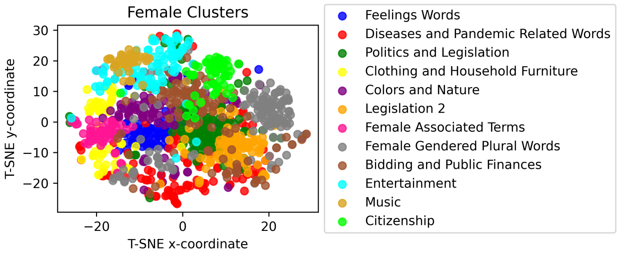

Figure 5. Female Clusters Visualization.

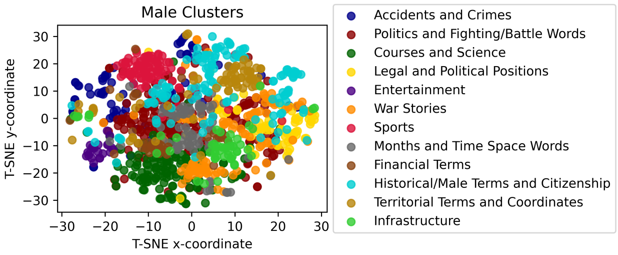

Figure 6. Male Clusters Visualization.

The aggregation of words around politics was noticed in both subsets. We attribute these agglomerations to the origin of the vocabulary used in the model and in our POS-tagging processing. Despite the existence of political concepts in both sets, the cluster *Legal and Political Positions*, consisting of profession names with a high dominance of men in Brazil, is exclusively present in the male dataset. Once again, this discrepancy in representation intensifies gender bias in the job market. Table 3 presents English translations of relevant words in each cluster. The clusters were found using KMeans with Elkan algorithm computed from two sets of 1000 words from Portuguese GloVe model. The original words are available in Appendices A and B. Since the RIPA results for the most frequent subsets and part-of-speech tagging aligned with WEAT, validating this metric, no word clustering was conducted.

5 Work Limitation

None of the authors are specialists in linguistics or sociology. However, they have a strong interest in evidencing

Table 1. Proportion of grammatical classes accordingly of Frequency Range for each set of Female Associated Words and Male Associated Words using SC-WEAT and RIPA metric.

Parts-of-Speech for the Gender-Associated Words Subsets – GloVe									
	POS	100		500		1000		1500	
		Female	Male	Female	Male	Female	Male	Female	Male
SC-WEAT	Adjectives	21 (21%)	9 (9%)	92 (18%)	77 (15.4%)	191 (19.1%)	158 (15.8%)	292 (19.4%)	223 (14.8%)
	Adverbs	0 (0 %)	11 (11%)	3 (0.6%)	28 (5.6%)	10 (1%)	36 (3.6%)	21 (1.4%)	41 (2.7%)
	Nouns	64 (64%)	49 (49%)	295 (59%)	271 (54.2%)	534 (53.4%)	545 (54.5%)	728 (48.5%)	799 (53.2%)
	Pronouns	7 (7%)	11 (11%)	12 (2.4%)	27 (5.4%)	14 (1.4%)	29 (2.9%)	14 (0.9%)	30 (2%)
	Verbs	4 (4%)	7 (7%)	83 (16.6%)	72 (14.4%)	230 (23%)	193 (19.3%)	418 (27.8%)	356 (23.7%)
	Others	3 (3%)	13 (13%)	15 (3%)	25 (5%)	21 (2.1%)	39 (3.9%)	27 (1.8%)	51 (3.4%)
RIPA	Adjectives	20 (20%)	6 (6%)	82 (16.4%)	59 (11.8%)	164 (16.4%)	114 (11.4%)	249 (16.6%)	187 (12.5%)
	Adverbs	0 (0%)	14 (14%)	5 (1%)	36 (7.2%)	22 (22%)	63 (6.3%)	30 (2%)	81 (5.4%)
	Nouns	64 (64%)	44 (44%)	304 (60.8%)	274 (54.8%)	558 (55.8%)	546 (54.6%)	796 (53%)	812 (54.1%)
	Pronouns	10 (10%)	9 (9%)	16 (3.2%)	26 (5.2%)	20 (2%)	38 (3.8%)	21 (1.4%)	43 (2.9%)
	Verbs	2 (2%)	12 (12%)	81 (16.2%)	77 (15.4%)	209 (20.9%)	203 (20.3%)	373 (24.9%)	327 (21.8%)
	Others	4 (4%)	15 (15%)	12 (2.4%)	28 (5.6%)	27 (2.7%)	36 (3.6%)	31 (2.1%)	50 (3.3%)

Table 2. Number of verbs grouped by their nominal form.

Parts-of-Speech – Verbs Annotation									
Verb's Nominal Forms		100		500		1000		1500	
		Female	Male	Female	Male	Female	Male	Female	Male
SC-WEAT	Finite	1	4	24	37	66	90	123	143
	Gerund	0	2	1	4	2	6	8	12
	Infinitive	0	1	9	12	33	39	59	70
	Participle	3	0	49	19	129	58	228	131
	Participle %	75%	0%	60%	26%	56%	30%	55%	37%
	Non Participle %	25%	100%	40%	74%	44%	70%	45%	63%
RIPA	Finite	0	8	31	45	74	104	144	171
	Gerund	0	1	2	4	5	6	7	14
	Infinitive	0	2	16	16	40	58	67	84
	Participle	2	1	32	12	90	35	155	58
	Participle %	100%	8%	40%	16%	43%	17%	42%	18%
	Non Participle %	0%	92%	60%	84%	57%	83%	58%	82%

how computing impacts society and how to measure these effects, especially in the Brazilian context. Due to the nature of the word embeddings used, with texts originated in Brazil and Portugal, it is not possible to directly attribute the biases found in our analysis only to historical sexism in Brazil. Nevertheless, our findings align with socio-statistical data and text analysis related to gender bias in the country, suggesting meaningful correlations.

In this research, the model and metrics highlighted the challenges of working with Indo-European Romance languages, which have complex grammatical structures involving gender, number, and class inflections. Unlike English, where nouns are generally neutral unless referring to living beings, Portuguese assigns a grammatical gender—either masculine or feminine—to all nouns. Additionally, Portuguese grammar follows a rule in which any mixed-gender group is automatically referred to using the masculine form, regardless of the number of feminine elements present. This linguistic characteristic can influence biases in language models and presents unique challenges for computational processing.

Constructing a methodology that adequately encompasses the intricacies of such languages proves to be challenging. The use of metrics originally designed for the English language, which lacks grammatical gender, revealed notable distinctions in results, particularly in the Part-of-Speech (POS-Tag) analysis, emphasizing peculiarities in verbal forms.

Drawing from Sabbaghi and Caliskan's insights [Omrani Sabbaghi and Caliskan, 2022], the analysis of inflective languages necessitates the "disentanglement" of word gender to enhance expressiveness in the language-specific context, with the adoption of conventional metrics. Nevertheless, the importance of employing metrics tailored for the Portuguese language, and Romance languages in general, is emphasized to enhance the accuracy and dependability of results. This holds significance not only for investigations into biases and prejudices in models, but also for the development of applications using Natural Language Processing models in regions where the language is spoken, stressing the importance of creating language-specific tools.

The limitation of using a smaller set of words for experiments, restricted to a maximum of 20.000 words in Frequency, was imposed due to the identification of meaningless words in the Portuguese model's vocabulary. These words became more prevalent as the set size increased. Post-processing mitigated this issue by reducing the subset size, hindering the exploration of experiment behaviour in larger groups. This limitation impacts the analysis, preventing a

Table 3. Female and Male Clusters Examples of words.

Female Clusters	**Examples**	**Male Clusters**	**Examples**
Feelings Words	love, surprise, soul, sensation, dramatic, sensitivity, purity	Accidents and Crimes	body, car, vehicle, fire, dead, police officer, iron
Diseases and Pandemic Related Words	cancer, surgery, vaccine, flu, abortion, doses, genetic	Politics and Fighting/Battle Words	democratic, movement, dialogue, forces, attack, combat, conflict
Politics and Legislation	initiative, interpretation, forecast, agenda, coalition, legislative, budget	Courses and Science	graduation, professional, science, medicine, engineering, computer, laboratory
Clothing and Household Furniture	room, table, clothes, bed, window, kitchen, dress	Legal and Political Positions	president, minister, deputy, governor, secretary, judge, lawyer
Colors and Nature	light, waters, beach, color, star, flowers, moon	Entertainment	film, career, actor, singer, producer, composer, guitar
Legislation (2)	administrative, institution, community, entity, license, contribution, regulation	War Stories	war, men, military, dead, troops, attacks, soldiers, civilians, allies
Female Associated Terms	mother, woman, lady, daughter, marriage, child, wife	Sports	game, points, football, team, field, match, player
Female Gendered Plural Words	two, small, localities, islands, substances, roots, magazines	Months and Time Space Words	year, days, january, period, june, past, march
Bidding and Public Finances	income, private, village, matrix, basin, railway, auction, plant	Financial Terms	total, value, money, product, price, average, payment
Entertainment	television, party, festival, exhibition, play, soap opera, carnival	Historical/Male Terms and Citizenship	lord, son, father, Portuguese, king, French, general
Music	music, version, voice, singer, artist, dance, tour	Territorial Terms and Coordinates	river, area, south, place, north, port, municipality
Female Gendered Citizenship Words	Brazilian, Portuguese, French, American, Italian, British, Spanish	Infrastructure	development, management, transport, investment, transit, traffic, automobile

more comprehensive understanding of the model across a broader spectrum.

It would be interesting to explore contextual and advanced models that account for factors such as polysemy, where word meanings vary based on context, as demonstrated by models like BERT [Devlin *et al.*, 2019]. For languages like Portuguese, a contextual analysis becomes essential to capture the nuances and intricacies of word usage in different contexts. Another possibility, considering educational and academic purposes, would be to organize workshops where members of the university community could conduct contextual analyses of the texts used in the WE.

6 Threats to Validity

Regarding external validity, we know that SC-WEAT can be used as an analysis for different groups and situations. Caliskan *et al.* [2017], the authors showed the WEAT metric and how it can be used to verify historical bias from moral concepts between flowers and insects, to race and gender bias. In our study, we used SC-WEAT for gender bias analysis, but race studies can also be analyzed using this metric.

A possible threat to internal validity would be the use of only one type of Word Embeddings model. We used the 300-dimensional GloVe from Hartmann *et al.* [2017] that was made freely available. However, we did not conduct experiments with other models they provide or at least use another model from a different source. At the time of our research, we found that the model we chose was the most efficient, and experiments with other available ones would have yielded much inferior results and possibly with much more amplified bias. The lack of good-performing model options available in Portuguese led us to limit our experiments to just one.

Regarding conclusion validity, we found that criticisms of WEAT/SC-WEAT usage exist where they claim that this metric can be manipulated to increase bias in a desired group. Ethayarajh *et al.* [2019] provide one of the works that criticizes its use. We also used their gender bias metric, RIPA, to verify if there was a significant difference in results. However, we found that the RIPA metric reached the same values as SC-WEAT and, in most cases, a higher value than the criticized metric. Nevertheless, we acknowledge that other metrics beyond these exist but were not the object of this study, and other comparisons could have been investigated.

7 On the Perpetuation of Bias

The United Nations Sustainable Development Goals (SDGs) include gender equality as a central priority for building more just and inclusive societies. SDG 5, "Gender Equality", aims to eliminate all forms of discrimination against women and girls, ensuring their full access to rights, opportunities, and participation in all spheres of society. Additionally, SDG 10, "Reduced Inequalities", reinforces the need to combat the marginalization of vulnerable groups, including women, by promoting equitable access to resources and decision-making power. These goals encourage the development of policies and practices that challenge gender stereotypes and foster a global environment of equality United Nations [2024].

Despite these provisions, observable advancements in compliance with these measures remain modest and hesitant. In Brazil, a study revealed that 84.5% of people hold at least one form of bias against women. The most concerning indicators relate to physical integrity, which includes intimate partner violence and the right to decide whether or not to have children, with 75.56% of respondents expressing such prejudice. Additionally, 39.91% of people believe that women are not as competent in politics as men. Finally, 31% of respondents consider men to be more suited for business [UNDP].

These statistics reported by communication news are also manifested in cultural expressions such as literature and music [Freitas and Martins, 2023; Firmino *et al.*, 2024].

Considering that all kinds of information are present on the World Web, the generation and continuous updating of large language models (LLMs) through data extraction from the Internet, without adequate scrutiny of their content, raise concerns regarding the permissible or advisable size limits for LLMs. Bender et al. [Bender *et al.*, 2021] illuminate perspectives on the environmental, social, economic, and cultural repercussions associated with the development and maintenance of LLMs employing extensive training datasets, particularly in their impact on marginalized groups. The authors analyze and discuss the selection and incorporation of textual data sourced from the Internet into the model, emphasizing that the data generated on the Web are not diverse and global, but rather hegemonic and limited, primarily originating from texts produced by young people. This reliance on a non-diversified dataset for model training raises critical questions about the representativeness and inclusivity of LLMs in reflecting a broader spectrum of perspectives.

Despite the proportional improvement in the performance of word embedding and language model (LM) systems with an increase in textual data, Bender *et al.* [2021] assert that the benefits and costs of these technological advancements are not equally distributed. The generation of sexist texts, along with other forms of discrimination and violent ideologies, coupled with the widespread implementation of LLMs and WE models in classification systems, can cause allocative damage. The representation of these biases significantly influences decision-making processes. Consequently, the authors advocate for counter-hegemonic solutions to disrupt the feedback loop of biases generated, disseminated across the Web, and reintegrated into new training data.

8 Conclusion

This research conducted a comprehensive analysis of gender bias within Portuguese Word Embedding models, employing a variety of experiments based on the framework of Caliskan *et al.* [2022]. The findings revealed significant bias patterns, particularly regarding word frequency and part-of-speech characteristics. Notable differences arose compared to studies in English, underscoring the importance of adapting the analyses to the specific linguistic nuances of Portuguese.

The association of most frequent words with male terms, particularly in the first frequency word groups, suggests a pervasive gender bias within the vocabulary. Interestingly, the prominence of adjectives in the female terms group and the lower presence of verbs in the male terms group deviate from English language patterns. The association of inflected verbs in the participle form with female terms points to a passive voice (i.e. women are not active participants in the sentence but rather recipients of action, in contrast to men), and adds depth to further understanding of gender stereotypes embedded in the language.

The clustering analysis further substantiated gender stereotypes, revealing distinct thematic clusters associated with female and male terms. These clusters, ranging from 'Feelings' and 'Nature' for female terms to 'Sports' and 'Political and Legal Positions' for male terms, underscored societal gender norms embedded in linguistic representations.

Our findings underline the necessity of using metrics that account for grammatical and morphosyntactic disparities in the Portuguese language. Moreover, this study emphasizes the significance of discussing gender bias in Portuguese word embedding models within the Brazilian context. The identified biases carry implications for societal perceptions and expectations, necessitating a nuanced approach to the development and application of language models in Portuguese. Therefore, this research advocates for an interdisciplinary lens in exploring biases within technological systems and encourages ongoing discourse on gender bias in language models, acknowledging its societal ramifications and the need for context-specific analyses.

Ultimately, it is interesting to note that the metrics used can also be employed and adapted for studies on different types of marginalized groups that lack research within the Portuguese language, as de Lima and de Araujo [2023] pointed out. Our findings and methodology can serve as a starting point for racial, class, and specific group analyses such as the LGBTQIA+ community. Beyond gender, and beyond Word Embeddings models, new research is necessary within the current context with the massive use of state-of-the-art models like ChatGPT and Gemini, and their biases toward different groups of interest mentioned need to be investigated and discussed. We understand that a study on biases with more current and contextual models is necessary. However, we believe the historical construction of research within the NLP field is necessary. We found that there were few works regarding gender bias in Word Embeddings in Portuguese, and the starting point was these fundamental models. New research is being planned to advance studies on contextual models and thus enable us to trace a historical evolution of models and their biases through this study.

A Female Clusters Portuguese Translations

Feelings Words: amor, surpresa, alma, sensação, dramática, sensibilidade, pureza. **Diseases and Pandemic Related Words**: câncer, cirurgia, vacina, gripe, aborto, doses, genética. **Politics and Legislation**: iniciativa, interpretação, previsão, agenda, coligação, legislativa, verba. **Clothing and Household Furniture**: quarto, mesa, roupas, cama, janela, cozinha, vestido. **Colors and Nature**: luz, águas, praia, cor, estrela, flores, lua. **Legislation (2)**: administrativa, instituição, comunitária, entidade, licença, contribuição, regulação. **Female Associated Terms**: mãe, mulher, senhora, filha, casamento, criança, esposa. **Female Gendered Plural Words**: duas, pequenas, localidades, ilhas, substâncias, raízes, revistas. **Bidding and Public Finances**: renda, privada, aldeia, matriz, bacia, ferroviária, leilão, usina. **Entertainment**: televisão, festa, festival, exibição, dança, novela, carnaval. **Music**: música, versão, voz, cantora, artista, dança, turnê. **Female Gendered Citizenship Words**: brasileira, portuguesa, francesa, americana, italiana, britânica, espanhola.

B Male Clusters Portuguese Translations

Accidents and Crimes: corpo, carro, veículo, fogo, morto, policial, ferro. **Politics and Fighting/Batlle Words**: democrático, movimento, diálogo, forças, ataque, combate, conflito. **Courses and Science**: formação, profissional, ciências, medicina, engenharia, computador, laboratório. **Legal and Political Positions**: presidente, ministro, deputado, governador, secretário, juiz, advogado. **Entertainment**: filme, carreira, ator, cantor, produtor, compositor, guitarra. **War Stories**: guerra, homens, militares, mortos, tropas, ataques, soldados, civis, aliados. **Sports**: jogo, pontos, futebol, equipe, campo, partida, jogador. **Months and Time Space Words**: ano, dias, janeiro, período, junho, passado, março. **Financial Terms**: total, valor, dinheiro, produto, preço, médio, pagamento. **Historical/Male Terms and Citizenship**: senhor, filho, pai, português, rei, francês, general. **Territorial Terms and Coordinates**: rio, área, sul, local, norte, porto, município. **Infrastructure**: desenvolvimento, gestão, transporte, investimento, trânsito, tráfego, automóvel.

Declarations

Acknowledgements

The present work was supported by the Federal University of Mato Grosso do Sul (UFMS), Brazil, and Leuphana Universität Lüneburg, Germany.

Authors' Contributions

Fernanda T. S. Taso contributed to the conceptualization, investigation, data curation, methodology, and writing of this work. Valéria Q. Reis and Fábio V. Martinez supervised Fernanda, collaborated in the methodology implementation, and reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data and code used for the analysis are available in the repository `https://doi.org/10.6084/m9.figshare.29485247.v2`. Last access: 07-13-2025.

Citation Diversity Statement

We acknowledge that the authors' backgrounds and life experiences may have influenced the direction of this research and the critical interpretation of the data. For this reason, we provide a brief introduction of who we are.

All authors hold academic degrees in Computer Science from the Federal University of Mato Grosso do Sul. Fernanda is a young computing enthusiast and an avid reader of books on technology and sociology.

Valéria is a professor whose research focuses on the societal impacts of computing. She teaches the course Computing and Society, where she frequently discusses gender discrimination in the field of Computer Science.

Fábio is a professor leading a research group on algorithmic neutrality and gender, with research interests that include social issues, focused on the ethical implications of technology in contemporary society.

References

Assi, F. and Caseli, H. (2024). Biases in gpt-3.5 turbo model: a case study regarding gender and language. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 294–305, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/stil.2024.245358.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proc. of ACM-FAccT*, page 610–623, Canada. Association for Computing Machinery. DOI: https://doi.org/10.1145/3442188.3445922.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Advances in neural information processing systems, 29. DOI: https://doi.org/10.5555/3157382.3157584.

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., and Banaji, M. R. (2022). Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 156–170, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3514094.3534162.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. DOI: https://doi.org/10.1126/science.aal4230.

Chaloner, K. and Maldonado, A. (2019). Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In Costa-jussà, M. R., Hardmeier, C., Radford, W., and Webster, K., editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W19-3804.

Chen, X., Li, M., Yan, R., Gao, X., and Zhang, X. (2022). Unsupervised mitigating gender bias by character components: A case study of Chinese word embedding. In *Proc. of GeBNLP*, pages 121–128, Seattle, Washington. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2022.gebnlp-1.14.

Cunha, C. and Cintra, L. (2016). *Nova Gramática do Português Contemporâneo*. Lexicon, NY, USA, 7 edition.

de Lima, L. F. and de Araujo, R. (2023). A call for a research agenda on fair NLP for Portuguese. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 187–192, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/stil.2023.233763.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics. DOI: https://doi.org/10.18653/V1/N19-1423.

Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding undesirable word embedding associations. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/P19-1166.

Falcão, C. (2021). Lentes racistas: Rui costa está transformando a bahia em um laboratório de vigilância com reconhecimento facial. `https://interc.pt/3nKVrw9`. Last access: 07-09-2025 (in Portuguese).

Firmino, V., Lopes, J., and Reis, V. (2024). Identificando padrões de sexismo na música brasileira através do processamento de linguagem natural. In *Anais do V Workshop sobre as Implicações da Computação na Sociedade*, pages 59–69, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/wics.2024.2968.

Fonseca, E., G Rosa, J., and Aluísio, S. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of Brazilian Computer Society*, 21(2). DOI: https://doi.org/10.1186/s13173-014-0020-x.

Freitas, C. and Martins, F. (2023). Bela, recatada e do lar: o que a mineração de textos literários nos diz sobre a caracterização de personagens femininas e masculinas. *Fórum Linguístico*, 20(3). DOI: http://dx.doi.org/10.5007/1984-8412.2023.e86749.

Freitas, C., Rocha, P., and Bick, E. (2008). Floresta sintá(c)tica: Bigger, thicker and easier. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Computational Processing of the Portuguese Language*, pages 216–219, Berlin, Heidelberg. Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-540-85980-2_23.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644. DOI: https://doi.org/10.1073/pnas.1720347115.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *J Pers Soc Psychol*, 74(6):1464–80. DOI: https://doi.org/10.1037/0022-3514.74.6.1464.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In Paetzold, G. H. and Pinheiro, V., editors, *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação. https://aclanthology.org/W17-6615/.

IBGE (2024). Estatísticas de Gênero: Indicadores sociais das mulheres no Brasil. Available in: `https://biblioteca.ibge.gov.br/visualizacao/livros/liv102066_informativo.pdf`. Last access: 07-09-2025.

Kurpicz-Briki, M. (2020). Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) 16th Conference on Natural Language Process-ing (KONVENS)*, volume 2624, Zurich, Switzerland. CEUR Workshop proceedings. DOI: https://doi.org/10.24451/arbor.11922.

LESFEM (2024). Monitor de Feminicídios no Brasil. Available in: `https://tinyurl.com/4dn7f36w`. Last access: 07-09-2025.

Noble, S. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, NY, USA. DOI: https://doi.org/10.2307/j.ctt1pwt9w5.

Omrani Sabbaghi, S. and Caliskan, A. (2022). Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proc. of AIES*, page 518–531, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3514094.3534176.

Park, J. H., Shin, J., and Fung, P. (2018). Reducing gender bias in abusive language detection. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/D18-1302.

Raymakers, T. (2020). Gender bias in word embeddings of different languages. Bachelor's thesis, Delft University of Technology. https://tinyurl.com/ykunzjj6. Last access: 07-13-2025.

Salles, I. and Pappa, G. (2021). Viés de gênero em biografias da wikipédia em português. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 211–216, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/brasnam.2021.16142.

Silva, T. (2022). Linha do tempo do racismo algorítmico. `http://bit.ly/3yFFrzw`. Last access: 07-09-2025 (in Portuguese).

Sogancioglu, G., Mijsters, F., van Uden, A., and Peperzak, J. (2022). Gender bias in (non)-contextual clinical word embeddings for stereotypical medical categories. https://doi.org/10.48550/arXiv.2208.01341.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/P19-1159.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proc. of EAAMO*, volume 17, pages 1–9, NY, USA. ACM Press. DOI: https://doi.org/10.1145/3465416.3483305.

Taso, F., Reis, V., and Martinez, F. (2023a). Algorithmic gender discrimination: Case study and analysis in the brazilian context. In *Proc. of WICS*, pages 13–25, João Pessoa, PB, Brazil. SBC. (*in Portuguese*). DOI: https://doi.org/10.5753/wics.2023.229980.

Taso, F., Reis, V., and Martinez, F. (2023b). Sexismo no Brasil: análise de um Word Embedding por meio de testes baseados em associação implícita. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 53–62, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/stil.2023.233845.

Torres Berrú, Y., Batista, V., and Zhingre, L. (2023). A data mining approach to detecting bias and favoritism in public procurement. *Intell Autom Soft Co*, 36(3):3501–3516. DOI: http://dx.doi.org/10.32604/iasc.2023.035367.

Trainotti Rabonato, R., Milios, E., and Berton, L. (2025). Gender-neutral english to portuguese machine translator: Promoting inclusive language. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 180–195, Cham. Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-79038-6_13.

(UNDP), U. N. D. P. (2023). Breaking down gender biases shifting social norms towards gender equality. https://www.undp.org/sites/g/files/zskgke326/files/2023-06/gsni202302pdf_0.pdf. Last access: 07-09-2025.

United Nations (2024). Sustainable Development Goals. Available in: `https://sdgs.un.org/goals`. Last access: 07-09-2025.

Wagner, J. and Zarrieß, S. (2022). Do gender neutral affixes naturally reduce gender bias in static word embeddings? In *Proc. of KONVENS*, pages 88–97, Potsdam, Germany. KONVENS 2022 Organizers. https://aclanthology.org/2022.konvens-1.10/.

Wairagala, E. P., Mukiibi, J., Tusubira, J. F., Babirye, C., Nakatumba-Nabende, J., Katumba, A., and Ssenkungu, I. (2022). Gender bias evaluation in Luganda-English machine translation. In *Proc. of AMTA*, pages 274–286, Orlando, USA. AMTA. https://aclanthology.org/2022.amta-research.21/.

Werneck, A. (2019). Reconhecimento facial falha em segundo dia, e mulher inocente é confundida com criminosa já presa. `http://bit.ly/3mSoNKy`. Last access: 07-09-2025 (in Portuguese).

Yee, K., Tantipongpipat, U., and Mishra, S. (2021). Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency. *Proc. of HCI*, 5:1–24. DOI: https://doi.org/10.1145/3479594.

Zajonc, R. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, 10(6):224–228. DOI: https://doi.org/10.1111/1467-8721.00154.