




Muses or Stereotypes? Identifying Historical Patterns of Sexism in a Corpus of Brazilian Lyrics

Janaina Nogueira de Souza Lopes   [Federal University of Mato Grosso do Sul | janaina.nogueira@ufms.br]


Vitória Pereira Firmino  [Federal University of Mato Grosso do Sul | vitoria.firmino@ufms.br]

Valéria Quadros dos Reis  [Federal University of Mato Grosso do Sul and Leuphana University of Lüneburg | valeria.reis@ufms.br]

 Faculty of Computing, Federal University of Mato Grosso do Sul, Av. Costa e Silva, s/n, Pioneiros, Campo Grande, MS, 79070-900, Brazil.

Received: 29 November 2024 • Accepted: 23 May 2025 • Published: 05 June 2025

Abstract: This study aims to identify gender bias in Brazilian songs by analyzing the most frequent predicatives used to describe women and the most referenced occupations associated with the feminine gender. To achieve this, we created a corpus containing 146,612 song lyrics and applied natural language processing techniques to extract sentences that describe women. The identified predicatives were annotated and used to train a machine learning model that categorizes them into five descriptive categories. Additionally, we compiled a list of occupations mentioned in the lyrics. From a distant reading perspective, the results reveal a persistent historical pattern of sexism: women are predominantly portrayed through physical and emotional traits, while men are more frequently associated with character and social roles. In the professional domain, women are mainly depicted in caregiving and entertainment roles. These findings align with previous studies and contribute to methodological advancements in identifying gender bias in Portuguese-language texts.

 The text contains potentially harmful and offensive examples.

Keywords: Gender Bias, Natural Language Processing (NLP), Corpus-based Analysis, Brazilian Music, Algorithmic Discrimination

1 Introduction

Gender bias, or sexism, refers to the judgment of an individual based solely on their gender [Moss-Racusin *et al.*, 2012; Sun *et al.*, 2019]. This form of discrimination has historical roots in Brazil, where even today, monitoring indicators from IBGE, reveal significant inequalities between men and women [IBGE, 2024]. Women account for 51.5% of the population, yet they hold only 39.3% of managerial positions. Moreover, female representation is concentrated in occupational groups that demand attention to needs and well-being, such as Education, Health, and Social Services. The IBGE survey also indicates that women earn 78.8% of men's wages for the same positions. This data highlights the challenges women face in integrating into diverse professional and educational environments, demonstrating the persistence of gender barriers in Brazilian society.

The media and culture play a fundamental role in the fight against gender inequality; however certain manifestations still carry biases that limit female representation, emphasizing women's physical and emotional characteristics over their professional qualifications or character. In their study, Salles and Pappa [2021] computationally analyze gender bias in Portuguese Wikipedia biographies, where only 16% of Wikipedia biographies are about women. Biases also manifest through the objectification of women, which exalts appearance.

Freitas and Martins [2023] analyze a corpus of Portuguese-language literary works. The data analysis was conducted

using Natural Language Processing (NLP) methods, and the results reveal a stereotypical construction of genders. Femininity is primarily characterized by a strong association with the body, especially beauty. In contrast, masculinity is presented with an emphasis on social roles and character traits.

Expanding on the discussion of gender representation in media and literary pieces, similar patterns can be observed in other cultural expressions. Music holds significant weight in the contemporary landscape and is highly useful for studying issues related to gender, particularly in terms of language [Feijó and Macedo, 2013]. In this context, Kong [1995] states that music possesses a compelling structure: both as a medium and as a result of experience, it serves to produce and reproduce social systems. The author emphasizes that music is also a means through which people convey their experiences, bringing to light social constructions, feelings, and other aspects. Duprat [2008] highlights the representation of women in Brazilian music, often portrayed in a sexualized manner and more commonly associated with adjectives related to beauty. The author states that the power of man has manifested in music through composition and production, while that of woman has manifested through inspiration and seduction. It is as if the consecration of a world where the feminine does not oppose the masculine but rather seduces it.

Several authors, through manual analysis, have highlighted the presence of bias, sexism, and gender violence in Brazilian music [Lima and Sanches, 2009; Assis, 2014; Schlösser and Fantin, 2022]. However, automated ap-

proaches capable of processing and evaluating large volumes of data remain rare, especially in the context of the Portuguese language.

The study published by Firmino *et al.* [2024] demonstrates, through the use of NLP techniques, that there is a stereotypical representation of women in Brazilian music, where the most frequently used descriptive words predominantly relate to physical appearance. In contrast, men are portrayed with an emphasis on good character. The work's findings raise two new questions in the context of a limited selection of Brazilian song lyrics:

- **What is the distribution of the words used to describe women?** To answer this question, it is necessary to consider all those characteristics quoted in the songs and how they fit into pre-determined descriptive categories. The hypothesis is that the pattern based on female appearance persists even when less common descriptions are considered.
- **What are the most frequently mentioned occupations by gender?** The hypothesis is that the songs reflect patterns of sexism historically present in society, with women primarily engaging in occupations requiring attention and care for others.

Thus, the objective of this study is to extend the findings of Firmino *et al.* [2024], demonstrating that sexist stereotypes are manifested both in the predicatives and in the professions most attributed to each gender. For this purpose, a classification model was trained and employed with NLP techniques, such as word frequency calculation and categorization, to analyze a large corpus composed of 146,612 Portuguese-language lyrics, compiled exclusively for this study.

The results quantify and qualify patterns of sexism found in Brazilian lyrics, going beyond common sense and allowing for a systematic and comprehensive analysis of the language used in the music. Negative trends were identified in the language used to describe women, also revealing the limitations of female characterization in the professional field.

Finally, it is important to highlight that this study specifically analyzes stereotypes attributed to women. However, we acknowledge that other gender groups, such as those included in the LGBTQIAPN+ community, would also benefit from similar analyses.

2 Theoretical Basis

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) dedicated to developing algorithms capable of understanding, interpreting, and generating human language in a natural way [Caseli and Nunes, 2024]. For this purpose, collections of texts are gathered to be processed by computers. Such a collection of linguistic data is referred to as a **corpus**.

Tokenization of words is part of the preprocessing step in linguistic analysis within NLP. It involves the process of dividing a text into words or smaller units known as *tokens*.

Part-of-Speech Tagging, or simply **POS-Tagging**, is a technique that assigns labels, known as **tags**, to each word

in a text, indicating its grammatical class.

In the context of NLP, **regular expressions** are algebraic notations commonly used to find specific patterns in texts [Jurafsky and Martin, 2023]. On the other hand, a **dependency tree** is a graphical representation of the grammatical structure of a sentence, highlighting the syntactic relationships between words.

NLP techniques are naturally employed in automated analyses of large volumes of data to identify language patterns, literary, social, and cultural phenomena. This type of analysis, which focuses on observing broad trends and structural patterns across vast quantities of texts – instead of meticulously analyzing a single source of information –, is known as **distant reading** [Moretti, 2008].

Throughout this text, we use the term **predicative** to refer to the part of a sentence or clause that provides information about the subject. In the context of this work, predicatives often consist of adjectives, nouns, and verbs in the participle form. “*Ela está cansada.*” (She is tired.), “*Maria é professora.*” (She is a teacher.) and “*Ela finalmente foi homenageada.*” (She was finally honored.) are examples of predicative sentences.

3 Related Work

The analysis of corpora with the aid of NLP is crucial for detecting and mitigating biases in textual data. The work of Freitas and Martins [2023] combines this method with distant reading to characterize male and female characters in literary texts. The authors explore a corpus of Brazilian literature using lexico-syntactic patterns for search and semantic classifications, observing predication and quantitative organization of occurrences. They distributed the predicatives by gender and categories, referred to as axes (appearance, character, emotion, and social role). The current work resembles theirs as it follows the same search method for bias characterization. Our results corroborate those obtained by Freitas and Martins [2023], even while using a corpus with significant differences in domain, language, and time.

Our conclusions relating women to physical appearance also align with those presented by Taso *et al.* [2023a] in their analysis of word embeddings trained on data primarily obtained from Brazilian news portals. Still considering word embeddings, Taso *et al.* [2023b] demonstrate the presence of occupational sexism in the Brazilian market, as reflected in the analyzed model. Women are often employed in roles involving aesthetics or caregiving. Preliminary searches in the sentence patterns we found show that occupational sexism is also present in song lyrics. Confirming this result contributes not only to social studies but also to validating the bias identification metric in word embeddings utilized by Taso *et al.* [2023a] and Taso *et al.* [2023b].

Salles and Pappa [2021] analyzed gender bias in Portuguese Wikipedia biographies, comparing the representation of men and women across two dimensions: metadata and language. The research revealed significant differences in how women are portrayed compared to men. Only 16% of Wikipedia biographies are about women. Furthermore, the women described in the pages tend to be more notable

on average than men, suggesting that female figures must achieve outstanding accomplishments to be featured in the encyclopedia. In the analysis of the words most associated with each gender, men are linked to sports, whereas women are associated with the arts. Traditionally female-dominated professions, such as nursing, have little representation on Wikipedia. The authors highlight that women's romantic relationships are more frequently discussed. Our conclusions are not entirely related to the results obtained by Salles and Pappa [2021], but they point in the same direction regarding the subordinate treatment of women in society, which is perpetuated in digital media.

3.1 Sexism in Songs

Through interviews, Schlösser and Fantin [2022] identified that Brazilian songs often contain explicit sexual content that reinforces the objectification of women and perpetuates gender stereotypes. Most of the women interviewed expressed dissatisfaction with how the female figure is portrayed in these song compositions, describing them as expressions of a patriarchal and sexist culture. They reported feeling disrespected and violated when listening to songs that sexualize and devalue women, particularly in genres such as Funk and Forró, which were linked to the explicit eroticization of the female body. The interviewees noted that such sexual content reinforces female submission and legitimizes abusive and violent male behaviors, contributing to the normalization of gender-based violence in interpersonal relationships. Furthermore, the study highlights that, although some women do not find these lyrics offensive, the majority of respondents associate such content with the perpetuation of gender prejudices, underscoring the need to continue fighting for women's rights and respect in society.

From a psychological and social perspective, the study by Assis [2014] reveals how Brazilian musical works influence and reinforce emotional and behavioral dynamics associated with gender violence. The author emphasizes that the social representations of love in these song lyrics often mix positive affections with aggression, perpetuating the notion that conflict and violence are natural components of romantic relationships. Assis argues that this emotional ambivalence, which intertwines "love and hate" or "affection and violence", contributes to the normalization of gender violence. Additionally, the study demonstrates that these representations reinforce unequal gender roles, particularly by emphasizing female submission and male dominance. The romanticization of violence in songs, according to the author, legitimizes aggressive and controlling behaviors, sustaining a patriarchal culture that normalizes violence against women as part of romantic relationships.

Firmino *et al.* [2024] used NLP techniques to identify gender bias in a corpus of Brazilian lyrics. The study revealed that women are often described with adjectives related to physical appearance, such as *bonita*, *linda* e *gostosa* (in English, pretty, beautiful, and sexy) while men are primarily portrayed with adjectives highlighting traits of character and abilities, such as *feliz*, *forte* e *capaz* (happy, strong, and capable). These results highlight a stereotypical construction of gender in the considered musical art, where women are

objectified and valued for their appearance, while men are represented by their personal and social qualities.

In Betti *et al.* [2023], an automatic analysis of 377,808 English song lyrics focused on expressions of sexism and the measurement of gender bias over five decades (1960–2010). The results showed that the lyrics of male solo artists progressively became more sexist over the years, while this trend was less noticeable in other categories of artists. Additionally, linguistic biases revealed that the lyrics of male artists contained more gender biases than those of female artists, especially in relation to stereotypes portraying men as strong and success-oriented, and women as more family-focused.

In the work of Huang [2022], over 237,000 English songs from various genres between the 1970s and 2010s were analyzed using word representations. Techniques of dynamic word embeddings, trained with BERT, were applied to identify changes in the meanings and connotations of terms over time. The study highlighted the diversification of the term *bitch*, with its usage varying between positive and negative connotations. Artists from different genres began to use the term for various purposes, reflecting both acceptance and contestation of patriarchal norms. Among the findings, a clear bias was identified in the usage by male artists, who frequently used *bitch* pejoratively to describe women, while female artists reappropriated the term to signify female empowerment.

No similar studies to Betti *et al.* [2023] and Huang [2022] were found in the Brazilian context.

Although previous studies have explored gender bias in literature, digital media, and song lyrics, our work advances this field in several ways. Similar to Firmino *et al.* [2024], who identified gender biases in Brazilian music using NLP techniques, our study expands their findings by incorporating a broader corpus, a comprehensive categorization of predicatives, and an analysis of representations of professions in song lyrics.

Furthermore, while Betti *et al.* [2023] and Huang [2022] examined sexism in English-language music over time, no equivalent large-scale study has been conducted in the Brazilian context. Our work fills this gap by applying a methodology to Portuguese lyrics, identifying historical patterns of gender stereotypes. By integrating distant reading, machine learning classification, and profession analysis, we provide a more comprehensive approach to quantifying sexism in Brazilian music.

4 Research Resources

The analyses described in this study are based on a large corpus of song lyrics and the sentences within it, which qualify men and women. This section outlines the methods used to create this corpus and identify its predicative sentences.

4.1 Corpus and Data Collection

For the development of this research, we employed four corpora of Brazilian song lyrics – previously available on the Internet – and integrated them into a corpus compiled exclusively for this study in order to consider more recent widely

successful musical works. The corpora obtained from third parties are: *Song Lyrics from 79 Musical Genres*¹, *Bossa Nova Lyrics*², *Funk Carioca Dataset*³ e o *Corpus Vagalume* [de Araújo Lima et al., 2020].

The corpus developed by the authors consists of the 1,627 most played Brazilian songs since 1947, according to Spotify⁴. Spotify is the largest music streaming service in Brazil, offering both free and paid options. The 55 most popular playlists on the platform were selected to ensure a diverse and representative sample.

During the unification process of the datasets, the columns from the different corpora were standardized, and titles deduplicated. Gospel songs were excluded as they predominantly used the pronoun “He” and masculine adjectives to reference God. In the end, the resulting corpus contained a total of 146,612 songs, with the following columns: *Song Name*, *Artist*, *Music Genre*, and *Song Lyrics*.

In total, 74 music genres were presented in the corpus. Their frequencies are presented in Figure 1, with genres such as Sertanejo, MPB, and Forró among the most frequent.⁵

4.2 Preprocessing and Search Patterns

We chose the SpaCy library⁶ for preprocessing song lyrics because of its established methods for segmentation and syntactic analysis. A large pre-trained model for natural language processing in Portuguese was applied.

During preprocessing, each song lyric was divided into sentences for subsequent analysis of their content. The goal was to identify sentences that qualify men and women. To achieve this, search patterns were applied to the sentences.

The search patterns were inspired by Freitas and Martins [2023] and subsequently refined to align with the syntactic structures recognized by SpaCy. Initially, a random set of sentences was analyzed, and representative examples containing feminine and masculine predicatives were selected from the lyrics. These examples were then compiled into a dictionary, which served as the foundation for further analysis.

The sentences in the dictionary underwent a detailed analysis of their morphological and syntactic structures. This process involved identifying the grammatical classes of words and the syntactic relationships between them. Based on the initial analysis of the dictionary’s sentences, our search patterns were validated and expanded whenever necessary.

Regular Expressions (REs) were employed to identify descriptive sentences. REs enable more precise searches by using specific patterns that combine pronouns, subjects, verbs, adverbs, adjectives, determiners, and nouns to capture various grammatical structures within sentences. In total, 120

search patterns were developed, which can be verified in our source code (see the Declarations section). A summary of these patterns is presented using a high-level syntax in Appendix A. Two examples of these patterns include:

Subject/Proper Name/You + Auxiliary + [Adv] + Adj/Noun
Subject/Proper Name/You + Punct + [Det] + Noun

The listed pattern would return sentences such as:

Você é/está bonita. (You are/look beautiful.)
Ela é/está muito/pouco velha. (She is/looks very/little old.)
Mariana, a puta/vaca. (Mariana, the whore/bitch.)
Ela é professora. (She is a teacher.)

We obtained 15,223 sentences with female predicatives and 15,081 sentences with male predicatives.

As an illustrative example, we analyzed the song “G G” by Luiz Caldas, which appeared 29 times in the search patterns for male, highlighting its linguistic relevance within the analyzed criteria. Its high recurrence made it a representative case in this study. Below, we present a snippet of the song’s lyrics, with the predicative parts identified by our REs underlined.

Gil é genial, Gilberto é doutor
Gil é jovial, Gilberto é senhor
Gil é carnaval, Gilberto é rigor
Gil é contralto, Gilberto é tenor

Following this, the classic sertanejo lyric “Panela Velha”, performed by Sérgio Reis, appeared 12 times in corresponding search patterns for female. Below, we can observe a snippet of the song’s lyrics.

Tô de namoro com uma moça solteirona,
a bonitona quer ser a minha patroa, os
meus parentes já estão me criticando estão falando
que ela é muito coroa, ela é madura já
tem mais de trinta anos

5 Research Methods

After selecting the corpus, the methodology followed a structured sequence of steps to meet the proposed objectives. This sequence, as well as the research questions addressed at each step, are detailed in the diagram illustrated in Figure 2.

5.1 Categorization of Predicatives

Following the methodology used by Freitas and Martins [2023] for the analysis of human characterization, the categorization of predications was carried out into five categories: *Appearance*, *Character*, *Emotion*, *Social Role*, and *Other*. We created the last category for sentences that did not fit clearly into the other categories. These other terms were identified throughout the categorization process but were omitted from the analyses to avoid distortions in comparisons between the other categories.

The selection process was carried out to capture variability in language use. Initially, all sentences describing men and women retrieved during the search phase were randomly

¹<https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres>. Last access: 23 May 2025.

²<https://www.kaggle.com/datasets/mcarujo/bossa-nova-lyrics/data>. Last access: 23 May 2025.

³<https://www.kaggle.com/datasets/slshes/funk-carioca-dataset>. Last access: 23 May 2025.

⁴<https://open.spotify.com/>. Last access: 23 May 2025.

⁵One can notice that some categories could be merged into broader groups, but for the sake of accuracy, we maintained the original information as provided in the music repositories.

⁶<https://spacy.io/>. Last access: 23 May 2025.

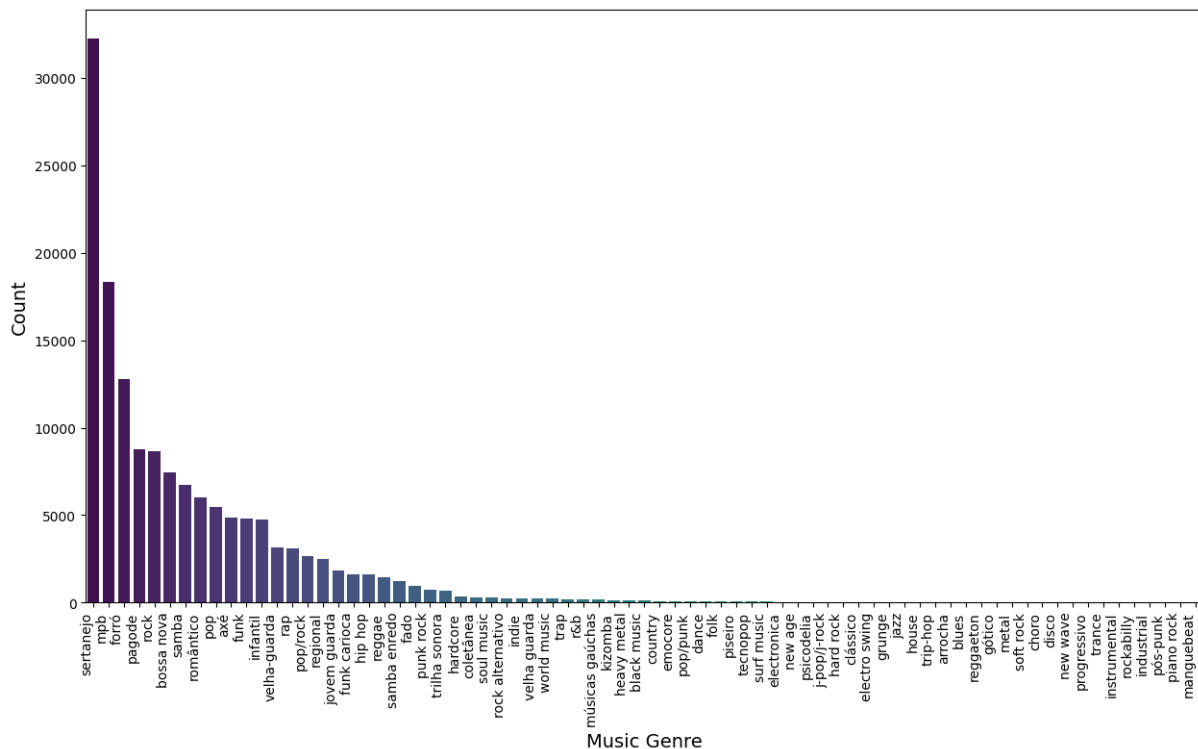


Figure 1. Distribution of the music genres presented in the corpus.

ordered. Sentences that had been erroneously included in the corpus due to syntactic analysis errors, as well as those containing unknown words, were excluded. Valid sentences were then manually annotated until a set of 1,000 predicate sentences was obtained for each gender, ensuring a balanced representation of labeled categories within each set.

The annotation was carried out by nine female volunteers, including seven undergraduate Psychology students from a public university, aged between 18 and 25, and two of the authors, aged between 23 and 43, both with experience in the fields of Information Technology. Before the annotation process began, an explanatory meeting was held with the annotators to clarify the categorization criteria and ensure a shared understanding among participants.

The Fleiss' Kappa agreement index among the seven annotators was 0.3395 for the feminine sentences and 0.3535 for the masculine sentences. Discrepancies were resolved based on the majority classification; in cases of a tie, one of the authors manually determined the final annotation.

The relatively low agreement rate may be attributed to the subjective nature of the categories analyzed, such as *character* and *emotion*, which require individual interpretation and contextual sensitivity. Despite the preliminary meeting to align categorization criteria, differences in the annotators' personal perceptions may have contributed to divergent interpretations of the same content. Additionally, the linguistic variety and semantic nuances present in the annotated texts may have hindered convergence in annotation decisions.

For comparison purposes, in the final category obtained from the joint annotation by the seven Psychology students and the annotation performed by the two authors – who jointly analyzed the 1,000 sentences – the Fleiss' Kappa agreement index was 0.5274 for the feminine sentences and

0.5879 for the masculine sentences.

Using the labeled predicatives, a machine learning model was trained with the objective of automating the classification of the remaining data. The BERTimbau Large model was adopted, a version of BERT pre-trained with data in Portuguese, especially suited for text classification tasks in this language. The choice of this model was motivated by its ability to capture linguistic and contextual nuances in Portuguese, making it particularly appropriate for the analysis of Brazilian song lyrics.

The model was loaded from the Transformers library, with a classification layer added to the end of the standard architecture. To transform the sentences into input vectors compatible with the model, the tokenizer corresponding to BERTimbau was used, applying truncation and padding so that all sequences had a fixed length of 128 tokens. The original dataset was initially filtered to remove instances with missing labels and, subsequently, categorical labels were converted into numerical values using label encoding. Given that the dataset showed class imbalance, the oversampling technique was applied using the *RandomOverSampler* algorithm, aiming to ensure a balanced distribution among the categories and avoid bias during training.

The architecture was trained using the AdamW optimizer, with a learning rate set to $1e-5$. To ensure the robustness and reliability of the results, stratified 10-fold cross-validation was employed, preserving class proportions in each subdivision. In each iteration of the validation process, the model was trained for five epochs and evaluated based on standard classification metrics, such as accuracy, precision, recall, and F1-score. At the end of each fold, the results were recorded and later aggregated for the calculation of overall averages. Training and evaluation were carried out using

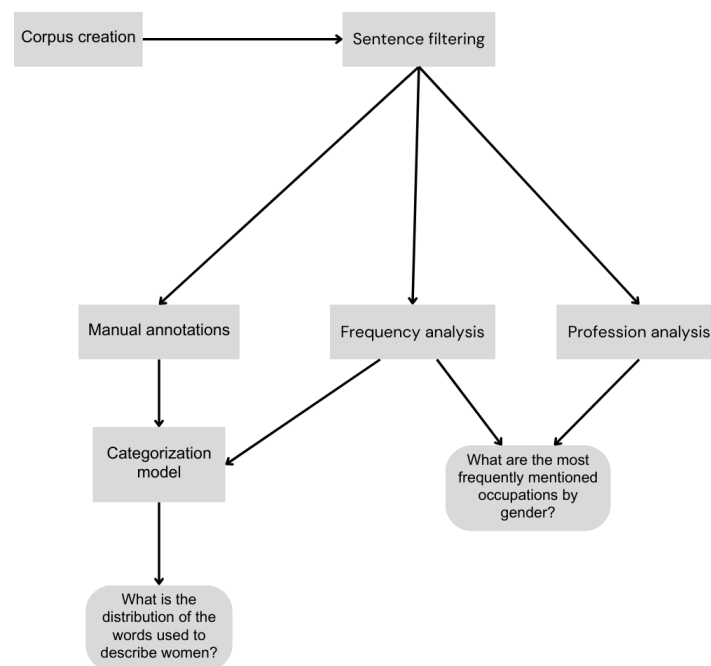


Figure 2. Methodological flow and research questions

GPU resources, whenever available, which provided greater efficiency during execution.

In the evaluation phase, four main metrics were computed: (i) Accuracy, which measures the proportion of correct predictions; (ii) Precision, which assesses the proportion of true positives among all positive predictions; (iii) Recall, which measures the model’s ability to correctly identify all positive instances; and (iv) *F1-Score*, which is the harmonic mean between precision and recall, offering a balanced view of overall performance.

These metrics were calculated using the `classification_report()` function from the *sklearn* library⁷, with class-weighted averaging, ensuring a fair evaluation even in scenarios with imbalanced classes.

The final performance of the model showed an accuracy of 92%, a weighted average precision of 91%, a weighted average recall of 92%, and a weighted average F1-score of 91%. When analyzing the results by class, the *Appearance* category achieved 92% precision, 93% recall, and 93% F1-score; *Character* reached 89% in all three metrics; *Emotion* obtained 92% precision, 96% recall, and 94% F1-score; *Other* scored 89% precision, 84% recall, and 86% F1-score; finally, the *Social Role* category recorded 95% precision, 96% recall, and 95% F1-score. These results demonstrate the model’s effectiveness in correctly classifying sentences into their respective semantic categories, highlighting its predictive capability and consistent performance across different classes.

Finally, the fine-tuned model was used to perform inference on a new set of previously tokenized sentences, and the predicted classes were then converted back to their original categories. The final results were saved and exported for statistical analysis and visualization. All source code used in

this process—including preprocessing, training, evaluation, and inference—was developed in Python and is available for inspection and reproduction⁸.

5.2 Analysis of Professions

In this analysis, we investigate the relationship between professional representation and gender. We compiled a list of professions based on the Brazilian Classification of Occupations (CBO-Domiciliar)⁹. Our goal was to understand how professions are attributed to men and women in song lyrics, looking for gender patterns in occupational associations through regular expressions and dependency trees.

To structure the analysis, professions were organized into three categories: those inflected in the feminine, those inflected in the masculine, and those common to both genders (where the gender was inconclusive without additional grammatical or semantic support). In the latter case, the gender distinction occurs through the placement of masculine or feminine determiners (pronouns, articles), such as *o presidente* or *a presidente*.

In the searches for professions inflected in the masculine, we used terms like *doutor* and *professor*, while for the feminine, terms like *doutora* and *professora* were used. Professions common to both genders were included only when the gender of the subject could be clearly identified, either as masculine or feminine, based on the morphology of the involved subject.

When the subject’s gender was not identified, as in “Você é piloto” (“You are a pilot”), we conducted searches that ex-

⁸See the Declarations section.

⁹<https://concla.ibge.gov.br/estrutura/ocupacao-estrutura>. Last access: 23 May 2025.

⁷<https://scikit-learn.org/>. Last access: 23 May 2025.

amined the profession along with its masculine or feminine inflection. To ensure greater accuracy in the results, duplicate sentences within the same song were removed.

6 Results

The results are presented as answers to each research question in this study.

6.1 *What is the distribution of the words used to describe women?*

The analysis of the distribution of predicatives found in the songs is conducted in three parts. First, we present the 10 most frequent characterizations for each genre, in order to highlight the contents of the most popular predicatives. Then, we extrapolate this quantity to 1,000 predicative sentences and use manual annotation to classify the terms into one of the 5 categories considered. Finally, we use the annotated predicative sentences to train a model to automatically categorize all descriptions found in the lyrics as referring to women and men.

The classification of the 10 most frequently used predicates for each gender, as well as their occurrence count, are presented in Table 1. It is observed that terms related to appearance, such as *bonita* (beautiful), *linda* (pretty), *novinha* (adjective often used to refer to young girls), *gostosa* (hot), and *feia* (ugly), are frequently attributed to women. This result suggests a focus on female aesthetics as the main attribute and is in line with the findings obtained by Firmino *et al.* [2024]. The constant emphasis on female appearance contributes to the perpetuation of beauty stereotypes that, although often presented as compliments — such as referring to a woman as *bonita* — restrict the role of women in society. As discussed by Wolf [1991], the excessive emphasis on female beauty acts as a tool of social control, limiting women’s development and full participation in various spheres. Such practice devalues other dimensions of their identity and competencies, particularly in areas such as the professional sphere, where intellectual attributes and skills are often overshadowed by aesthetic criteria.

Table 1. Ranking of the ten most frequent adjectives found in the predicative sentences.

Female		Male	
Adjective	Occurrence	Adjective	Occurrence
bonita	852	feliz	335
linda	794	bom	299
louca	263	pobre	189
novinha	207	forte	158
gostosa	183	capaz	141
boa	159	triste	124
feia	153	louco	114
feliz	145	amor	114
dona	123	diferente	114
mãe	110	casado	95
Total:	2989	Total:	1683

For men, on the other hand, the most common adjectives

are associated with emotional and character traits. Terms such as *feliz* (happy), *bom* (good), and *capaz* (capable) point to a representation that values personality and male competence. *Triste* (sad) and *louco* (crazy) are bad qualifications used to describe men inside these traits. This tendency reinforces the idea that, while descriptions of women focus on external attributes, those of men emphasize subjective qualities and capabilities. It is also important to highlight the frequency of the most common predicatives: 2,989 in the case of women and 1,683 in the case of men. In other words, women are much more “adjectivized” than men.

Figures 3 and 4 display the word clouds of all feminine and masculine predicatives identified in the corpus, respectively. Among the 3,419 feminine predicatives, 1,910 occur only once in the dataset. This phenomenon reveals the need to understand all the characteristics used to describe women throughout the lyrics.



Figure 3. Cloud of female predicatives



Figure 4. Cloud of male predicatives

Figure 5 presents the frequencies of the categories for the 1,000 annotated predicative sentences, where the categories were, during the selection of sentences, assumed to be approximately balanced. The distribution in the figure allows us to visualize the predominance of each category in the characterization of men and women.

It can be seen that the categories *Character* and *Social Role* were the most frequent for both genders, although with different proportions. The *Social Role* category, for example, presented a significantly higher frequency in the representation of men (33.7%) than of women (28.7%).

On the other hand, the category *Character* was slightly more attributed to women (30.2%) than to men (29.8%), although the difference is small.

The category *Appearance*, although less frequent, presents similar proportions between genders, with a slight predominance in feminine sentences (23.6% versus 22.5% in masculine sentences).

The category *Emotion*, on the other hand, is the one that shows the greatest disparity between genders, being more re-

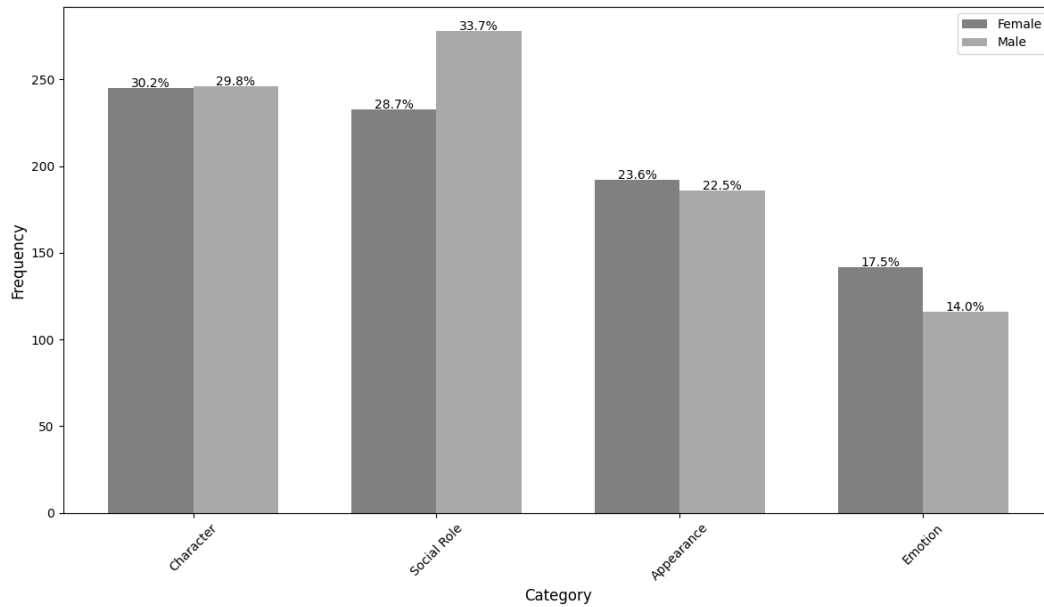


Figure 5. Frequency of categories for the 1000 annotated predicative sentences.

current in feminine sentences (17.5%) compared to masculine sentences (14.0%).

To verify the general pattern of characterization of women and men, we used a machine learning model developed specifically for this purpose, trained to categorize all predicative phrases present in the analyzed corpus. The results can be seen in Figure 6.

The predominance of the *Appearance* category in sentences associated with the female gender is striking, accounting for 27.43% of the occurrences. This figure is more than double the proportion observed for the male gender, where the same category represents only 12.25%. Such a discrepancy highlights a significant emphasis on the physical appearance of female characters.

On the other hand, there is a predominance of the *Social Role* category in sentences associated with the male gender, with 24.12%, in contrast to 16.75% for the female gender. This pattern suggests a stronger association between men and social positions, aligning with previous studies.

The *Character* category shows a fairly balanced distribution between genders, with 21.17% for females and 20.99% for males.

Finally, the *Emotion* category constitutes the smallest proportion in both distributions, with 6.12% for females and 6.55% for males.

Overall, the data reveal distinct representational trends between genders. While women are more frequently associated with physical traits, men appear more linked to social roles. These differences likely reflect pervasive cultural norms. They underscore how such associations shape social perceptions of femininity and masculinity.

6.2 What are the most mentioned occupations by gender?

A total of 281 sentences referencing professions in the masculine form and 131 in the feminine form were identified.

After the search, the professions were ranked according to their frequency, with the results presented in Table 2.

Table 2. Ranking of the ten most frequent occupations by gender.

Female		Male	
Occupation	Frequency	Occupation	Frequency
Atriz	32	Doutor	31
Professora	10	Presidente	28
Modelo	10	Capitão	14
Chefe	6	Artista	13
Babá	5	Professor	13
Prostituta	4	Cantor	12
Lavadeira	4	Caçador	12
Enfermeira	4	Delegado	11
Costureira	3	Locutor	11
Artista	3	Soldado	11
Total	131	Total	281

The data reveal that the main occupations associated with the male gender include *doutor* (doctor), *presidente* (president), and *capitão* (captain). For the female gender, the most common professions are *atriz* (actress), *professora* (teacher), and *modelo* (model). This difference suggests a division of professional roles between genders, reflecting cultural and social expectations present in song lyrics, while also highlighting the impact of these perceptions on career choices and opportunities for men and women.

The observed division also reflects the perpetuation of gender stereotypes in professional occupations. While men are predominantly associated with positions of authority, women are represented in professions related to entertainment, teaching, and care. This contrast points to a continuation of traditional gender representations, in which men are seen as leaders or figures of power, while women are associated with roles linked to appearance and caregiving.

IBGE [2024] provides data that highlight this occupational segregation. In the “transportation, storage, and mail” sector, for instance, 80.5% of managerial positions are held by men,

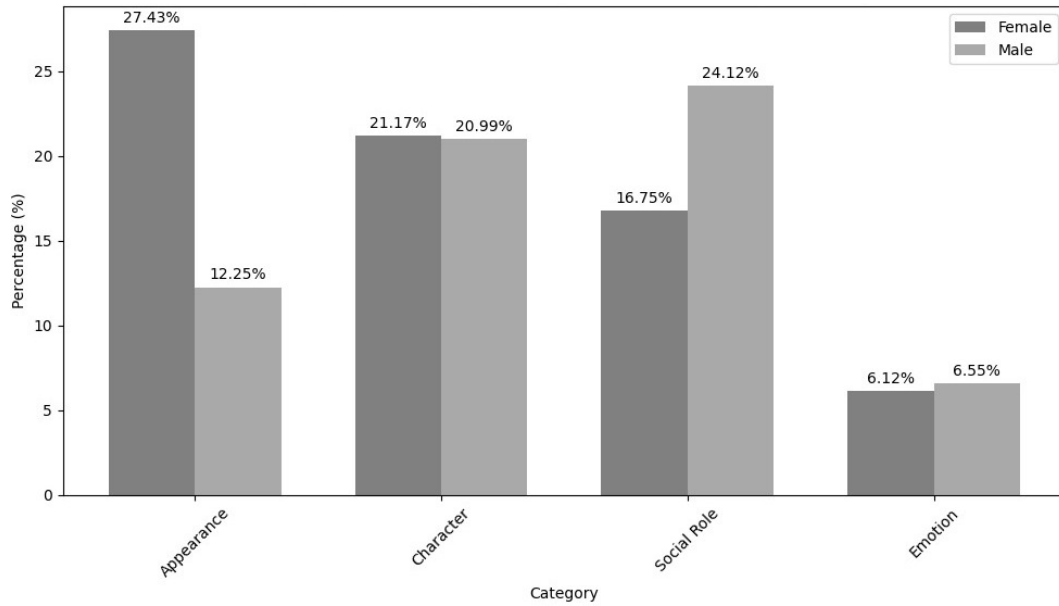


Figure 6. Frequency of categories for all the predicatives in the corpus

while women represent only 19.5%. Similarly, in the “construction” sector, 85.9% of managerial positions are male, compared to just 14.1% held by women. On the other hand, in sectors traditionally associated with women, such as “human health and social services”, 63.5% of managerial positions are held by women, while men occupy 36.5% of these roles.

These data corroborate gender stereotypes, where men dominate technical and infrastructure management areas, while women are more present in sectors related to care and education.

Regarding the occupation of power and managerial positions in general, men occupy a significantly higher percentage than women (75%). This scenario can be interpreted as a reflection of deeply rooted historical biases, where media and cultural representations reinforce social expectations and behaviors tied to traditional gender roles.

7 Stereotypes in Society and Algorithmic Discrimination

Artificial Intelligence (AI) has become expansive and influential in modern society. However, as technological advances progress, cases where algorithms reproduce or amplify discrimination become increasingly evident. This highlights the importance of observing and questioning the ethical and social implications related to AI, as well as the need to understand how stereotypes manifest in models to develop fairer computational systems.

According to Barocas and Selbst [2016], the effectiveness of data usage by computational systems is inherently related to the quality of the data analyzed. If this data contains biased behavior, the models trained from it will learn from the poor example these decisions set. Thus, an algorithmic decision is highly dependent on the data that feeds it. In this sense, bias can manifest when analyses are conducted, and models are trained on datasets reflecting existing societal inequali-

ties. This can affect specific groups unequally, especially concerning characteristics such as gender, race, and age.

Discussing NLP, Caseli and Nunes [2024] emphasize that processing textual data for interaction with computers does not simply involve organizing a set of characters or structuring formal orthographic representation. Text processing also involves engaging with linguistic resources that represent human experience. When this experience carries social biases, decisions based on it are also susceptible to often discriminatory tendencies.

Moss-Racusin *et al.* [2012] define gender bias as a preference or prejudice towards one gender at the expense of another. Sun *et al.* [2019] categorize gender bias in NLP tasks into four forms: “denigration”, referring to culturally or historically derogatory terms; “stereotyping”, which reinforces existing social stereotypes; “recognition”, involving algorithmic inaccuracies in recognition tasks; and “underrepresentation”, describing the low representation of a specific group.

The replication of discriminatory tendencies, such as the underrepresentation or stereotyping of women, can further exacerbate the gender inequality scenario. The absence of rigorous monitoring and precise adjustments in these models can lead to even more lost opportunities for women, reinforcing cycles of exclusion and perpetuating inequality.

The presence of gender biases in large language models (LLMs), such as GPT-3, is demonstrated by Lucy and Baman [2021], where female characters in stories were often associated with family and appearance, while male characters were linked to power and action. Additionally, Wan *et al.* [2023] highlighted a strong disparity in the adjectives used to describe men and women in automatically generated recommendation letters. While letters for men often described leadership and competence qualities, those for women tended to highlight warmer, more supportive characteristics. In the Brazilian context, Soares *et al.* [2023] investigated automatic translation performed by GPT-3.5 Turbo, observing gender bias in translating professions from English to Portuguese. In sentences ambiguous in terms of gender, the Portuguese

translation tended to assign masculine gender to professions like *physician*, while professions like *nurse* were translated in the feminine.

“Desvelar – Justiça Racial na Inteligência Artificial e TICs” is a project, which maintains a platform gathering articles and news on algorithmic discrimination [Silva, 2023]. It presents an interactive timeline that allows the visualization of the impact of incidents involving AI in different social contexts. Examples include: streaming platform recommendations where male artists receive more than 55% of the recommendations compared to only 20% for women; an image generator that includes weapons in prompts related to black women in favelas; a system for assessing domestic violence risk that denied protection to a woman who was subsequently murdered, with 50 similar cases identified; and Google’s keyboard suggesting sexual terms for the word *neguinha* (little black girl).

Using lyrics in machine learning models can present significant risks without careful and diverse data curation. In songs using derogatory descriptions to portray women, algorithms can learn and perpetuate such stereotypes, reflecting and amplifying these biased representations. Thus, the repeated presence of stereotyped representations can propagate these conceptions in subsequent tasks, influencing everything from content classification to the automatic generation of texts and images.

To mitigate these effects, it is crucial that datasets used in machine learning models include a representative diversity of popular culture while being carefully examined to prevent specific patterns from perpetuating prejudice. Such meticulous selection aids in developing models that respect cultural plurality and simultaneously remain attentive to the nuances of social representations. Ultimately, training with culturally informed yet balanced datasets allows technology to reflect a more ethical and inclusive view, avoiding the inadvertent propagation of stereotypes and fostering a more equitable interpretation of content derived from popular culture.

8 Future Work

Although this study has advanced the understanding of the characterization of women in Brazilian music, there are still many issues to be explored. The most important is to analyze each occurrence of a predicative within its context and classify it as a positive, negative, or neutral sentiment. This analysis would allow for a closer and more accurate reading of the songs.

It was not possible to measure the evolution of female stereotyping over time because the data used for analysis does not provide enough information to perform an analysis that considers the release date of the songs. It is necessary to further enrich the corpus in order to enable this type of research.

It would also be relevant to investigate the interactions between record labels, composers, and performers, analyzing how these relationships influence the choice of themes and the way gender roles are represented in the songs. Expanding the scope to include other forms of media would allow us to verify whether the patterns of objectification and gender

stereotypes are also present in other domains, such as cinema, television, and social networks. These new analyses could provide a broader understanding of gender representations in popular culture and their influence on the formation of stereotypes.

Additionally, based on the verbs most frequently used by female characters in the songs, it is possible to identify the most common actions for each gender. The hypothesis is that the verbs used in the male context suggest discursive or intellectual roles, while verbs used in the female context reveal passivity, focusing on expressions of desire or aspiration. There is also the prospect of studying the use of emotional control metrics to qualify the actions described in the songs.

9 Conclusions

The objective of this study was to analyze the presence of gender biases in Brazilian music. To achieve this, a corpus comprising over one hundred thousand national song lyrics was constructed. Using a combination of distant reading supported by NLP techniques and manual annotations, the research revealed that women are predominantly characterized by their appearance, while men are more frequently described in terms of social roles. We employed a trained classification model to expand the manually conducted analysis, allowing for the categorization of a significantly larger volume of predicates and providing more robust evidence of recurring patterns of gender representation in Brazilian song lyrics.

The limitation of women’s social role in society was also evident in the music. The analysis of professions and the grouping of terms in Brazilian song lyrics revealed significant patterns that reinforce gender stereotypes. Occupations associated with power and authority are predominantly masculine, while occupations linked to care, teaching, and entertainment are mostly attributed to the feminine gender. This demonstrates a division of professional roles and leadership and management perspectives, reflecting cultural issues about the boundaries of gender roles pointed out by IBGE.

Thus, we conclude that historical patterns of sexism in the treatment of women in songs have been identified. Our results contribute and align with those obtained by Freitas and Martins [2023], even using a corpus with different domain, language, and seasonality. Our conclusions also corroborate Firmino *et al.* [2024], confirming and reaffirming the presence of stereotypical representations of women in society through music. However, this study advances by integrating professional aspects, conducting a detailed analysis of the distribution of adjectives used in sentences to describe women, and employing the BERTimbau model for a more robust and precise categorization of sentences. Thus, we provide an expanded and updated perspective on gender biases in the Brazilian musical context.

This work presents limitations related to the NLP library used, whose algorithms may influence the quality of the lyrics analysis. This issue becomes even more pronounced due to the unconventional linguistic structures found in the songs, which may compromise the accuracy of information extraction. Additional challenges arise from the low

level of agreement observed during the annotation process. For future work, we intend to revise the categorization of predicatives using an incremental methodology, supported by clearer annotation guidelines for cases with low inter-annotator agreement.

Finally, muses or stereotypes? Women in Brazilian songs are predominantly portrayed through stereotypes that reinforce historical sexist patterns. Although the role of muse may seem like a celebration of female beauty and emotion, it often reduces women to idealized and superficial representations, neglecting their complexity and other characteristics, such as intelligence, strength, and capability. The stereotypes go further, consolidating narratives that restrict them to traditional and subordinate roles, perpetuating gender inequalities and cultural patterns that hinder a broader, fairer, and more empowering vision of women.

A Search for Descriptive Sentences

Subj/Name/Pron + Aux + Adj
 Subj/Name/Pron + Aux + Adv + Adj
 Subj/Name/Pron + Aux + Aux + Adj
 Subj/Name/Pron + Aux + Aux + Adv + Adj
 Subj/Name/Pron + Aux + Verb + Adj
 Subj/Name/Pron + Aux + Verb + Adv + Adj
 Subj/Name/Pron + Verb + Aux + Adj
 Subj/Name/Pron + Verb + Aux + Adv + Adj
 Subj/Name/Pron + Aux + Verb_Part
 Subj/Name/Pron + Aux + Adv + Verb_Part
 Subj/Name/Pron + Verb + Adj
 Subj/Name/Pron + Verb + Adv + Adj
 Subj/Name/Pron + Adj
 Subj/Name/Pron + Pron + Verb + Adj
 Subj/Name/Pron + Pron + Verb + Adj
 Subj/Name/Pron + Aux + Pron + Verb + Adj
 Subj/Name/Pron + Aux + Pron + Verb + Adv + Adj
 Subj/Name/Pron + Verb + Pron + Verb + Adj
 Subj/Name/Pron + Verb + Pron + Verb + Adv + Adj
 Subj/Name/Pron + Punct + Adj
 Subj/Name/Pron + Punct + Det + Adj
 Subj/Name/Pron + Aux + Noun
 Subj/Name/Pron + Aux + Adv + Noun
 Subj/Name/Pron + Aux + Aux + Noun
 Subj/Name/Pron + Aux + Aux + Adv + Noun
 Subj/Name/Pron + Aux + Verb + Noun
 Subj/Name/Pron + Aux + Verb + Adv + Noun
 Subj/Name/Pron + Verb + Aux + Noun
 Subj/Name/Pron + Verb + Aux + Adv + Noun
 Subj/Name/Pron + Verb + Noun
 Subj/Name/Pron + Verb + Adv + Noun
 Subj/Name/Pron + Noun
 Subj/Name/Pron + Pron + Verb + Noun
 Subj/Name/Pron + Pron + Verb + Adv + Noun
 Subj/Name/Pron + Aux + Pron + Verb + Noun
 Subj/Name/Pron + Aux + Pron + Verb + Adv + Noun
 Subj/Name/Pron + Verb + Pron + Verb + Noun
 Subj/Name/Pron + Verb + Pron + Verb + Adv + Noun
 Subj/Name/Pron + Punct + Noun
 Subj/Name/Pron + Punct + Det + Noun

Declarations

Acknowledgements

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

Authors' Contributions

Janaina N. S. Lopes contributed to the conceptualization, investigation, data curation, methodology, and writing of this work. Vitória

P. Firmino carried out part of the investigation, data curation, and methodology phases. She was solely responsible for the software development and its validation. Valéria Q. Reis supervised Janaina and Vitória, 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://github.com/firminovitoria/profissoes-categorias>.

References

- Assis, C. L. (2014). “Entre tapas e beijos”: representações sociais sobre a violência de gênero para adolescentes. 2:229–242. DOI: <https://doi.org/10.12957/psi.saber.soc.2013.8796>.
- Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3):671–732.
- Betti, L., Abrate, C., and Kaltenbrunner, A. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1). DOI: <https://doi.org/10.1140/epjds/s13688-023-00384-8>.
- Caseli, H. M. and Nunes, M. G. V. (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition.
- de Araújo Lima, R., de Sousa, R. C. C., Lopes, H., and Barbosa, S. D. J. (2020). Brazilian Lyrics-Based Music Genre Classification Using a BLSTM Network. In Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., and Zurada, J. M., editors, *Artificial Intelligence and Soft Computing*, pages 525–534, Cham. Springer International Publishing.
- Duprat, R. (2008). Fruição, sedução e produção: o papel da mulher na música. *Música em Perspectiva*, 1(1). DOI: <https://doi.org/10.5380/mp.v1i1.11717>.
- Feijó, M. and Macedo, R. M. S. d. (2013). Gênero, cultura e rede social: a construção social da desigualdade de gênero por meio da linguagem. *Nova Perspectiva Sistêmica*, 21(44):21–34.
- 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, Brasília, DF, Brasil. SBC. DOI: <https://doi.org/10.5753/wics.2024>.
- 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*, 20:9118–9138. DOI: <https://doi.org/10.5007/1984-8412.2023.e86749>.
- Huang, J. (2022). Changing semantics of gendered insults in music lyrics. (THESIS). DOI: <https://doi.org/10.6082/uchicago.3762>.

- 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: 23 May 2025.
- Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing*, volume 3.
- Kong, L. (1995). *Popular Music in Geographical Analyses*. Progress in Human Geography, 15th edition.
- Lima, C. D. and Sanches, N. P. L. (2009). A construção do eu feminino na música popular brasileira. *Caderno Espaço Feminino*, 21:181–205.
- Lucy, L. and Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. In Akoury, N., Brahman, F., Chaturvedi, S., Clark, E., Iyyer, M., and Martin, L. J., editors, *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2021.nuse-1.5>.
- Moretti, F. (2008). *A literatura vista de longe*. Editora Arquipélago.
- Moss-Racusin, C., Dovidio, Brescoll, V., Graham, M., and Handelsman, J. (2012). Science faculty’s subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109:16474–16479. DOI: <https://doi.org/10.1073/pnas.1211286109>.
- 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>.
- Schlösser, A. and Fantin, G. (2022). Percepção de mulheres acerca de conteúdos sexuais na música popular do Brasil. *Ayvu: Revista de Psicologia*, 9.
- Silva, T. (2023). Mapeamento de Danos e Discriminação Algorítmica. Available in <https://desvelar.org/casos-de-discriminacao-algoritmica/>. Last access: 23 May 2025.
- Soares, T., Gumiel, Y., Junqueira, R., Gomes, T., and Pagano, A. (2023). Viés de gênero na tradução automática do GPT-3.5 turbo: avaliando o par linguístico inglês-português. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 167–176, Porto Alegre, RS, Brasil. SBC. DOI: <https://doi.org/10.5753/stil.2023.234186>.
- 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>.
- Taso, F., Reis, V., and Martinez, F. (2023a). Discriminação Algorítmica de Gênero: Estudo de Caso e Análise no Contexto Brasileiro. In *Anais do IV Workshop sobre as Implicações da Computação na Sociedade*, pages 13–25, Porto Alegre, RS, Brasil. SBC. 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>.
- Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., and Peng, N. (2023). “Kelly is a Warm Person, Joseph is a Role Model”: Gender Biases in LLM-Generated Reference Letters. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2023.findings-emnlp.243>.
- Wolf, N. (1991). *The Beauty Myth: How Images of Beauty Are Used Against Women*. HarperCollins.