NLP Pipeline for Gender Bias Detection in Portuguese Literature

Mariana O. Silva[1], **Mirella M. Moro**[1]

[1]Department of Computer Science
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{mariana.santos,mirella}@dcc.ufmg.br

Abstract. *We present a novel Natural Language Processing (NLP) pipeline designed to analyze gender bias in Portuguese literary works. Our pipeline comprises five processing steps, culminating in gender bias detection across different linguistic dimensions. We apply it to a corpus of Portuguese literary texts and evaluate its effectiveness in uncovering gender bias. Our findings reveal prevalent gender stereotypes in character descriptions, with female characters often associated with appearance and emotion, while male characters are depicted in terms of social status and personality traits. Furthermore, our analysis of physical traits stereotypes indicates a more equitable representation across genders in such a dimension.*

1. Introduction

Literature can be a rich source of information about human behavior and society, reflecting and shaping cultural norms and beliefs. Within the domain of literature, gender bias is a relevant issue that not only mirrors societal inequalities and stereotypes but also influences how gender is perceived and constructed [Lucy and Bamman 2021, Silva et al. 2023, Freitas and Santos 2023]. Therefore, understanding and addressing gender bias in literary works are crucial to promoting inclusivity, equity, and social change. However, manually identifying and analyzing cases of gender bias can be challenging, especially considering a large corpora.

In this context, computational methods in literary studies, particularly distant reading, have revolutionized how researchers approach literary texts' analysis [Jänicke et al. 2017]. Distant reading (i.e., characterized by applying computational techniques to large-scale textual corpora) enables researchers to uncover patterns and insights that traditional close reading methods may overlook. Natural Language Processing (NLP) plays an essential role in distant reading by providing computational tools and techniques to process and extract information from textual data efficiently and effectively, thus facilitating the identification of subtle patterns across extensive literary corpora.

In addition to these advancements, applying NLP to literary texts presents unique challenges compared to other text types. Literary texts often contain complex language, including figurative language, dialects, and archaic or poetic forms [Goldman and Lee 2014], which can be challenging for NLP models to interpret accurately. Additionally, literary texts may contain ambiguous or subjective content that requires contextual understanding and background knowledge to interpret correctly [Silva 2021, Freitas and Santos 2023].

While NLP has been employed in various literary studies, including character analysis, sentiment analysis, and topic modeling, its application in detecting gender bias within literary works has only recently attracted interest [Xu et al. 2019, Lucy and Bamman 2021, Silva 2021]. Scholars have used NLP techniques to explore the portrayal of gender in literature, identifying and analyzing gender stereotypes, language use, and representation of characters. However, much of this research exists for English and other languages, leaving a notable gap for exploring gender bias within Portuguese literature [Freitas and Santos 2023, Silva et al. 2024].

This paper addresses this research gap by presenting a novel NLP pipeline tailored to detect gender bias in Portuguese literary works. By leveraging domain-specific linguistic resources and computational techniques, our pipeline offers a systematic and efficient approach to identifying instances of gender bias within Portuguese texts. Through automated analysis, our pipeline aims to provide researchers and literary analysts with a scalable and efficient tool for gender bias assessment in Portuguese literature. The main contributions of our work are summarized as follows:

1. We introduce a novel NLP pipeline designed to identify and analyze instances of gender bias within Portuguese literary works.
2. Our pipeline is open-source, enabling accessibility for researchers interested in analyzing gender bias in literary works.[1]
3. We provide a corpus comprising over 1,200 literary works in Portuguese, spanning several centuries and authored by different writers.
4. We demonstrate the relevance of our pipeline in detecting gender bias by applying it to a diverse corpus of Portuguese literary works.

2. Related Work

Distant reading is an approach in literary studies that employs computational methods to analyze literary data. It is instrumental in uncovering patterns and insights that traditional close reading methods may overlook [Jänicke et al. 2017]. Typically, distant reading involves analyzing extensive collections of texts, requiring computational resources to handle large volumes of textual data efficiently. Such computational tools enable researchers to conduct in-depth analyses across vast corpora, facilitating text exploration [Gusmão et al. 2021, Real et al. 2021, Zahn et al. 2021].

Indeed, NLP has been applied across various literary issues, including character analysis [Labatut and Bost 2019, Silva et al. 2021], sentiment analysis [Maharjan et al. 2018], topic modeling [Chu et al. 2022], and even gender bias analysis [Lucy and Bamman 2021, Silva et al. 2023, Freitas and Santos 2023, Silva et al. 2024]. Researchers have employed computational methods to identify and analyze gender bias present in textual data, exploring themes such as stereotype portrayal [Silva 2021, Freitas and Santos 2023] and disparities in character representation [Casey et al. 2021, Kejriwal and Nagaraj 2024].

Despite the advancements in computational literary analysis, research specifically focusing on gender bias in Portuguese literature remains relatively limited [Santana et al. 2018]. For instance, [Silva 2021] explores the portrayal of male and female characters, considering predicates used in character descriptions and actions, and

[1] Available in: `https://github.com/marianaossilva/gender_pipeline`.

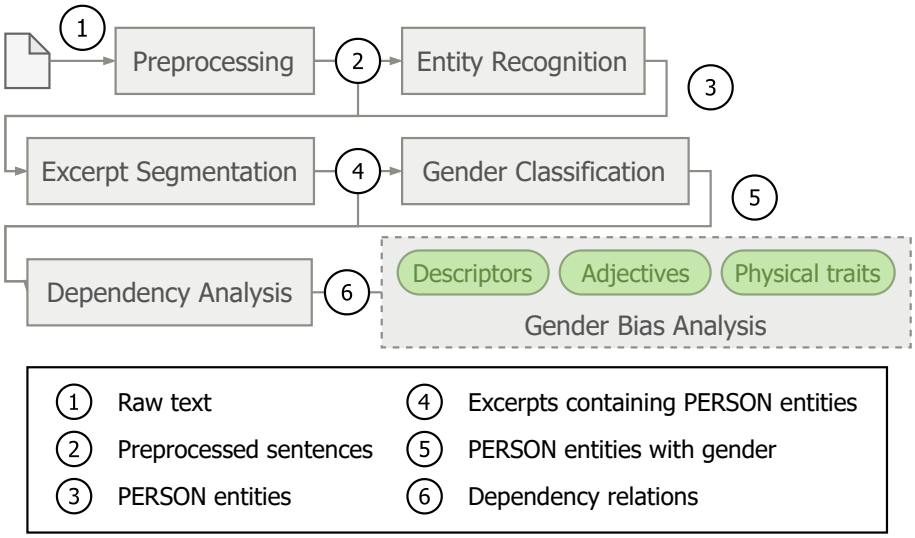

Figure 1. Overview of the proposed pipeline.

critically analyzing gender representations. Similarly, [Freitas and Santos 2023] investigate words describing human characters, classifying these words into social, emotional, character, and physical attributes. Their findings reveal a higher tendency to describe female characters by appearance than male characters, along with distinct patterns in using gender-specific attributes.

Then, focusing on the descriptions of male and female body parts, [Silva et al. 2023, Silva et al. 2024] present a quantitative analysis of gender representation within literature in Portuguese. The authors investigate a corpus of 34 literary works to quantify body part descriptions' frequency, specificity, and objectification and provide empirical evidence of gender portrayal. Their findings reveal specific differences in the frequency and choice of adjectives used for male and female body parts, shedding light on prevalent gender stereotypes in literary works.

While the aforementioned studies offer valuable insights into gender depiction within Portuguese literature, our work goes beyond presenting a novel NLP pipeline tailored for detecting gender bias in Portuguese literary works. By integrating domain-specific techniques and linguistic resources, our pipeline offers a comprehensive framework for analyzing gender bias, thereby enriching NLP research and literary studies.

3. Pipeline

We now describe our proposed NLP pipeline for gender bias detection in Portuguese literature. As illustrated in Figure 1, it consists of six interconnected steps, each designed to preprocess, identify, and analyze instances of gender bias within the text. Starting from raw text input, the pipeline systematically extracts meaningful insights regarding gender representation and bias over different dimensions, as detailed next.

3.1. Preprocessing

The pipeline receives raw text as input, which goes through preprocessing to ensure consistency and facilitate subsequent analysis. First, it addresses formatting irregularities

commonly found in PDF files, such as removing line breaks to ensure the text is continuous and cohesive. Next, extra spaces, special characters, email addresses, and website URLs are removed to focus solely on textual content, while essential punctuation marks and hyphens are retained for linguistic coherence.

The preprocessing step also removes noisy headers containing metadata unrelated to the core textual content. To do so, it uses regular expressions to match specific textual elements such as "Chapter 1" or "Chapter I" marking the beginning of the work's content. Any headers preceding this limit are then disregarded. However, not all texts contain explicit chapter divisions, and thus, alternative methods are required to identify and remove headers effectively. It employs heuristic approaches based on common header patterns and textual cues in such cases.

Finally, the text is segmented into individual sentences, and the preprocessing step filters out works with fewer than ten sentences. Such a threshold helps ensure that the analyzed texts are sufficiently substantive for meaningful analysis while removing shorter texts that may not provide adequate context for gender bias detection.

3.2. Entity Recognition

In this step, we employ a named entity recognition (NER) model to extract PERSON entities, which represent individuals mentioned in the text.[2] These entities are used in the subsequent steps of the pipeline, including gender classification and bias detection. Specifically, we used the LitBERT-CRF model [Silva and Moro 2024a] for NER, which is tailored for analyzing Portuguese literary texts. It is built upon the general-domain BERT-CRF model [Souza et al. 2019] and fine-tuned on the NER task using a literary annotated corpus in Portuguese [Silva and Moro 2024b], achieving an F1-Score of 78%.

3.3. Excerpt Segmentation

This pipeline step uses the sentence segmentation performed during preprocessing to segment the text into individual excerpts. Each excerpt encapsulates the sentence where each PERSON entity is mentioned. It begins by storing the sentence containing the entity and iteratively appends two additional sentences. This process ensures that each excerpt contains enough textual content to detect patterns and trends in portraying entities while minimizing the risk of analyzing fragmented or insufficiently contextual excerpts. Thus, such excerpts act as contextual units for subsequent analysis, facilitating a comprehensive investigation of the surrounding text associated with each identified entity

3.4. Gender Classification

After identifying PERSON entities and segmenting the texts, the gender classification step aims to assign genders to these entities based on contextual and linguistic cues. To do so, we integrate two distinct approaches previously evaluated in [Silva et al. 2024] (with F1-Scores over 90%). The first approach considers the *genderBR* package,[3] specifically designed to predict gender based on Brazilian first names. This package leverages data

[2]These individuals can be explicitly referred to by their proper names (e.g., "Capitu") or by anaphoric noun phrases referring to characters (e.g., "Bentinho's wife"). [Silva and Moro 2024b]

[3]*genderBR*: `https://github.com/meirelesff/genderBR`

from the IBGE's 2010 Census,[4] which provides information on the number of females and males associated with particular names in Brazil or a specific Brazilian state.

The *genderBR* package calculates the proportion of females who use a given name and classifies the name as either male or female based on a predefined threshold. For example, if the proportion of females using a name exceeds a threshold of 0.9, the name is classified as female. Conversely, if the proportion is less than or equal to 0.1, the name is classified as male. Names with proportions below both thresholds are classified as missing. Such an approach performs well for names that are clearly associated with a specific gender, as indicated by the census data. However, it may encounter limitations when handling non-traditional or ambiguous names that deviate from typical gender norms.

In cases where the *genderBR* package cannot provide a clear gender assignment or for entities with ambiguous gender markers, we employ a second context-based approach. This method analyzes descriptive text passages associated with PERSON entities to determine their gender. It involves performing dependency analysis on the entity's surrounding text, checking grammatical structures, semantic relationships, and contextual cues to infer gender more accurately. If most gender-bearing words associated with a character are marked with masculine gender morphology, the character is considered male. Conversely, if most gender-bearing words have feminine gender morphology, the character is classified as female.

3.5. Dependency Analysis

In this step, a dependency analysis is performed on the excerpt surrounding the identified PERSON entities to extract relevant linguistic dependencies and syntactic patterns. First, it focuses on extracting dependency relations such as "nsubj" (nominal subject), "amod" (adjectival modifier), "conj" (conjunct), and "appos" (apposition) related to the PERSON entities. These relations provide insights into the text's grammatical structure and semantic relationships, aiding in inferring gender-related information.

Additionally, the pipeline detects occurrences of body parts within the text excerpts, which are integral components of character descriptions and offer significant contextual clues for gender classification. To extract mentions of BPs associated with the identified PERSON entities, regular expressions are used, along with a manually compiled dictionary of body parts (BPs). The complete dictionary contains 55 BPs and 104 synonyms and is available in the pipeline's GitHub repository.[1]

3.6. Gender Bias Detection

The final step of the pipeline is a comprehensive analysis to detect gender bias within Portuguese literary works based on the attributes identified and returned from the previous steps. To quantitatively assess gender bias, we calculate a gender skewness measure (proposed in [Silva et al. 2024]), denoted as $S(x)$, for each analyzed attribute x:

$$S(x) = \frac{P(x|F) - P(x|M)}{P(x|F) + P(x|M)}, \tag{1}$$

where $P(x|F)$ is the average probability of attribute x occurring in excerpts with female PERSON entities, and $P(x|M)$ represents the average probability of attribute x occurring

[4] *Instituto Brasileiro de Geografia e Estatistica*'s 2010 Census: `https://censo2010.ibge.gov.br/`

in excerpts with male PERSON entities. From the calculated percentages, we can further analyze the gender skew of each attribute, $S(x)$, yielding significant insights:

$$S(x) = \begin{cases} \text{skewed to female entities} & \text{if } S(x) > 0 \\ \text{skewed to male entities} & \text{if } S(x) < 0 \\ \text{equally assigned to both genders} & \text{if } S(x) = 0 \end{cases}$$

Moreover, since the formula normalizes the difference by the sum of the probabilities, the range of possible values for $S(x)$ is constrained between -1 and 1. Such normalization allows for easier comparison between topics with different overall probabilities. If one gender dominates the corpus, it can still influence the resulting skewness measure. By analyzing the distribution of $S(x)$ across different attributes, researchers can identify patterns of gender bias within the text. In total, three attributes are considered to uncover different dimensions of gender bias, each described as follows.

Descriptors. This analysis focuses on the **descriptors** associated with PERSON entities identified in the dependency analysis step. Considering the lexicon proposed by Freitas et al. (2023), each associated word is categorized into four types of descriptors: "social", "emotional", "physical" (appearance), and "character".[5] Note that these categories are non-mutually exclusive, meaning a word can belong to multiple categories depending on its context. Words that do not fall into any predefined category are classified as "other". Following a similar approach applied in [Lucy and Bamman 2021], the main objective is to assess the alignment of individual descriptions with established gender stereotypes.

Adjectives. This analysis focuses on the **adjectives** associated with PERSON entities identified in the dependency analysis step. It provides insights into how adjectives are used to characterize and depict individuals, deepening the understanding of potential biases in portraying gender within the text.

Physical traits. This analysis focuses on the **physical traits** associated with PERSON entities identified in the dependency analysis step, using the dictionary of body parts. Through this analysis, the goal is to reveal how physical attributes are depicted concerning gender, offering insights into potential biases in depicting characters' physical appearances within the text.

4. Evaluation

In this section, we assess the effectiveness of our proposed NLP pipeline for detecting gender bias in Portuguese literary works. First, we describe the corpus employed (Section 4.1) and present statistics resulting from pipeline steps (Section 4.2). Next, we delve into the analysis of gender bias across three types of attributes (Section 4.3). Finally, we discuss the importance of the proposed pipeline and its possible limitations (Section 4.4).

[5]"Social" denotes professions, occupations, and social status descriptors, while "emotional" represents feelings, emotions and emotional tendencies. "Physical" refers to descriptions of physical appearance, while "character" refers to personality traits, including cognitive properties such as intelligence or its absence [Freitas and Santos 2023].

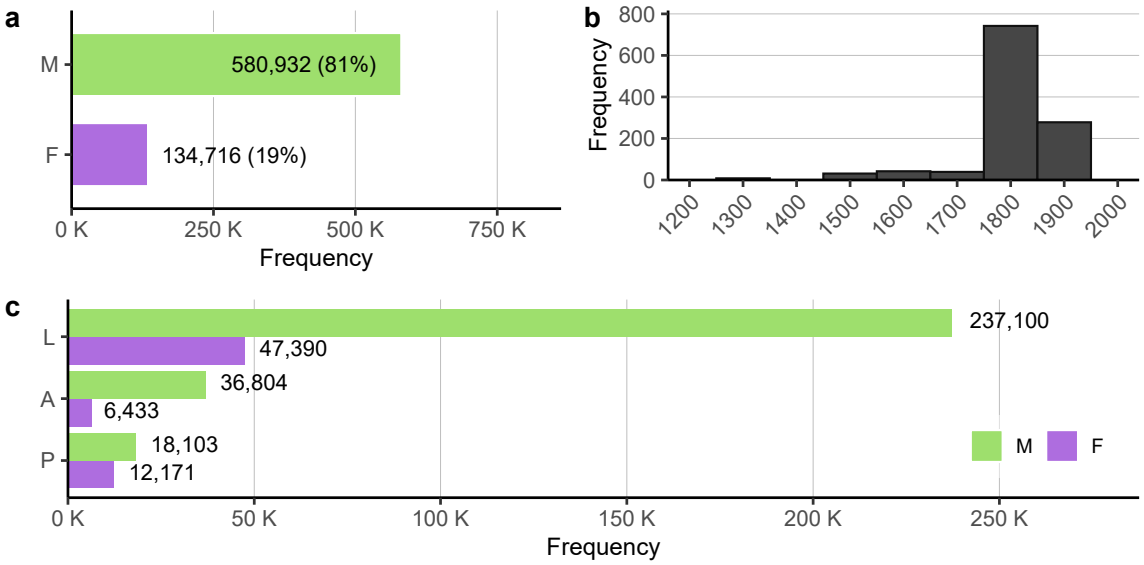

Figure 2. Overall statistics: distributions of (a) the gender classification, (b) the publication century of the literary works, and (c) the evaluated attributes. (D: descriptors, A: adjectives, PT: physical traits).

4.1. Corpus

Our evaluation considers a subset of 1,500 public-domain literary works in Portuguese sourced from the *PPORTAL* [Silva et al. 2022]. This dataset comprises an extensive metadata collection, including download links of over 9,000 public domain literary works predominantly from Brazil and Portugal. The subset selected includes full-length documents from various authors and literary genres. After applying the first step of the pipeline (i.e., preprocessing), 264 works are filtered out for having less than ten sentences. The resulting subset has a total of 1,236 literary works published from 1250 to 2019, with the majority published in the 18th and 19th centuries. In total, the corpus contains over 24 million tokens and is freely available for download.[1]

4.2. Pipeline Statistics

Once preprocessing all works, the next step is entity recognition. In this step, a total of 715,648 PERSON entities are identified and extracted. As depicted in Figure 2(**a**), 81% of the entities are classified as "male", indicating a significant dominance of male entities within the corpus, whereas only 19% are classified as "female". This skewed representation may reflect historical biases in literature, where male characters often dominate narratives while female characters are comparatively underrepresented. Such a notable gender disparity in the distribution of identified entities prompts further investigation into potential gender biases in Portuguese literary works.

Figure 2(**b**) shows the publication century distribution of the literary works. Such information is essential for understanding the temporal context in which these works were produced, as gender bias is usually influenced by the socio-cultural norms prevailing during a specific time period. The distribution reveals that most evaluated literary works were published between the 18th and 19th centuries. This temporal concentration underscores the significance of investigating gender bias across different historical epochs to discern evolving patterns and attitudes toward gender portrayal in literature.

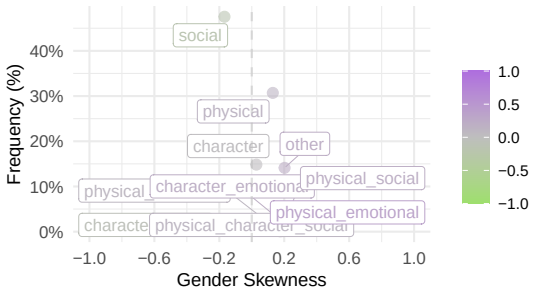

Figure 3. Distribution of categories occurrences.

Category	**Freq (%)**	$S(x)$
physical emotional	0.15	0.344
other	14.12	0.202
character emotional	2.10	0.191
physical social	0.13	0.189
emotional social	0.07	0.189
physical character social	30.70	0.103
physical character	2.31	0.090
character	14.83	0.028
character social	1.58	-0.064
social	47.55	-0.168

Table 1. Top skewed categories.

The following pipeline step is the dependency analysis, which extracts linguistic dependencies and syntactic patterns related to the identified PERSON entities. Such linguistic features are then used as input to the final pipeline step (gender bias detection). Such features are processed during this step to compute the gender skewness across three distinct dimensions: descriptors, adjectives, and physical traits. Figure 2(**c**) shows the distribution of each attribute by gender.

Overall, the physical traits dimension is the only attribute with balanced results, indicating similar frequencies regardless of gender. Conversely, the other two dimensions show noticeable disparities, presenting a higher frequency of adjectives and lexicons associated with male individuals. Such observations can be justified by the greater presence of individuals classified as male in the corpus. To obtain a more standardized assessment, we proceed to the gender bias analysis, as follows.

4.3. Gender Bias Analysis

In total, we evaluate three attributes to uncover different dimensions of gender bias. Figures 3 to 5 show the resulting gender skewness, $S(x)$, of each one. This measure provides a concise representation of the relative difference between the probabilities of an attribute occurring in excerpts with female and male PERSON entities, thereby facilitating a quantitative analysis of gender bias. Next, we discuss the results of each attribute.

4.3.1. Gender skewness for *descriptors*

The **descriptor** analysis identified 10 different types of descriptors formed by combinations of the four pre-defined categories (in addition to the "others" category). As illustrated in Figure 3, the most frequent categories include "social", "physical", "character" and "other". Overall, most categories exhibit positive skewness, suggesting a tendency for specific types of words to be more frequently associated with female entities. Table 1 highlights the top ten skewed categories, with only two demonstrating a negative measure ("social" and "character social"). Although most skewness values are close to zero, such findings indicate a subtle bias in the portrayal of gender in literary works.

Corroborating previous findings [Freitas and Santos 2023], our results reveal that female individuals are more likely to be characterized by their appearance, emotions, and personality traits, whereas social characteristics are more commonly associated with male

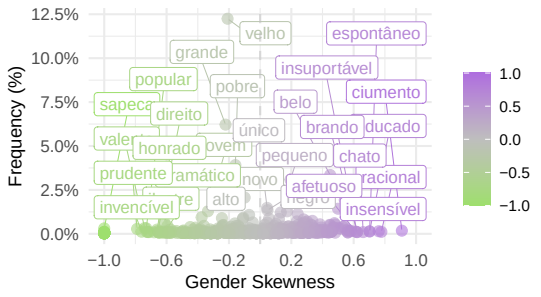

Figure 4. Distribution of adjective occurrences.

Adjective	**Freq (%)**	$S(x)$
ciumento (jealous)	0.18	0.909
racional (rational)	0.11	0.778
insensível (insensitive)	0.15	0.760
educado (polite)	0.13	0.702
espontâneo (spontaneous)	0.15	0.702
legal (cool)	0.08	-1.000
revolucionário (revolutionary)	0.09	-1.000
venerável (venerable)	0.11	-1.000
leal (loyal)	0.12	-1.000
sapeca (naughty)	0.31	-1.000

Table 2. Top skewed adjectives.

individuals. These findings suggest that gender-specific stereotypes may be prevalent in Portuguese literary works. For females, emphasizing appearance, emotions, and character traits may perpetuate traditional gender roles and expectations, potentially limiting their portrayal to superficial attributes. On the other hand, the association of social characteristics with male individuals might reinforce stereotypes related to authority, leadership, and societal roles traditionally attributed to men.

4.3.2. Gender skewness for *adjectives*

The **adjective** analysis identified 489 unique adjectives, providing insights into the linguistic portrayal of individuals in the corpus (Figure 4). Table 2 highlights the top ten skewed adjectives, showcasing their gender skewness values. Among these, the most positively skewed adjectives, such as “ciumento” (jealous), “racional” (rational), and “insensível” (insensitive), predominantly characterize female individuals. Conversely, adjectives like “legal” (cool) and “sapeca” (naughty) exhibit strong negative skewness values, suggesting a tendency to describe male individuals.

As observed in the previous analysis, emotional and character descriptors are more frequently associated with female individuals. This trend is also revealed in the gendered adjective analysis, where the most positively skewed adjectives predominantly refer to emotional and character attributes commonly associated with female gender stereotypes. For example, “ciumenta” (jealous) and “educada” (polite) are among the top skewed adjectives, reflecting the portrayal of traits often stereotypically attributed to women.

Conversely, for male individuals, the most frequently attributed adjectives refer to personality traits and social status. That is, descriptors from the “social” and “character” categories, as observed in the previous analysis. For example, “sapeca” (naughty) and “leal” (loyal) are among the top negative skewed adjectives, reflecting the portrayal of male characters with attributes associated with their behavior and social standing.

4.3.3. Gender skewness for *physical traits*

The last analysis provided by the pipeline is the exploration of physical traits stereotypes. The portrayal of the human body holds significant weight in literary works, allowing read-

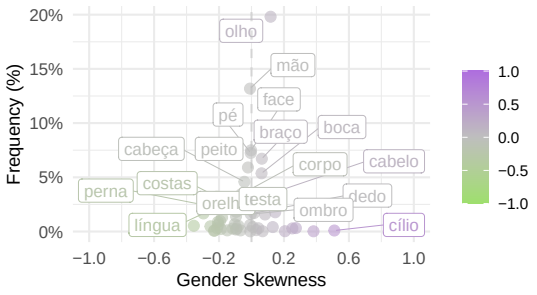

Figure 5. Distribution of physical traits occurrences.

Physical trait	Freq (%)	$S(x)$
cílio (eyelash)	0.09	0.510
antebraço (forearm)	0.02	0.382
cintura (waist)	0.32	0.273
pupila (pupil)	0.29	0.252
tornozelo (ankle)	0.03	0.206
nuca (nape)	0.09	-0.227
bunda (butt)	0.07	-0.231
punho (fist)	0.48	-0.253
língua (tongue)	1.67	-0.296
coluna (spine)	0.53	-0.355

Table 3. Top skewed physical traits.

ers to visualize characters vividly and immerse themselves fully in the narrative. However, the depiction of male and female bodies in literature can carry profound implications, potentially reinforcing and perpetuating societal gender norms and biases. Therefore, the primary aim of this analysis is to investigate the frequency of physical traits regarding the gender of characters and evaluate any correlated gender bias.

In total, 49 unique physical traits are identified within the literary works. As depicted in Figure 5, the distribution of gender skewness values for these physical traits tends to cluster around zero, indicating a balanced representation across male and female characters. The most skewed physical trait ($S(x) = 0.51$) is "cílio" (eyelash), which seems to be slightly more associated with female characters. This finding suggests that, unlike other dimensions of gender bias, the representation of physical traits in the evaluated corpus exhibits a relatively impartial distribution between male and female characters.

4.4. Discussion

Overall, our pipeline offers a comprehensive method for analyzing gender bias in literary works, providing a quantitative tool to uncover nuanced biases across various dimensions. The results of the descriptor and adjective analyses highlight prevalent gender-specific stereotypes in character descriptions: female individuals are often depicted with attributes related to appearance and emotions, while male ones are associated with social status and personality traits, reflecting traditional gender norms. Conversely, regarding physical traits stereotypes, a more balanced representation across genders is revealed, suggesting less pronounced bias in such a dimension.

Despite the relevant results, it is important to acknowledge the limitations of our pipeline. While it quantitatively assesses gender bias, it may not capture the full complexity of gender representation in literature. Factors such as authorial intent, narrative context, and reader interpretation can significantly influence the portrayal and perception of gender in literary texts, and these nuances may not be fully captured by automated analysis alone. Moreover, for the purpose of this study, we focus on binary values for gender (i.e., in individuals identified as male or female), recognizing that a more nuanced analysis of gender roles could be explored in future research.

Finally, the pipeline relies on predefined categories and linguistic features, which may not encompass the full range of gendered language and representation present in literary works. Additionally, the pre-trained NER models, the dependency parser, and the

gender detection methods applied during the different pipeline steps may not be perfectly accurate, potentially leading to misclassifications or oversimplifications of gender identities. Such limitation highlights the importance of ongoing refinement and validation of NLP algorithms for gender analysis in literary texts.

5. Conclusion

We introduced a novel NLP pipeline designed to unveil and analyze gender bias in Portuguese literary works. Our pipeline comprises six steps: preprocessing, entity recognition, excerpt segmentation, gender classification, dependency analysis, and gender bias detection. Its comprehensive approach considers multiple dimensions of gender representation, including linguistic stereotypes and physical traits, providing researchers with a quantitative method to explore nuanced biases in literature.

By evaluating our pipeline on a corpus of Portuguese literary works, we showed its effectiveness in identifying and quantifying gender bias across multiple dimensions. Our findings reveal prevalent gender stereotypes embedded within character descriptions, with female characters often associated with appearance and emotion, while male characters are depicted in terms of social status and personality traits. Furthermore, our analysis of physical traits stereotypes suggests a more balanced representation across genders in this dimension. Overall, our work contributes to the growing body of literature aimed at promoting gender equity and fostering more inclusive representations in literary texts.

While our pipeline provides a quantitative assessment of gender bias, it is important to acknowledge its potential limitations in capturing the full complexity of gender representation in literature. Future work could incorporate additional statistical tests, such as chi-squared tests or t-tests. Such statistical rigor would help validate the observed patterns and offer more concrete evidence of systemic biases. Moreover, refining the pipeline to address these limitations and exploring additional dimensions of gender representation beyond those covered in this study could be valuable avenues for future investigation.

Acknowledgments. The work is supported by CNPq, CAPES, and FAPEMIG, Brazil.

References

Casey, K., Novick, K., and Lourenco, S. F. (2021). Sixty years of gender representation in children's books: Conditions associated with overrepresentation of male versus female protagonists. *Plos one*, 16(12):e0260566.

Chu, K. E., Keikhosrokiani, P., and Asl, M. P. (2022). A topic modeling and sentiment analysis model for detection and visualization of themes in literary texts. *Pertanika Journal of Science & Technology*, 30(4):2535–2561.

Freitas, C. and Santos, D. (2023). Gender Depiction in Portuguese: Distant reading Brazilian and Portuguese literature. In *CCLS*, pages 1–27.

Goldman, S. R. and Lee, C. D. (2014). Text complexity: State of the art and the conundrums it raises. *The Elementary School Journal*, 115(2):290–300.

Gusmão, C., Figueiredo, K., and Brito, W. (2021). Técnicas de processamento de linguagem natural em denúncias criminais: Automatização e classificação de texto em português coloquial. In *SEMISH*, pages 172–182. SBC.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2017). Visual text analysis in digital humanities. *Computer Graphics Forum*, 36.

Kejriwal, M. and Nagaraj, A. (2024). Quantifying gender disparity in pre-modern english literature using natural language processing. *Journal of Data Science*, 22(1):77.

Labatut, V. and Bost, X. (2019). Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5):89:1–89:40.

Lucy, L. and Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. In *NUSE*, pages 48–55. ACL.

Maharjan, S. et al. (2018). Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *ACL*, pages 259–265.

Real, L., Johansson, K., Mendes, J., Lopes, B., and Oshiro, M. (2021). Generating e-commerce product titles in Portuguese. In *SEMISH*, pages 299–304. SBC.

Santana, B. S., Woloszyn, V., and Wives, L. K. (2018). Is there gender bias and stereotype in Portuguese word embeddings? In *PROPOR*, pages 24–26. Springer.

Silva, F. M. (2021). Diferenciações de gênero na caracterização de personagens: uma proposta metodológica e primeiros resultados. Master's thesis, Departamento de Letras, PUC-Rio.

Silva, M. et al. (2021). Exploring brazilian cultural identity through reading preferences. In *BraSNAM*, pages 115–126. SBC.

Silva, M., Melo-Gomes, L., and Moro, M. (2023). Gender representation in literature: Analysis of characters' physical descriptions. In *KDMiLe*, pages 17–24. SBC.

Silva, M. O., de Melo-Gomes, L., and Moro, M. M. (2024). From words to gender: Quantitative analysis of body part descriptions within literature in portuguese. *Information Processing & Management*, 61(3):103647.

Silva, M. O. and Moro, M. M. (2024a). Evaluating Pre-training Strategies for Literary Named Entity Recognition in Portuguese. In *PROPOR*, pages 384–393. ACL.

Silva, M. O. and Moro, M. M. (2024b). PPORTAL_ner: An Annotated Corpus of Portuguese Literary Entities. In *LREC*. ELRA. to appear.

Silva, M. O., Scofield, C., de Melo-Gomes, L., and Moro, M. M. (2022). Cross-collection dataset of public domain portuguese-language works. *JIDM*, 13(1).

Souza, F., Nogueira, R. F., and de Alencar Lotufo, R. (2019). Portuguese named entity recognition using BERT-CRF. *CoRR*, abs/1909.10649.

Xu, H., Zhang, Z., Wu, L., and Wang, C.-J. (2019). The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS one*, 14(11):e0225385.

Zahn, N., Molin, G. D., and Musse, S. (2021). Cross-media sentiment analysis on German blogs. In *SEMISH*, pages 114–122, Porto Alegre, RS, Brasil. SBC.