Gender Bias in Portuguese Literary Texts: A Masked Language Model Approach

Mariana O. Silva, Michele A. Brandão, Mirella M. Moro

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

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

Abstract. *In this work, we investigate how a corpus of Portuguese literary texts shapes the gender bias of a language model using a masked language modeling approach. We fine-tune a BERT-based model on a curated corpus of 592 literary works and analyze gender associations in adjective and verb predictions. Our results show that fine-tuning shifts gender associations: female-denoting targets showed a significant reduction in negative associations, while male-denoting targets retained a negative bias. For verbs, gender disparities decrease, though male subjects retain stronger links to intellectual/work-related verbs. These findings highlight how literary texts can shape gender representations in language models, reinforcing or reshaping biases based on training data.*

1. Introduction

Language models have become fundamental to Natural Language Processing (NLP), driving applications from text generation [Dong et al. 2022] to literary analysis [Silva and Moro 2024]. Still, research shows these models are not neutral; they encode and can even amplify bias present in their training data [Bolukbasi et al. 2016, Kurita et al. 2019]. These biases often manifest as gender, racial, and social stereotypes, influencing downstream applications in ways that may reinforce historical inequalities and perpetuate discriminatory language patterns [Stanczak and Augenstein 2021].

Early studies on bias in NLP focused on static word embeddings (e.g., Word2Vec), showing how they encode gender stereotypes [Bolukbasi et al. 2016, Garg et al. 2018]. However, the arrival of transformer-based models, such as BERT [Devlin et al. 2019], introduced contextualized word representations, making bias more dynamic and context dependent [Kurita et al. 2019, Bartl et al. 2020]. There are methods to measure bias in such models, e.g., log probability differences in masked language models (MLMs) [Kurita et al. 2019]. Still, most studies focus on English [Ding et al. 2025], leaving a gap in understanding how bias manifests in gendered languages, such as Portuguese.

Portuguese, with its gendered grammatical system, requires explicit morphological agreement (e.g., *ela é bonita* vs. *ele é bonito*), which may reinforce gender stereotypes through obligatory gender marking [Zhou et al. 2019]. Existing research on bias in Portuguese NLP has primarily analyzed general-domain corpora [Santana et al. 2018, Taso et al. 2023], overlooking literary texts despite their cultural significance. Literature encodes gendered associations through character portrayals, narrative structures, and linguistic choices, shaping societal perceptions across generations [Xu et al. 2019, Stuhler 2024]. For instance, female characters are often described in

terms of physical appearance and emotions, whereas male characters are linked to agency and intellect [Freitas and Santos 2023, Silva and Moro 2024].

Here, we investigate how a corpus of Portuguese literary texts shapes the gender bias of a pre-trained language model. Our study leverages masked language modeling (MLM) in two main capacities: first, for domain adaptation, by fine-tuning the BERTimbau model [Souza et al. 2020] on our curated corpus of 592 prose works (1804–1998); and second, for bias measurement, by using masked token probing to quantify gender associations. To do so, we build a template-based sentence corpus with gendered targets and specific attributes (adjectives and verbs). We then apply the log probability difference metric [Kurita et al. 2019] to compare gender associations before and after fine-tuning, assessing whether exposure to literary narratives amplifies, mitigates, or reshapes existing biases within the model. Our main contributions are summarized as follows.

- We adapt template-based bias measurement to Portuguese, accounting for grammatical gender agreement in adjectives and verbs;
- We compile and pre-process a diverse literary corpus, balancing Brazilian and European Portuguese works, and build a gender bias evaluation corpus of Portuguese-language templates derived from common linguistic patterns in literary texts; and
- We compare gender bias in pre-trained and fine-tuned versions of BERTimbau to evaluate the impact of literary corpora on contextualized word associations.

2. Related Work

Bias in language models has been extensively studied in NLP, with growing research interest in both static and contextualized word representations. This section summarizes previous work on bias in language models and explores the intersection of NLP and literary analysis, focusing on Portuguese-language texts to contextualize our contributions.

Bias in Language Models. Early research on bias, now considered foundational, predominantly focused on static word embeddings (e.g., Word2Vec, GloVe). These studies demonstrated how embeddings encode gender stereotypes through vector space associations, such as linking "doctor" to men and "nurse" to women [Bolukbasi et al. 2016]. The introduction of the Word Embedding Association Test (WEAT) [Caliskan et al. 2017] provided a crucial tool for systematic bias measurement, while subsequent work explored various mitigation techniques [Gonen and Goldberg 2019, Zhao et al. 2018]. The advent of transformer-based models brought new complexities; the shift to contextualized embeddings meant word representations became dynamic and context-dependent [May et al. 2019], necessitating new methods like template-based or probabilistic approaches to measure associations [Kurita et al. 2019, Bartl et al. 2020].

Building on these methods, Kurita et al. (2019) adapted WEAT for masked language models (MLMs), using log-probability differences in sentence templates to show that BERT also incurs gendered associations similar to those in static embeddings. Bartl et al. (2020) extended this approach, showing that fine-tuning on specific corpora can alter biases encoded in contextualized models. Despite these advancements, much of the research on bias in NLP has centered on English, a language with relatively weak grammatical gender marking. Studies on gendered languages, where grammatical gender is explicitly encoded in morphology and syntax, reveal additional complexities. For instance, Zhou et al. (2019) investigated bias in bilingual embeddings and proposed

an evaluation metric for languages requiring gender morphological agreement, such as Spanish and French. More recently, Omrani Sabbaghi and Caliskan (2022) found that grammatical gender signals can interfere with measurements of social gender bias in word embeddings of gendered languages.

For Portuguese, a grammatically gendered language, research on gender bias is a growing but still relatively limited field [Lima and Araujo 2023]. Early studies primarily focused on static word embeddings in general-domain corpora. For example, Hartmann et al. (2017) evaluated various embedding models using syntactic and semantic analogy tasks, while Santana et al. (2018) examined associations between professions and gender stereotypes in Portuguese word embeddings. Similarly, Taso et al. (2023) analyzed sexism in GloVe embeddings trained on Portuguese data, demonstrating that these models capture both commonsense knowledge and gender-based biases. More recent work has also investigated how such biases manifest in large generative language models [Rodrigues et al. 2023, Assi and Caseli 2024].

Bias in Literary Texts. Computational approaches are increasingly used in literary analysis [Silva and Moro 2024]. In the context of gender bias, studies indicate that literary texts often reflect and reinforce traditional stereotypes, with female characters disproportionately described through physical appearance and male characters through agency [Cheng 2020, Schulz and Bahník 2019, Luo et al. 2024]. For example, Xu et al. (2019) explored fairy tales, finding remarkable stability in stereotypical gender roles over time. Stuhler (2024) employed large-scale computational methods to track historical shifts in gender representation in contemporary fiction, identifying trends toward greater balance but also persistent stereotypes.

While most research in this area focuses on English, recent Portuguese studies reveal similar patterns. Silva et al. (2023, 2024) conducted a quantitative analysis of gendered descriptions in a corpus of 34 Portuguese literary works, finding that female characters are objectified through body part descriptions. Similarly, Freitas and Santos (2023) explored lexical patterns in Portuguese literary corpora, showing that descriptions of female characters tend to focus on physical appearance, whereas male characters are more often defined by their actions or social roles. Silva and Moro (2024) further studied bias by using an NLP pipeline to analyze gender representation in fiction.

Research Gaps. Despite recent advances in gender bias research in NLP, several issues remain open. First, most existing studies focus on English, with fewer efforts dedicated to understanding how bias manifests in gendered languages. Second, research on bias in Portuguese embedding models has primarily analyzed general-domain corpora, with limited focus on literary texts, despite their crucial role in shaping societal narratives and cultural significance. Our work addresses both research gaps by investigating gender bias in Portuguese literary texts using a masked language modeling approach and template-based methods, specifically with the BERTimbau model.

3. Corpus

To investigate how literary texts influence gender representations in language models, we built a curated corpus of Portuguese literary works sourced from four well-established corpora, selected to ensure a broad temporal span (1804–1998) and balanced representation of both Brazilian and European Portuguese. The dataset integrates texts from *Colonia*

Table 1. Composition of Portuguese literary corpora used in this study.

Corpus	Period	#Works	#Sentences	#Tokens
OBras [Santos et al. 2018]	1855–1984	23	54,317	1,005,266
Colonia [Zampieri and Becker 2013]	1844–1948	35	171,741	2,797,503
ELTeC-por [Santos 2021]	1844–1973	37	209,614	3,139,395
PPORTAL [Silva et al. 2021]	1804–1998	497	788,542	10,641,252
Total	**1804–1998**	**592**	**1,224,214**	**17,583,416**

[Zampieri and Becker 2013], *ELTeC-por* [Santos 2021], *OBras* [Santos et al. 2018], and *PPORTAL* [Silva et al. 2021, Silva et al. 2022]. Together, these corpora offer a diverse selection of novels and short stories across different literary movements.

From an initial collection of 840 works, we excluded non-narrative genres (poetry and plays) to ensure analytical consistency. This selection criterion is needed for two main reasons: (*i*) poetry often employs non-standard syntax, symbolic language, and unconventional structures that would hinder meaningful comparison with prose works; and (*ii*) dramatic texts are primarily composed of dialogue and typically lack the extended narrative descriptions essential for analyzing gendered language patterns in character portrayals [Freitas and Santos 2023].

The final corpus comprises 592 prose works (70% of the initial collection), totaling 1.2 million sentences and 17.6 million tokens. Text pre-processing includes text cleaning and sentence segmentation tailored for Portuguese texts [Silva and Moro 2024]. Table 1 provides an overview of the corpus.

4. Fine-tuning

To study how literary texts shape gender representations, we first adapt a pre-trained language model to the literary domain. We fine-tuned BERTimbau [Souza et al. 2020] on our literary corpus (Section 3) using the standard masked language modeling (MLM) task. This domain adaptation is not meant to directly address bias but to capture literary-specific linguistic patterns (e.g., archaic vocabulary, stylistic variations), producing a model shaped by literary narratives that can later be probed for gender bias.

Setup. Fine-tuning the BERTimbau Base model[1] uses the Hugging Face *Transformers* library. We employed the standard MLM approach, where 15% of the input tokens were randomly masked for prediction. The fine-tuning process runs for 10 epochs with a batch size of 16, a learning rate of 5×10^{-5}, and a weight decay of 0.01.

Training. We use the Hugging Face *Trainer* API with dynamic token masking applied at each training step. The MLM loss function is used for optimization and computed over the masked tokens in each batch. We employ the AdamW optimizer with weight decay to mitigate overfitting. Training is performed on an NVIDIA GeForce RTX 4050 (6GB VRAM), and gradient accumulation is used to manage memory constraints.

Evaluation. We assess adaptation quality through *perplexity* (PPL) on a held-out validation set (20% of corpus). After fine-tuning, the model achieved a PPL of 2.71, compared to 3.16 pre-fine-tuning, reflecting a 14.2% reduction. This decrease, though

[1] `https://huggingface.co/neuralmind/bert-base-portuguese-cased`

moderate, indicates the model has improved its predictive performance and adapted to some extent to the linguistic patterns of literary texts.

5. Gender Bias Assessment

Building on previous methodologies [Bartl et al. 2020, Kurita et al. 2019], we develop a framework to quantify gender bias in Portuguese literary texts using masked language model probing. Our approach extends previous studies by: (*i*) incorporating Portuguese-specific grammatical gender constraints, and (*ii*) distinguishing between descriptive stereotypes (evaluated through adjectives) and agency stereotypes (evaluated through verbs). Our framework consists of two main components: template design (Section 5.1) and bias measurement (Section 5.2).

5.1. Template Design

To capture different types of gender bias, we design two categories of sentence templates: *adjective-based* and *verb-based*. *Adjective-based* templates capture descriptive stereotypes by associating gendered noun phrases with positive or negative adjectives, while *verb-based* templates evaluate action-related stereotypes by analyzing gendered subject-verb associations across cognitive, perceptual, occupational, and social interactions. Such templates provide a broad coverage of gender bias, with the former emphasizing personality traits and emotional characteristics, and the latter focusing on role-based distinctions in activities and behaviors.

Gendered Noun Phrases (`<person>`). We select ten gendered noun phrases (five male and five female) from the BP-LIWC2015 [Carvalho et al. 2019, Carvalho et al. 2024], a Brazilian Portuguese adaptation of the LIWC dictionary. Specifically, we extract terms from its *female* and *male* categories, ensuring all selections are in the singular third person for grammatical consistency. To maintain semantic and syntactic balance, we manually verify each gendered pair (e.g., "ela" for female and "ele" for male).

Adjectives and Verbs (`<adj>` and `<verb>`). For adjectives and verbs, we use PortiLexicon-UD [Lopes et al. 2022], a lexicon containing 1.2 million Portuguese word forms with detailed morphological tags. Unlike English, Portuguese is a gender-marking language, meaning that adjectives and verbs often agree in gender with their subjects, a key distinction from English. Consequently, to ensure fair bias measurement, we inflect all adjectives in both masculine and feminine forms and align verb conjugations with their respective subject gender.

Sentence Templates. For *adjective-based* templates, we extract 180 adjectives (90 positive, 90 negative) from the *adj*, *posemo*, and *negemo* categories of BP-LIWC2015 to measure descriptive stereotypes related to sentiment and personality. These adjectives are combined with four commonly used linking verbs, including "ser" (to be), "estar" (to be in a temporary state), "ficar" (to become), and "continuar" (to remain), with multiple conjugations to ensure linguistic diversity. For the *verb-based* templates, we select 120 verbs to evaluate stereotypes related to agency and social roles, drawing from the *cogproc* (cognitive processes), *percept* (perception), *social*, and *work* categories of BP-LIWC2015. These specific semantic axes were chosen as they are central to gender bias research, allowing for a nuanced analysis of associations with intellect and work versus social interaction and perception.

Table 2. Template specifications and examples.

Type	Template	Examples	Count
Adjective-based	<person> *(é, era, será, está, estava, ficou, continua, continuou, continuava)* <adj>	*ela é feliz* *ele era feliz*	10,800
Verb-based	<person> <verb>	*ela sentiu* *ele sentiu*	1,200
Total			**12,000**

In total, ten sentence templates are designed to cover both *adjective-based* and *verb-based* structures, as samples in Table 2. Using such templates and systematically combining gendered noun phrases, adjectives, and verbs, we generate a total of 12,000 unique sentences. The complete generated sentences are available at [Silva et al. 2025].

5.2. Bias Measurement

To quantify gender bias, we adopt a methodology based on masked token prediction, following the approach of Kurita et al. (2019), which extends the Word Embedding Association Test (WEAT) [Caliskan et al. 2017] to masked language models. This methodology allows us to assess how specific attributes (e.g., adjectives and verbs) impact the likelihood of gendered person words appearing in a sentence. By using masked language models, we can directly evaluate how the model predicts gendered words in context, offering a straightforward way to measure gender bias in sentence generation.

In our experimental setup, *targets* (T) refer to gendered person words (e.g., "ele" for male, "ela" for female), while *attributes* (A) consist of adjectives and verbs. We hypothesize that in a masked language model, the probability of a target word is influenced by the surrounding context, so the presence of an attribute should affect the likelihood of the target: $P(T) \neq P(T|A)$. Furthermore, we assume that the same attribute will impact male- and female-denoting targets differently: $P(T_{female}|A) \neq P(T_{male}|A)$. To measure this association, we use the sentence templates from Section 5.1.

For each template, we compute the target probability P_T of the masked target when the attribute is present, and the prior probability P_{prior} of the masked target when the attribute is absent. Both probabilities are derived from BERT-based language models (both pre-trained and fine-tuned). We apply the softmax function to the predicted logits at the masked position, generating a probability distribution over all vocabulary tokens. From this distribution, we extract the probability of the target word by locating its index in the model's vocabulary. The association score between a target (T) and an attribute (A) is computed through the following six steps [Bartl et al. 2020]:

1. Select a sentence containing a target and an attribute, e.g., "ela era feliz"
2. Mask the target word: "[MASK] era feliz"
3. Calculate the target probability: $P_T = P(ela = [MASK]|sent)$
4. Mask both the target and the attribute: "[MASK] era [MASK]"
5. Compute the prior probability: $P_{prior} = P(ela = [MASK]|masked\_sent)$
6. Compute the association score: $Association(T, A) = \log(\frac{P_T}{P_{prior}})$

The resulting association score quantifies the influence of the attribute on the likelihood of the gendered target word. A negative association score indicates that the

Table 3. Gender association scores across model versions.

Category	Pre-trained	Fine-tuned	Δ	p_{pre}	p_{pos}
Adjective-based (female)	-0.79 ± 1.70	0.15 ± 1.49	$+0.94$	0.56	< 0.001
Adjective-based (male)	-0.71 ± 1.52	-0.64 ± 1.51	$+0.08$		
Verb-based (female)	-1.91 ± 3.42	0.15 ± 2.06	$+2.06$	< 0.001	0.0005
Verb-based (male)	-0.88 ± 2.79	0.57 ± 1.74	$+1.45$		

where: values have mean $\pm$ standard deviation; Δ is the difference between pre-trained and fine-tuned scores; p_{pre} and p_{pos} are the statistical significance of gender differences (female vs. male) before and after fine-tuning, respectively

attribute reduces the likelihood of the target compared to the prior probability, suggesting that the attribute is less associated with the gendered target. In contrast, a positive association score indicates a stronger association between the attribute and the gendered target. The magnitude of the association score reflects the strength of this relationship: larger absolute values suggest a more pronounced association, while scores near zero indicate a neutral or weak connection between the attribute and the target.

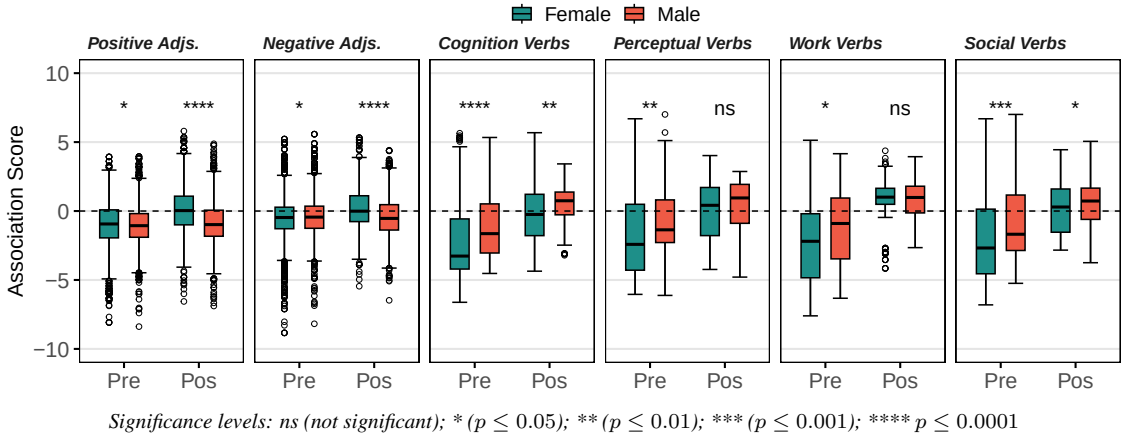
6. Experimental Results

In this section, we present the results of our gender bias measurement experiments. We analyze how gender associations manifest in the masked language model's predictions by comparing the pre-trained BERTimbau model with the version fine-tuned on our Portuguese literary corpus. To do so, we computed association scores for each *adjective-based* and *verb-based* template to assess the impact of fine-tuning.

Overall Trends. Table 3 presents the average association scores for female and male target words in both model versions. For *adjective-based* templates, the pre-trained model showed negative associations for both female (-0.79 ± 1.70) and male (-0.71 ± 1.52) targets, with no statistically significant difference between them (Wilcoxon, $p_{pre} = 0.56$). After fine-tuning, the average association score for female-related targets increased from -0.79 to 0.15 ($\Delta = +0.94$), while the score for male-related targets remained relatively stable (-0.71 to -0.64, $\Delta = +0.08$). This indicates that fine-tuning increased the model's likelihood of associating female-denoting words with adjectives.

For *verb-based* templates, the pre-trained model exhibited a more pronounced negative average association for female-denoting words (-1.91 ± 3.42) compared to male targets (-0.88 ± 2.79). This suggests that verbs across the tested categories (cognition, perception, social, and work) were initially less likely to be predicted when the masked subject was female. Following fine-tuning, the average association scores for verbs increased for both genders: for male-denoting words, from -0.88 to 0.57 ($\Delta = +1.45$), and for female-denoting words, more substantially from -1.91 to 0.15 ($\Delta = +2.06$). This marked change suggests that fine-tuning reduced the initial disparity for female subjects in *verb-based* contexts, though male subjects retained a higher average association score with verbs post-fine-tuning.

Category-level Analysis. Figure 1 displays the average association scores for female and male targets across specific attribute categories. For *adjective-based* templates in the pre-trained model, both positive and negative adjectives yielded negative average association scores for both male and female targets, with no statistically significant



*Significance levels: ns (not significant); * ($p \leq 0.05$); ** ($p \leq 0.01$); *** ($p \leq 0.001$); **** $p \leq 0.0001$*

Figure 1. Gender bias measurement across different adjective polarities and verb categories, comparing (pre) pre-trained and (pos) fine-tuned models. Significance levels are indicated, with *ns* denoting non-significant results.

difference between genders (Wilcoxon, $p_{pre} \leq 0.05$). After fine-tuning, female targets became more associated with positive adjectives, while male targets remained negatively associated with them. For negative adjectives, female targets showed a small positive shift in association, while male targets maintained a slight negative association.

For *verb-based* templates, the pre-trained model showed negative average association scores across all four verb categories (work, social, perceptual, and cognition) for both genders. This effect was more pronounced for female-denoting words, particularly in cognition (female: -2.04 ± 3.33 vs. male: -0.86 ± 2.63), social (female: -1.99 ± 3.38 vs. male: -0.93 ± 2.84), and work (female: -2.08 ± 3.69 vs. male: -1.13 ± 2.96) categories. These results confirm that, before fine-tuning, the model was significantly less likely to associate female subjects with verbs related to thinking, working, or social interactions compared to male subjects. After fine-tuning, association scores for verbs generally became positive or closer to neutral. Male subjects, in particular, showed a significant increase in association scores across all verb types, shifting from negative to positive average associations. For female targets, while scores also increased, the increase was comparatively weaker for certain categories, with cognition-related verbs, for instance, remaining slightly negatively associated.

Qualitative Analysis. To further explore the impact of fine-tuning, we analyzed specific adjectives with the highest disparity shifts (Figure 2). For positive adjectives associated with female targets (Figure 2a), all top five adjectives showed a substantial increase in association scores ($\Delta > +2.5$), including "leal" (loyal), "favorita" (favorite), and "saudável" (healthy). For male targets (Figure 2b), shifts for positive adjectives like "formidável" (formidable) and "leal" (loyal) were comparatively smaller. Regarding negative adjectives, for female targets (Figure 2a), "perigosa" (dangerous) and "pobre" (poor) showed notable increases in association scores. Conversely, for male targets (Figure 2b), most of the top five high-disparity negative adjectives shifted towards more neutral or even positive associations, such as "desgraçado" (wretched), "insignificante" (insignificant), and "podre" (rotten).

Discussion. Our findings show that fine-tuning BERTimbau on a curated corpus of

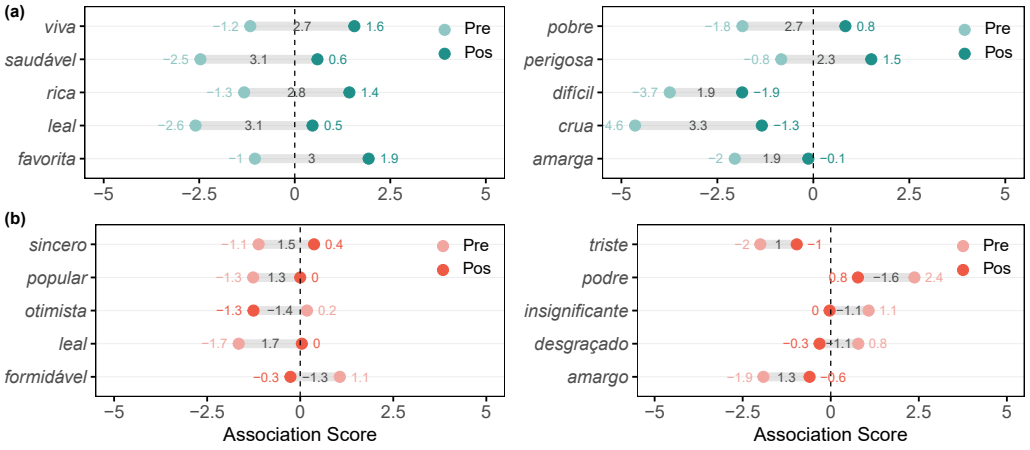


Figure 2. Top five high-disparity (left) positive and (right) negative adjective associations for (a) female and (b) male targets, comparing pre-trained (Pre) and fine-tuned (Pos) models.

Portuguese literary texts significantly alters its gender association patterns, though in complex and sometimes diverging ways for female and male targets. A prominent finding is the increased association of female-denoting words with adjectives after fine-tuning (Table 3). This shift, particularly towards positive adjectives (Figure 1), suggests that exposure to the literary corpus encouraged a more descriptively rich representation of female targets. This may reflect the stylistic tendencies within the literary texts themselves, where female characters are more frequently described with descriptive attributes [Freitas and Santos 2023, Silva and Moro 2024]. The qualitative analysis further supports this trend, showing strong positive shifts for adjectives like “leal” (loyal) and “saudável” (healthy) for females (Figure 2a).

For verb-based associations, fine-tuning reduced gender disparity by countering the pre-trained model’s tendency to disassociate female subjects from actions and intellectual roles (Table 3). Both genders showed higher verb association scores, indicating a reduced reluctance to link gendered subjects with actions after exposure to narrative texts. Post-fine-tuning scores for female targets approached neutrality for adjectives (0.15 ± 1.49) and verbs (0.15 ± 2.06), reflecting a correction of the strong negative biases in the original model rather than the absence of association. Exposure to literary narratives, where female characters act and are described, balanced these scores.

However, the correction was not uniform. Male-denoting words consistently retained higher association scores with verbs, particularly those related to work and cognition (Figure 1). This persistence suggests that the literary corpus, while diversifying female roles to some extent, still carries stronger traditional associations of male characters with agency and intellect, a pattern noted in previous studies of literary texts [Freitas and Santos 2023, Silva and Moro 2024].

The qualitative analysis of adjectives revealed a gendered asymmetry in how negative attributes are treated post-fine-tuning (Figure 2). While female targets became more associated with certain negative adjectives like “perigosa” (dangerous), indicating an intensification of specific negative stereotypes, male targets saw a shift where adjectives with negative semantics (e.g., wretched and insignificant) became more neutrally or

even positively associated with them, with their association scores moving from negative values toward zero or above. This intriguing finding suggests that the literary domain, as captured by the model, might reframe or soften certain negative masculine stereotypes while potentially reinforcing or specifying others for females.

Overall, our findings demonstrate that fine-tuning on literary texts does not simply increase or decrease bias, but rather reshapes it in complex ways. The model's final state is a hybrid, reflecting stereotypes from both its original general-domain training data and the specific literary corpus. For instance, the fine-tuning process mitigated a strong pre-existing bias that disassociated female subjects from verbs, yet it also reinforced stereotypes prevalent in the literature, such as the stronger association of female characters with descriptive attributes and the persistent link between male characters and agentic roles. This highlights a crucial takeaway: AI systems inherit, combine, and transform the biases present in all human-generated text they are exposed to, rather than simply erasing them [Bolukbasi et al. 2016, Kurita et al. 2019].

7. Conclusion

This study investigated how a literary corpus reshapes the gender bias of a pre-trained Portuguese language model. By fine-tuning BERTimbau on 592 works and probing the resulting model, we found that exposure to literary texts does not simply eliminate or intensify bias but rather transforms it. Overall, fine-tuning increased the positive association of female-denoting targets with adjectives, while male-denoting targets maintained a negative association. For verbs, gender disparities decreased, but male-denoting targets still showed stronger associations with action-oriented and intellectual roles. These findings highlight the impact of literary texts on shaping gender representations in language models, which can simultaneously challenge some pre-existing biases while reinforcing others native to the literary domain.

Limitations. While this study provides valuable insights, certain limitations should be acknowledged. First, our probing methodology, though controlled, relies on artificial templates that may not fully capture how bias manifests in natural language. Second, our corpus, while diverse, reflects specific historical periods, and its findings may not generalize to all Portuguese literature, especially contemporary works. The choice of attributes, though systematic, is also finite.

Future Directions. A natural next step is to conduct a diachronic analysis to track how gender representations change over time, potentially correlating these shifts with societal transformations. Furthermore, an analysis incorporating author demographics, such as gender, could reveal how the identity of the writer influences the textual portrayal of characters and, consequently, the biases learned by the model. Finally, exploring alternative bias metrics beyond template-based probing would provide a more holistic understanding of how language models internalize and reproduce cultural biases.

Acknowledgments. This work was funded by *Fundação de Amparo à Pesquisa do Estado de Minas Gerais*, (FAPEMIG), and supported by *Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação* (INCT-TILD-IAR).

References

Assi, F. M. and Caseli, H. d. M. (2024). Biases in GPT-3.5 Turbo model: a case study regarding gender and language. In *Simp. Bras. de Techologia da Informação e da Linguagem Humana*, STIL, pages 294–305. SBC.

Bartl, M., Nissim, M., and Gatt, A. (2020). Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *GeBNLP*, pages 1–16. ACL.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., et al. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *NeurIPS*, volume 29. Curran Associates, Inc.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Carvalho, F., Junior, F. P., Ogasawara, E., et al. (2024). Evaluation of the brazilian portuguese version of linguistic inquiry and word count 2015 (BP-LIWC2015). *Language Resources and Evaluation*, 58(1):203–222.

Carvalho, F., Rodrigues, R., Santos, G., et al. (2019). Avaliação da versão em português do liwc lexicon 2015 com análise de sentimentos em redes sociais. In *BRASNAM*, pages 24–34. SBC.

Cheng, J. (2020). Fleshing Out Models of Gender in English-Language Novels (1850 – 2000). *Journal of Cultural Analytics*, 5(1).

Devlin, J., Chang, M.-W., Lee, K., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Ding, Y., Zhao, J., Jia, C., Wang, Y., Qian, Z., Chen, W., and Yue, X. (2025). Gender bias in large language models across multiple languages: A case study of ChatGPT. In *TrustNLP*, pages 552–579. ACL.

Dong, C., Li, Y., Gong, H., et al. (2022). A Survey of Natural Language Generation. *ACM Comput. Surv.*, 55(8):173:1–173:38.

Freitas, C. and Santos, D. (2023). Gender depiction in portuguese. In *CCLS*, pages 4–30.

Garg, N., Schiebinger, L., Jurafsky, D., et al. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, 115(16):E3635–E3644.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *WiNLP*, pages 60–63. ACL.

Hartmann, N. S. et al. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In *STIL*, pages 122–131. SBC.

Kurita, K. et al. (2019). Measuring bias in contextualized word representations. In *GeBNLP*, pages 166–172. ACL.

Lima, L. F. F. P. d. and Araujo, R. M. d. (2023). A call for a research agenda on fair NLP for Portuguese. In *STIL*, pages 187–192. SBC.

Lopes, L., Duran, M., Fernandes, P., et al. (2022). PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model. In *LREC*, pages 6635–6643. ELRA.

Luo, K. et al. (2024). Reflecting the Male Gaze: Quantifying Female Objectification in 19th and 20th Century Novels. In *LREC-COLING*, pages 13803–13812, Torino, Italia. ELRA and ICCL.

May, C., Wang, A., Bordia, S., et al. (2019). On Measuring Social Biases in Sentence Encoders. In *NAACL*, pages 622–628. ACL.

Omrani Sabbaghi, S. and Caliskan, A. (2022). Measuring Gender Bias in Word Embeddings of Gendered Languages Requires Disentangling Grammatical Gender Signals. In *AIES*, pages 518–531. ACM.

Rodrigues, G., Albuquerque, D., and Chagas, J. (2023). Análise de vieses ideológicos em produções textuais do assistente de bate-papo chatgpt. In *Anais do IV Workshop sobre as Implicações da Computação na Sociedade*, WICS, pages 148–155. SBC.

Santana, B. S., Woloszyn, V., and Wives, L. K. (2018). Is there Gender bias and stereotype in Portuguese Word Embeddings? arXiv:1810.04528.

Santos, D. (2021). Portuguese Novel Corpus (ELTeC-por): April 2021 release. Zenodo. doi:10.5281/zenodo.4288235.

Santos, D., Freitas, C., and Bick, E. (2018). OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in public domain. *CorLex*. `https://www.linguateca.pt/OBRAS/OBRAS.html`.

Schulz, D. and Bahník, Š. (2019). Gender associations in the twentieth-century English-language literature. *Journal of Research in Personality*, 81:88–97.

Silva, M., Brandão, M., and M. Moro, M. (2025). Gender Bias in Portuguese Literary Texts: A Masked Language Model Approach. In *Zenodo*. doi:10.5281/zenodo.16748552.

Silva, M. and Moro, M. (2024). NLP Pipeline for Gender Bias Detection in Portuguese Literature. In *SEMISH*, pages 169–180. 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., 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., Scofield, C., de Melo-Gomes, L., et al. (2022). Cross-collection dataset of public domain portuguese-language works. *Journal of Information and Data Management*, 13(1):95–110.

Silva, M. O., Scofield, C., and Moro, M. M. (2021). PPORTAL: Public domain Portuguese-language literature Dataset. In *SBBD - DSW*, pages 77–88. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *BRACIS*, pages 403–417. Springer.

Stanczak, K. and Augenstein, I. (2021). A Survey on Gender Bias in Natural Language Processing. arXiv:2112.14168.

Stuhler, O. (2024). The gender agency gap in fiction writing (1850 to 2010). *PNAS*, 121(29):e2319514121.

Taso, F. T. d. S., Reis, V. Q., and Martinez, F. V. (2023). Sexismo no Brasil: análise de um Word Embedding por meio de testes baseados em associação implícita. In *STIL*, pages 53–62. SBC.

Xu, H., Zhang, Z., Wu, L., et al. (2019). The Cinderella Complex: Word embeddings reveal gender stereotypes in movies and books. *PLOS ONE*, 14(11):e0225385.

Zampieri, M. and Becker, M. (2013). Colonia: Corpus of historical portuguese. In Zampieri, M. and Diwersy, S., editors, *Special Volume on Non-Standard Data Sources in Corpus-Based Research*, volume 5 of *ZSM Studien*, pages 77–84. Shaker Verlag, Aachen, Germany.

Zhao, J. et al. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *NAACL*, pages 15–20. ACL.

Zhou, P. et al. (2019). Examining gender bias in languages with grammatical gender. In *EMNLP-IJCNLP*, pages 5276–5284. ACL.