UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Mariana de Oliveira Santos Silva

A Computational Framework for Measuring and Analyzing Gender Bias in Portuguese-language Literary Texts

Belo Horizonte
2025

Mariana de Oliveira Santos Silva

A Computational Framework for Measuring and Analyzing Gender Bias in Portuguese-language Literary Texts

Final Version

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Mirella Moura Moro
Co-Advisor: Michele Amaral Brandão

Belo Horizonte
2025

Silva, Mariana de Oliveira Santos.

S586c A Computational framework for measuring and analyzing gender bias in portuguese-language literary texts [recurso eletrônico] / Mariana de Oliveira Santos Silva – 2025.
1 recurso online (214. il, color.) : pdf.

Orientadora: Mirella Moura Moro.
Coorientadora: Michele Amaral Brandão.

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.

Referências: f. 152-174

1. Computação – Teses. 2. Recuperação da informação – Teses. 3. Processamento de linguagem natural – Teses. 4. Mineração de dados (Computação) – Teses. 5. Identidade de gênero na literatura – Teses. 6. Língua portuguesa – Análise de Textos – Teses. I. Moro, Mirella Moura. II.Brandão, Michele Amaral. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6*73(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

A Computational Framework for Measuring and Analyzing Gender Bias in Portuguese-language Literary Texts

MARIANA DE OLIVEIRA SANTOS SILVA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. MIRELLA MOURA MORO - Orientadora
Departamento de Ciência da Computação - UFMG

PROFA. MICHELE AMARAL BRANDÃO - Coorientadora
Departamento de Ciência da Computação - UFMG

PROFA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - UFMG

PROFA. KARIN BECKER
Instituto de Informática - UFRGS

PROFA. TICIANA LINHARES COELHO DA SILVA
Programa de Pós-Graduação em Modelagem e Métodos Quantitativos - UFC

PROF. EVANDRO LANDULFO TEIXEIRA PARADELA CUNHA
Faculdade de Letras - UFMG

Belo Horizonte, 11 de dezembro de 2025.

To my family and all who believed in me. You made this possible — please keep believing, because I still have no idea what I'm doing next.

Acknowledgments

Pursuing a Ph.D. was never a childhood dream of mine, but as I reach the end of this journey, I feel deeply fulfilled and proud of the choice I made. These past years have been a period of immense learning, both academically and personally, and I am grateful for every step of the way.

First and foremost, I would like to thank my family. To my mother, Luciana, who has always been by my side, ready to help with anything I need. To my father, Valdivino, whose work ethic, integrity, and honesty have always been an inspiration, and who has continuously supported and guided me in the best way possible. To my sister, Gabriela, who always cheers for me and is not only family but also a friend. And to my dog, Forró, who has comforted me in the moments I needed it most, even without saying a word. And finally, though certainly not least, to my partner, Matheus, who has been by my side every step of the way, supporting me, encouraging me, and making me feel truly special. Without them, I would not have made it this far.

I am profoundly grateful to my advisor, Professor Mirella Moro, who has been like an academic mother to me. She has always been there, supporting me, pushing me forward and upward, and constantly making me feel capable of achieving anything I set my mind to. It has been an honor to be her student and to work with her throughout all these years (*and there were many!*). Mirella's energy is unmatched: always positive, always uplifting, and always bringing brilliant and insightful ideas. I am deeply thankful for all her academic, professional, and personal guidance. Having her by my side during this journey was fundamental, and I would not have chosen anyone else to guide me. I am also grateful to my co-advisor, Professor Michele Brandão, for her support, encouragement, and friendship. She consistently brings great ideas and advice and has supported me in countless ways. I feel genuinely proud to have been mentored by two incredible women.

I also want to thank the friends and colleagues I met during my Ph.D. in the CS+X lab, who made the entire process lighter and more enjoyable. In particular, my heartfelt thanks go to my academic brothers, Gabriel P. Oliveira and Danilo Seufitelli. Danilo, the "older brother", with his calm and steady energy, made me believe that everything would work out in the end (*and he was right!*). And Gabriel, the "younger one", my daily companion, who endured all my complaints and crises and stood by me through both the best and the toughest moments. Without their support and friendship, none of this would have begun—nor ended.

Finally, I would like to thank UFMG, especially the Department of Computer

Science, its professors, and its staff, for providing an excellent educational and research environment. Throughout my Ph.D., I had the opportunity to travel to conferences and present the work developed in this dissertation, with essential support from the department. I also extend my gratitude to the research projects I was part of, the *Analytical Capabilities Program* and the *hDataPLM Project*, which allowed me to learn extensively and work alongside remarkable people.

Thanks to everyone who directly or indirectly contributed to this work.

“*A word after a word after a word is power.*”
(Margaret Atwood)

Resumo

A literatura tem operado como um meio através do qual as sociedades reproduzem e transformam vieses sociais, sendo o viés de gênero um dos mais persistentes. Tal viés é linguisticamente codificado nas narrativas por meio de padrões recorrentes, incluindo adjetivos, verbos e estruturas sintáticas que definem como os personagens são descritos, reforçando, assim, estereótipos culturais de feminilidade e masculinidade. Enquanto a crítica literária tradicional, fundamentada na *close reading*, oferece insights interpretativos detalhados, ela permanece limitada em escopo e escalabilidade. Apesar dos avanços na interseção entre estudos de gênero e linguística computacional, a pesquisa existente continua concentrada no inglês e em outras línguas de grande disponibilidade de recursos. A pesquisa sobre textos literários em português é consideravelmente sub-representada em comparação com outras línguas, tanto na disponibilidade de recursos computacionais quanto em análises em larga escala da representação de gênero. Para preencher essas lacunas de pesquisa, esta tese propõe um *framework* computacional para medir e analisar o viés de gênero em textos literários de língua portuguesa. O *framework* integra a análise literária interpretativa com métodos computacionais e Processamento de Linguagem Natural (PLN) dentro de um paradigma de *distant reading*, combinando etapas como identificação de personagens, inferência de gênero e quantificação de viés. Além da investigação literária, compreender como o viés é linguisticamente codificado é crucial para o desenvolvimento ético de modelos de linguagem, que frequentemente herdam assimetrias históricas e culturais a partir de seus dados de treinamento. Ao revelar padrões sistemáticos de gênero em diferentes períodos históricos e gêneros literários, este trabalho contribui para os estudos literários e avança a pesquisa em PLN ao fornecer um *framework* replicável e recursos para a análise de viés de gênero em textos literários em português.

Palavras-chave: viés de gênero; processamento de linguagem natural; mineração de texto; análise literária; literatura em língua portuguesa; leitura à distância.

Abstract

Literature has long operated as a medium through which societies reproduce and transform social biases, with gender bias being one of the most pervasive. Such bias is linguistically encoded in narratives through recurring patterns, including adjectives, verbs, and syntactic structures that define how characters are described, thereby reinforcing cultural stereotypes of femininity and masculinity. While traditional literary criticism grounded in *close reading* offers nuanced interpretive insights, it remains limited in scope and scalability. Despite advances at the intersection of gender studies and computational linguistics, existing research remains concentrated on English and other high-resource languages. Research on literary texts in Portuguese is considerably underrepresented compared to other languages, both in the availability of computational resources and in large-scale analyses of gendered representation. To bridge these research gaps, this dissertation proposes a computational framework for measuring and analyzing gender bias in Portuguese-language literary texts. Our framework integrates interpretive literary analysis with computational methods and Natural Language Processing (NLP) within a *distant reading* paradigm, combining stages such as character identification, gender inference and bias quantification. Beyond literary inquiry, understanding how bias is linguistically encoded is crucial for the ethical development of language models, which frequently inherit historical and cultural asymmetries from their training data. By revealing systematic gendered patterns across historical periods and literary genres, this work contributes to literary scholarship and advances NLP research by providing a replicable framework and resources for gender bias analysis in Portuguese-language literary texts.

Keywords: gender bias; natural language processing; text mining; literary analysis; literature in Portuguese; distant reading.

List of Figures

List of Tables

List of Acronyms

BERT Bidirectional Encoder Representations from Transformers. 120

DisCo Discovery of Correlations. 34

DR Demographic Representation. 36

GBPCs Gendered Body Part Clusters. 47

LLMs Large Language Models. 22, 30, 79, 86, 90, 120, 126, 127, 136, 137, 145

LMB Language Model Bias. 35

LPBS Log-Probability Bias Score. 34, 35, 123

MLM Masked Language Modeling. 74, 120–122

NER Named Entity Recognition. 22, 23, 42, 59, 61, 62, 67, 70–72, 74, 75, 77, 79, 80, 82, 84, 94, 95, 140–142, 146, 148, 188, 189, 199–203, 205, 206, 208

NLP Natural Language Processing. 9, 21, 22, 24–28, 30, 37, 38, 40, 56–58, 61, 67, 71, 79, 96, 126, 137, 139, 142, 145, 180, 182, 186, 187, 196, 199, 201, 202

PLL Pseudo-log-likelihood. 35

PLN Processamento de Linguagem Natural. 8

PMI Pointwise Mutual Information. 32

SBS Sentence Bias Score. 34

SMS Systematic Mapping Study. 37, 45, 175

SRL Semantic Role Labeling. 45

WEAT Word Embedding Association Test. 33, 34, 51, 53, 66, 123, 130, 131, 134

Contents

Chapter 1

Introduction

Literature is not only an artistic expression but also a cultural system through which societies narrate, reproduce, and challenge their own values [1]. As Eagleton [2] observes, literary works are embedded within ideological and historical frameworks that shape their production and interpretation. Far from being neutral, literature participates in the construction of worldviews that reflect and sustain social hierarchies, often encoding and reinforcing norms related to gender, race, and class [3].

In this sense, literature functions both as a reflection of societal norms and as a means of shaping cultural perceptions [1, 4, 5]. The stories told and the voices silenced are not random choices, but reflections of who has historically been entitled to speak and be represented [5, p.29]. Indeed, feminist literary criticism has long emphasized that canonical traditions tend to privilege male authors and perspectives, rendering women's experiences marginal or stereotyped [6, 7].

In Brazil, literature has played a crucial role in building social and national identities [8]. The writings of Machado de Assis, Clarice Lispector, and Graciliano Ramos provide clear examples of how literature engages with questions of subjectivity, class, and gender. In *Memórias Póstumas de Brás Cubas*, Machado de Assis criticizes patriarchal and classist structures; in *A Hora da Estrela*, Clarice Lispector foregrounds the subjectivity and struggles of female protagonists under socially constraining norms; and in *Vidas Secas*, Graciliano Ramos portrays the intersection of poverty and gender roles in the Brazilian Northeast.

Across these and many other examples, gender bias emerges not merely as a matter of representation but as a structural feature of narrative discourse. How female and male characters are described, the actions they are allowed to perform, and the emotions they express are shaped by deep cultural assumptions about gender [9–11]. Even when subtle, these asymmetries remain across centuries of literary production, creating enduring patterns that can shape readers' perceptions and cultural norms [12, 13].

In the broader social context, addressing gender bias in literary texts provokes a more equitable cultural landscape. By making visible the gendered patterns embedded in texts, research can challenge inherited stereotypes and foster critical engagement with cultural production. Moreover, this understanding urges the responsible development of

Natural Language Processing (NLP) models, ensuring that computational systems trained on cultural data do not reproduce inequalities that literature has historically encoded.

During the development of this research, the field of Natural Language Processing underwent rapid and significant transformations, particularly with the emergence of large-scale pre-trained and instruction-tuned language models. These advances have reshaped the landscape of text analysis, enabling tasks that were previously unattainable or impractical, especially for high-resource languages. At the same time, they have introduced new challenges related to interpretability, control, reproducibility, and the amplification of social biases.

This Ph.D. dissertation was conceived and developed in a context where such models were either unavailable, immature, or poorly supported for Portuguese. Consequently, many of the methodological choices made throughout this work reflect the state of the field at that time, prioritizing transparency, modularity, and analytical control over black-box performance. While contemporary models may now perform some of the proposed tasks more directly, the framework introduced here remains relevant by offering **explicit modeling decisions**, **interpretable metrics**, and a **systematic perspective** on how gender bias manifests across textual, representational, and generative levels.

Rather than being superseded by recent advances, the contributions of this dissertation provide a structured foundation upon which newer models can be critically evaluated, adapted, and audited. In this sense, the framework is designed not as a closed solution, but as a methodological baseline that remains applicable as language technologies continue to evolve.

1.1 Motivation

Literature has long acted as a means of representing and interrogating the structures of human experience. Within literary texts, linguistic and narrative choices establish symbolic hierarchies that define who speaks, how they are described, and which perspectives are granted legitimacy. These mechanisms make literature a key ground for understanding how social biases are reproduced and transformed over time [2, 5, 7]. Among the various forms of inequality embedded in cultural discourse, *gender bias* is one of the most pervasive, shaping both literary representation and critical reception across centuries [6].

In literary texts, gender bias is manifested through language, including the adjectives used to describe characters, the verbs that define their agency, and the semantic associations that link gender to emotions, morality, or power [14, 15]. As Tavares and Quixadá [9] argue, for instance, gendered patterns and morphosyntactic structures often reinforce cultural stereotypes, reproducing traditional expectations of femininity and masculinity. Such patterns persist across literary traditions, shaping readers' perceptions and leading to the reinforcement and cultural normalization [13, 16].

The study of gender bias in literature has a long tradition in the humanities, especially within feminist literary criticism [6, 7]. However, these analyses have historically relied on *close reading*—the detailed, interpretive examination of selected texts. While such an approach offers rich insight into narrative form and ideology, it is inherently limited in scope. The rise of computational methods has introduced complementary paradigms such as *distant reading* [17], which enable the large-scale examination of literary corpora through quantitative and linguistic modeling. This shift does not replace human interpretation but extends it, revealing recurring lexical, syntactic, and thematic patterns that might remain invisible to traditional methods [18–20].

Natural Language Processing provides the tools to operationalize such analyses [21]. By processing large collections of texts, NLP methods can extract features related to character representation, lexical associations, and dependency relations that reflect underlying gender asymmetries. Techniques such as Named Entity Recognition (NER), coreference resolution, and dependency parsing allow the identification of narrative agents, their grammatical roles, and the semantic relations among them [22]. When applied systematically, these tools enable the quantification of linguistic bias and investigate how it evolves across time, genres, and authors [15, 23, 24].

The relevance of computationally investigating gender bias extends beyond literary studies. In the field of Computer Science, similar concerns have emerged regarding the fairness of machine learning systems, particularly Large Language Models (LLMs) [25]. These models are trained on vast text corpora that often include literary and cultural materials, inheriting the biases present in their sources [26–28]. Understanding how bias is linguistically encoded in literature thus informs both cultural analysis and the ethical development of computational models trained on human language [29, 30].

Despite the growing intersection between gender studies and computational linguistics, most existing research focuses on English or other high-resource languages [31]. In contrast, Portuguese remains underrepresented in the development of linguistic resources and tools for bias analysis. Few studies have systematically explored gendered patterns in Portuguese-language literary texts [10, 32]. Addressing such gaps is essential to ensure that computational models and cultural analyses reflect the linguistic diversity of the communities they aim to represent.

This Ph.D. dissertation tackles such challenges by proposing a framework for measuring and analyzing gender bias in Portuguese-language literary texts. To the best of our knowledge, this is the first framework specifically designed for the large-scale and systematic analysis of gender bias in Portuguese-language literary texts. By integrating literary interpretation with NLP-based modeling, the proposed framework advances digital humanities through large-scale cultural analysis, while simultaneously expanding computational linguistics research beyond English-centric bias studies.

1.2 Research Goals and Contributions

The overall goal of this dissertation is:

> *To develop a computational framework for automatically measuring and analyzing gender bias in Portuguese-language literary texts.*

To achieve this goal, the dissertation begins with a systematic mapping of related work, which constitutes the first stage of the research (Chapter 2). This review explores existing approaches, resources, and methods for gender bias analysis and assessment across digital humanities and computational linguistics. From this mapping, we identify four main challenges that motivate and guide the development of the proposed framework. These challenges are addressed through four research goals (**RGs**), each corresponding to a major chapter of the dissertation, as follows.

[RG1] Develop a computational framework for measuring and analyzing gender bias in Portuguese-language literary texts. This research goal focuses on designing a flexible and modular framework that integrates linguistic and computational techniques for gender bias analysis. The framework is structured around three analytical dimensions: *Character Identification*, *Gender Inference*, and *Gender Bias Measuring* (Chapter 3). **Key results.** The direct outcome of this research goal is the conference paper [33], which presents the definition and evaluation of the developed framework.

[RG2] Develop and evaluate domain-adaptive strategies for identifying literary characters in Portuguese-language texts. This research goal addresses the challenges of recognizing literary characters in Portuguese-language literary texts. It involves the creation of annotated corpora, the design and evaluation of pre-training and fine-tuning Named Entity Recognition models. The main goal is to enhance the precision and robustness of character identification, which is a critical foundation for subsequent bias analyses (Chapter 4). **Key results.** The direct outcomes of this research goal include three publications: [34], which introduces domain-adapted NER models for identifying entities—particularly characters—in Portuguese-language literary texts; [35], an extended version of the previous paper with additional analyses and detailed result analyses; and [36], which presents the annotated corpus that is developed to train the models.

[RG3] Develop and evaluate an automated approach for gender inference in Portuguese-language literary texts. This research goal focuses on the gender inference task, which assigns gender labels to literary characters in a scalable and accurate manner. It combines name-based heuristics with contextual linguistic cues. The approach is evaluated quantitatively and qualitatively to ensure high accuracy and robustness, enabling reliable downstream analyses of gender bias (Chapter 5). **Key results.** The direct outcomes of this research goal include two publications: [37] and [38], which apply the proposed gender inference heuristic to quantitatively analyse descriptions of male and female body parts in literary works in Portuguese.

[RG4] Measure and analyze gender bias in both literary language and computational methods. This research goal investigates how gender is represented in Portuguese-language literary texts and how these representations are reflected or amplified in computational models. Analyses include quantifying disparities in character portrayal, identifying linguistic stereotypes, and evaluating bias in both word embeddings and generative language models. This comprehensive assessment provides insights into both textual and methodological sources of gender bias (Chapter 6). **Key results.** The direct outcomes of this research goal include three publications: [33], which evaluates gender bias using lexical features; [39], which investigates gender bias in generative language models; and [40], which examines gender bias in Portuguese word embeddings.

1.3 Organization

The remainder of this dissertation is organized as follows.

- **Chapter 2** presents related work on interdisciplinary approaches to gender bias assessment in literature and computational linguistics.
- **Chapter 3** introduces the computational framework for measuring and analyzing gender bias in literary works. The chapter details its modular design, comprising three analytical dimensions: *Character Identification*, *Gender Inference*, and *Gender Bias Measuring*.
- **Chapter 4** focuses on the identification of literary characters in Portuguese-language literary texts. It presents the creation of an annotated corpus, the design of domain-adaptive pre-training and fine-tuning strategies for transformer-based models, and a comprehensive evaluation of these approaches.
- **Chapter 5** details the development and evaluation of an automated approach for inferring the gender of named entities in Portuguese-language literary texts.
- **Chapter 6** presents the measurement and analysis of gender bias in both literary language and computational methods. The chapter examines disparities in character portrayal, linguistic patterns, and stereotype associations, as well as how these biases are captured and potentially amplified by embeddings and generative language models. Findings provide a comprehensive view of gender bias in Portuguese-language literature and its propagation through NLP models.
- **Chapter 7** concludes the dissertation by summarizing the main contributions, reflecting on the potential limitations, and outlining future research directions.

Chapter 2

Related Work

This dissertation has an inherently interdisciplinary nature, located within Computer Science while drawing upon insights from literary and gender studies. From the computational perspective, the primary focus is on developing methods, resources, and a framework for processing textual data to measure and analyze gender bias in literary works. Literary theory and gender studies, in turn, provide the critical context necessary to interpret the patterns uncovered through computational methods.

To establish a coherent foundation for the research, this chapter introduces key concepts that are essential for understanding the intersections of these domains. Section 2.1 defines concepts that are relevant for the analysis of textual gender bias. Section 2.2 reviews prior work on computational approaches that have been applied to assess gender bias in Natural Language Processing. Section 2.3 focuses on gender bias in literary texts, mapping studies on gender bias in literature. Finally, Section 2.4 summarizes related work, discussing existing gaps and opportunities, and motivating the development of the framework proposed in this dissertation.

2.1 Fundamental Concepts

This dissertation lies at the intersection of Computer Science and Digital Humanities. Therefore, in this section, we present concepts particularly relevant for understanding both the computational and literary aspects of gender bias assessment.

Gender vs. Sex. In this dissertation, *gender* is understood as a socially constructed category formed by roles, behaviors, and identities of women, men, and gender-diverse individuals within specific historical and cultural contexts, which may vary across societies and over time [41, 42]. Hence, we acknowledge that gender is *separate* from biological sex, which generally refers to biological attributes associated with physical and physiological features (e.g., chromosomal genotype, hormonal levels, internal and external anatomy). We also recognize gender as a multifaceted concept that includes a spectrum of identities and experiences beyond the traditional binary categories of male and female.

Gender Bias. Following prior studies in the field [24, 43–45], this dissertation defines *gender bias* as the unequal treatment, misrepresentation, linguistic asymmetry, or stereotyping of individuals based on gender. Although *gender stereotypes* and *gender prejudice* are closely related to gender bias, they differ in scope and manifestation. *Gender stereotypes* involve generalized beliefs or expectations about traits, behaviors, or social roles typically associated with a particular gender [46], whereas *gender prejudice* refers to negative attitudes or judgments based solely on gender, which may lead to biased representation or even the marginalization of specific individuals or groups in a text [16].

Close and Distant Readings. *Close reading* is a traditional literary method involving detailed, interpretive analysis of specific texts or passages [47]. It provides insight into narrative structures, character development, and ideological content, but it is limited in scale. In contrast, *distant reading* [17] uses computational techniques to analyze large volumes of text, uncovering patterns in language use, character representation, and thematic structures that may not be apparent through close reading. This dissertation leverages distant reading via NLP methods to complement literary interpretation and quantify gender bias. This complementary perspective underlies the design of the computational framework proposed in this dissertation.

Literary Works. In this dissertation, *literary works* are understood as creative written expressions that employ language as an artistic and communicative medium to explore human experience, emotion, and social meaning. They can take many forms, including novels, poems, short stories, plays, and other forms of creative writing. Beyond their aesthetic dimension, literary works engage with the social and historical contexts from which they emerge, reflecting and shaping cultural values, ideologies, and power relations [1, 2, 5]. As Showalter [3] argues, the notion of literature itself has often been defined through gendered and exclusionary perspectives, reinforcing the importance of re-examining what and whose voices are represented in the literary canon. In this sense, literature is considered both an art form and a sociohistorical document that not only mirrors reality but also participates in its construction.

Literary Characters. The notion of *character* in literary studies remains conceptually diverse. As Bamman et al. [48] note, scholars differ on whether characters should be understood as representations of real or imagined persons (referential view) or as structural functions within the narrative (formalist view). In the widest sense, a *character* can be defined as any entity—individual or collective, human or human-like—participating in the storyworld [49]. This broad definition may include not only human characters, but also animals, collective entities (e.g., social groups or peoples), and, in some literary traditions, abstract or inanimate entities that are explicitly personified (such as death, time, or solitude). Following this inclusive perspective, in this dissertation, literary characters are operationally defined as narrative entities that play an identifiable role in the narrative or are explicitly referenced within it. From a computational standpoint, the proposed framework does not impose ontological distinctions between human, non-human, or per-

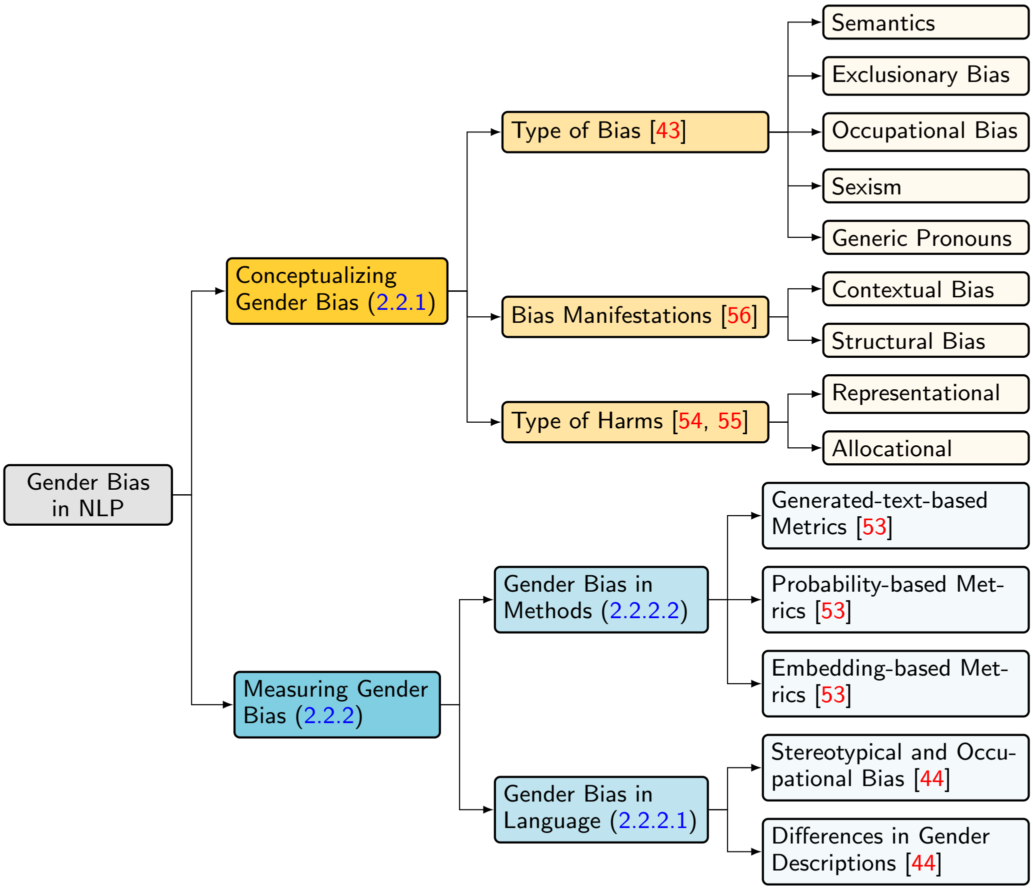


Figure 2.1: Conceptual map of gender bias in NLP – sections that explain each category are presented between parentheses.

sonified entities; instead, it treats them uniformly as narrative entities, whose relevance emerges from their linguistic realization and narrative function. This definition aligns with previous approaches to character identification in literature [50–52] and provides a flexible foundation for the modeling decisions adopted in this dissertation.

2.2 Gender Bias in Natural Language Processing

Building upon the previous discussion of how gendered dynamics emerge in language and discourse, this section turns to computational perspectives, exploring how bias is conceptualized and measured within Natural Language Processing. To guide this review of related work, we synthesized insights from major survey papers published over the past five years that address gender bias and related fairness concerns in language technologies. These works collectively provide a structured understanding of how bias arises in data and models, how it can be measured, and how the conceptualization of bias has evolved alongside advances in language models. Figure 2.1 provides an overview of how this section is organized, connecting conceptual foundations with measurement frameworks.

2.2.1 Conceptualizing Gender Bias

As Natural Language Processing systems become increasingly integrated into decision-making solutions across domains (such as healthcare, recruitment, criminal justice, and other high-stakes contexts), understanding how bias manifests and propagates through these systems has become a central concern [57]. Yet, conceptualizing bias remains one of the most persistent challenges in this field. Blodgett et al. [55] reviewed 146 studies addressing "bias" in NLP and observed that many lacked conceptual rigor: motivations were often vague or inconsistent, and few studies articulated explicit normative reasoning about why specific system behaviors were harmful, in what ways, and to whom. Moreover, most reviewed works failed to engage with relevant literature outside NLP, particularly from the social sciences and linguistics.

To clarify the types of harms that bias can produce, Blodgett et al. [55] draw on the taxonomy proposed by Crawford [54], which distinguishes between *allocational* and *representational* harms. *Allocational harms* occur when automated systems unfairly affect the distribution of resources or opportunities across social groups. For example, language models used in hiring processes that disproportionately favor one gender. These harms are economic or material in nature, reflecting situations where systems inequitably allocate resources among groups [44, 58]. In contrast, *representational harms* arise when systems perpetuate negative stereotypes, reinforce normative assumptions, or misrepresent certain social groups. They include the propagation of biased associations, the amplification of stereotypes, and the omission or underrepresentation of marginalized identities in datasets.

In the context of gender bias in NLP, *allocational harm* typically emerges when models perform better on data associated with the majority gender [44, 58]. For example, Prates et al. [59] show that Google Translate systematically defaults to masculine forms, especially in domains characterized by unequal gender representation or stereotypically masculine connotations, such as STEM (Science, Technology, Engineering, and Mathematics) professions. Such disparities often reflect imbalanced training corpora, in which male voices, names, and narratives are overrepresented, resulting in systems that systematically privilege one gender over others.

Conversely, *representational harm* refers to the ways gendered associations are encoded in linguistic representations or model parameters [44, 58]. Models may, for instance, associate words such as *nurse* or *teacher* with female identities and *engineer* or *leader* with male ones, thereby reinforcing traditional gender stereotypes [26, 60]. Although these biases do not directly define material outcomes, they shape perceptions of gender roles and attributes, influencing how individuals and groups are represented in language.

For both types of harms, biased data are a central source of the problem [25, 61],[1]

[1]Biases may also stem from model design choices, training objectives, or deployment contexts. Data sampling, preprocessing, and weighting can privilege particular linguistic forms or demographic groups, while evaluation and real-world deployment choices can further exacerbate disparities [61].

since language models learn directly from human-produced text. These data inherently reflect the social, cultural, and historical contexts in which they were created, including existing prejudices and inequalities. Consequently, gendered asymmetries in source data are translated into algorithmic representations, allowing models to reflect, reproduce, or even amplify societal biases [28, 44, 53].

Textual data, whether drawn from literature, journalism, or social media, encode how gendered meanings are built and reproduced through language. From this perspective, gender bias in text can be understood as the presence of linguistic or representational patterns—lexical, syntactic, or semantic—that implicitly or explicitly convey societal stereotypes or prejudices toward a particular gender [43, 56]. Building on this view, Hitti et al. [56] introduced a taxonomy of gender bias, distinguishing between *structural* and *contextual* manifestations, or combinations of both:

- **Structural Bias** arises from grammatical constructions or syntactic patterns that impose gender assumptions in otherwise neutral contexts. Structural bias can be divided into *gender generalization*, when a gender-neutral term is syntactically referred to by a gender-exclusive pronoun, therefore, making an assumption of gender; and *explicit marking of sex*, which appears with the use of gender-exclusive keywords when referring to an unknown gender-neutral entity or group.

- **Contextual Bias** manifests itself in a tone, the words used, or the context of a sentence. Unlike structural bias, this type of bias cannot be observed through grammatical structure but requires contextual background information and human perception. Contextual bias can be divided into *societal stereotype*, which showcases traditional gender roles that reflect social norms; and *behavioral stereotype*, which contains attributes and traits used to describe a specific person or gender.

While Hitti et al. [56] offer a structural–contextual distinction that focuses primarily on grammatical and semantic cues, Doughman et al. [43] expand this view by situating linguistic manifestations of bias within their broader social implications, proposing five interrelated types of gender bias:

- **Generic Pronouns**, which refer to the use of gendered pronouns or nouns to represent individuals of unspecified gender (e.g., *generic he, generic she, gendered generic man*). Such expressions contribute to biased mental imagery by implicitly privileging one gender as the default referent.

- **Sexism**, containing both *hostile* and *benevolent* forms. Hostile sexism manifests through explicitly negative portrayals of women as less competent or emotional, while benevolent sexism reinforces traditional gender hierarchies through seemingly positive but patronizing depictions of women as caring, innocent, and in need of men's protection.

- **Occupational Bias**, which arises when professions or social roles are stereotypically associated with one gender (e.g., "doctors are men, and nurses are women"), reflecting and reinforcing the sexual division of labor and leading to gender disparities in workforce participation.

- **Exclusionary Bias**, covering linguistic forms that explicitly or implicitly exclude gender diversity, such as gender-exclusive terms (*chairman, mankind*), gender-based neologisms (*man-bread, man-sip*), and gendered word ordering (*men and women*), all of which produce representational harms.

- **Semantics**, which includes metaphorical, lexical, or proverbial expressions that encode sexist meanings (e.g., referring to women as animals or food, or sayings such as "a woman's tongue three inches long can kill a man six feet high").

Taken together, these taxonomies illustrate that gender bias in text is not merely a computational artifact but a linguistic and cultural phenomenon that becomes observable and measurable through NLP methods. Bias can manifest through subtle asymmetries or overtly stereotypical expressions, posing enduring challenges for detection and mitigation [44]. Texts embody the social meanings and power relations through which gender is represented, and therefore function as both the primary input and the central object of analysis in this PhD dissertation.

2.2.2 Measuring Gender Bias

Measuring gender bias in natural language systems entails translating complex social phenomena into quantifiable computational constructs. This translation is not merely technical but epistemological: it determines which aspects of bias become visible to computational analysis and which remain invisible. Related work includes a wide range of evaluation approaches, varying in scope, granularity, and in the linguistic or representational layer at which bias is observed. Early works in bias measurement focused on explicit co-occurrence patterns within texts, while recent studies have formalized multi-level taxonomies for measuring bias in language models [44, 53, 83, 84]. Table 2.1 synthesizes the main approaches and serves as a guide for the discussion that follows.

Stanczak and Augenstein [44] distinguish between two principal domains of bias measurement: *bias in language* and *bias in methods*. Bias in language is typically measured through statistical associations that reveal how gendered expressions are represented or described, whereas bias in methods concerns the behavior of computational models trained on such data. Gallegos et al. [53] extend this discussion by organizing bias evaluation techniques, particularly those focused on LLMs, into a three-tiered taxonomy based on the computational layer at which bias is observed: *embedding-based metrics*,

Table 2.1: Summary of gender bias measurement approaches in NLP.

Category	**Metric**	**Description**
	Gender Bias in Language	
Differences in Gender Descriptions	PMI [20, 62, 63]	Measures asymmetric co-occurrence patterns between gendered terms and descriptive attributes.
Stereotypical and Occupational Bias	Gender stereotype score [64–66]	Captures biased associations between gender and stereotypical occupations or roles.
	Gender Bias in Methods	
Embedding-based Metrics	Direct bias [26]	Quantifies how much gender-neutral word vectors align with a gender direction in embedding space.
	WEAT [67]	Computes association strength between target and attribute word sets using cosine similarity.
	WEAT* [68]	Adaptation of WEAT for contextual embeddings, analyzing token-level bias in sentence contexts.
	SWEAT [69], CEAT [70]	Sentence-level variants of WEAT measuring contextual associations in full-sentence embeddings.
	SBS [71]	Aggregates gender-related semantic information across sentence embeddings.
Probability-based Metrics	DisCo [72]	Detects gender-specific continuations favored by the model via masked-token probability disparities.
	LPBS [73]	Computes log-probability gaps of gendered word predictions in controlled contexts.
	CBScore [74]	Measures the conditional probability of generating gendered terms given occupation or role contexts.
	PLL [53]	Compares pseudo-log-likelihood of gendered sentence variants to detect model preference.
	CrowS-Pairs [75], CAT [76], AUL/AULA [77]	Sentence-pair tests comparing stereotypical vs. anti-stereotypical continuations.
	LMB [78]	Likelihood-based metric capturing reinforcement of gender-stereotypical associations.
Generated-text-based Metrics	Co-Occurrence Bias Score [65]	Identifies asymmetric gender–attribute or gender–occupation co-occurrence patterns in generated text.
	Demographic Representation [79]	Measures frequency and prominence of gendered entities in generated outputs.
	Marked Persons framework [80]	Evaluates disparities in how gendered groups are referenced in model outputs.
	Classifier-based metrics [53]	Captures differences in sentiment, toxicity, or social regard by gender.
	HONEST [81]	Lexicon-based measure of harmful or offensive content targeting gendered entities.
	Gender Polarity [82]	Measures asymmetric sentiment or evaluative polarity for masculine vs. feminine prompts.

probability-based metrics, and *generated-text-based metrics*. Other surveys have proposed alternative but conceptually similar classifications [83, 84].

Following this taxonomy, we organize the main bias measurement methods into two complementary dimensions: those that focus on *gender bias in language*, observable directly in corpora and linguistic representations, and those that measure *gender bias in methods*, observable in how models internalize and reproduce such patterns. Taken together, these frameworks provide complementary perspectives that can be integrated into a unified view of gender bias measurement.

2.2.2.1 Gender Bias in Language

Bias in language primarily manifests in textual representations and can be identified through linguistic and extra-linguistic cues, such as lexical associations, co-occurrence patterns, or syntactic structures.

Differences in Gender Descriptions. Differences in depictions of men and women have been quantified using the Pointwise Mutual Information (PMI), a widely used association measure in distributional semantics [44]. PMI measures the strength of co-occurrence between gendered terms (e.g., *man*, *woman*) and descriptive attributes (e.g., *strong*, *beautiful*). Formally, it is defined as:

$$PMI(gender, \mathbf{word}) = \ln\left(\frac{P(gender, \mathbf{word})}{P(gender)\,P(\mathbf{word})}\right)$$

A higher PMI score indicates a stronger associative link between a gendered term and a given attribute. This measure has been applied to identify stereotypical lexical patterns across adjectives, verbs, and thematic contexts involving male and female entities in large textual datasets [20, 62, 63]. However, Rudinger et al. [62] caution that co-occurrence-based metrics, while interpretable and model-agnostic, risk overgeneralizing or obscuring context-specific meanings, as PMI captures only surface-level associations.

Stereotypical and Occupational Bias. Stereotypical or occupational bias has been quantified through the association between gendered terms and gender-neutral nouns referring to professions or social roles (e.g., *doctor*, *nurse*, *teacher*) [44]. Researchers compare co-occurrence frequencies or distributional similarities to derive a stereotype score, capturing how strongly a given occupation is linguistically gendered [64, 65]. For instance, Qian [66] defines the *gender stereotype score* of a word w as:

$$b(w) = \left| log \frac{c(w, m)}{c(w, f)} \right|,$$

where m and f denote sets of male- and female-associated words (e.g., *he, father, actor* vs. *she, woman, girl*), and $c(w, g)$ represents the number of times a gender-neutral word w co-occurs with gendered terms from group g within a fixed window of 10 tokens. A score of $b(w) = 0$ indicates perfectly balanced co-occurrence, while higher scores reflect stronger gender associations.

In summary, bias in language reflects systematic patterns in how gendered entities are described and represented within textual data. Measures such as PMI and gender stereotype scores reveal asymmetries in lexical associations, co-occurrence patterns, and occupational representations, highlighting both overt and subtle manifestations of gender bias. While these corpus-based metrics are essential for identifying representational disparities, they capture only surface-level signals and do not account for how such biases may be internalized or amplified by computational models. Consequently, understanding bias in language provides the foundational step for subsequent analyses of *bias in methods*, where these linguistic asymmetries are encoded, propagated, and potentially exacerbated through model representations and outputs.

2.2.2.2 Gender Bias in Methods

While bias in language reflects how gendered meanings are encoded in text, bias in methods captures how these meanings are internalized, amplified, or transformed through computational representations. In this sense, the focus shifts from what is written to how models learn, represent, and reproduce such associations.

Embedding-based Metrics. Embedding-based metrics quantify bias in the representational space of language models by analyzing geometric relations between gendered and gender-neutral terms. In their seminal study, Bolukbasi et al. [26] distinguished between two forms of bias in word embeddings: *direct* and *indirect*. The *direct bias* of a word embedding $\vec{w}$ is quantified as:

$$DirectBias_c = \frac{1}{|N|} \sum_{w \in N} \left| \cos(\vec{w}, g) \right|^c,$$

where N is the set of gender-neutral words, g represents the gender direction, and c controls the strictness of the bias definition. Direct bias captures the alignment of neutral words with the gender subspace. *Indirect bias*, in contrast, reflects latent associations among neutral words mediated by gender, which may influence seemingly unrelated semantic relations. For example, if neutral terms such as *businessman* and *genius* appear closer to *football* than to *nurse*, it signals latent gendered structures [85]. Subsequent research has highlighted that indirect bias is often neglected and not adequately mitigated by existing de-biasing methods.

To capture associations more systematically, Word Embedding Association Test (WEAT) [67] quantify the strength of association between two sets of target words (e.g., male–female) and two sets of attribute words (e.g., career–family). Formally, for protected attribute sets A_1 and A_2 and target sets W_1 and W_2, the test statistic is:

$$f(A_1, A_2, W_1, W_2) = \sum_{a_1 \in A_1} s(a_1, W_1, W_2) - \sum_{a_2 \in A_2} s(a_2, W_1, W_2),$$

where $s(a, W_1, W_2) = \text{mean}_{w_1 \in W_1} \cos(a, w_1) - \text{mean}_{w_2 \in W_2} \cos(a, w_2)$, and the effect size is computed as:

$$WEAT(A_1, A_2, W_1, W_2) = \frac{\text{mean}_{a_1 \in A_1} s(a_1, W_1, W_2) - \text{mean}_{a_2 \in A_2} s(a_2, W_1, W_2)}{\text{std}_{a \in A_1 \cup A_2} s(a, W_1, W_2)},$$

where a larger effect sizes indicate a stronger stereotypical association between gendered and attribute word sets. Dev et al. [68] later introduced WEAT*, a variant that replaces W_1 and W_2 with definitionally masculine and feminine words (e.g., *gentleman*, *matriarch*) to better capture gendered semantics. Contextualized embeddings motivated extensions such as *SEAT* [69] and *CEAT* [70], which estimate distributions of bias scores across diverse sentence embeddings rather than relying on fixed templates.

Finally, the Sentence Bias Score (SBS) [71] offers a complementary approach by aggregating word-level bias into a single sentence-level value. Given a sentence S and a list of gendered words A, the metric computes the cosine similarity between each token embedding $s \in S$ and a gender direction vector v_{gender} derived from the principal component of feminine–masculine word pairs. Each word-level bias is weighted by its semantic importance α_s, obtained from the number of times the sentence encoder's max-pooling operation selects the representation at position t:

$$SentenceBias(S) = \sum_{s \in S,\, s \notin A} \left| \cos(s, v_{\text{gender}}) \cdot \alpha_s \right| .$$

Probability-based Metrics. Probability-based metrics evaluate bias through a model's internal probability distributions, comparing how likely it is to produce gendered alternatives under equivalent contexts. These methods directly probe a model's decision surface, revealing disparities in the likelihood assigned to gendered or neutral terms in otherwise identical sentences. Gallegos et al. [53] categorize these methods into two main subgroups: *masked-token methods*, designed primarily for masked language models, and *pseudo-log-likelihood methods*, which are typically applied to autoregressive models.

Masked-token methods estimate token probabilities by masking specific words in a sentence and prompting the model to predict the missing term. Webster et al. [72] introduced the Discovery of Correlations (DisCo) framework, which compares model completions for template-based sentences such as "`[X] is [MASK]`" or "`[X] likes to [MASK]`". Each template contains two slots: the first is manually filled with a bias trigger representing a social group (e.g., gendered names or nouns), while the second is completed with the model's top-k predicted tokens. The bias score is then computed as the average number of differing predictions across social groups and templates, enabling systematic comparison of model behavior in controlled, minimal contexts.

A related approach, the Log-Probability Bias Score (LPBS) proposed by Kurita et al. [73], quantifies bias as the normalized difference in token likelihoods across gendered contexts. Using template-based sentences of the form "`[MASK] is a [NEUTRAL ATTRIBUTE]`", LPBS estimates a token's conditional probability p_a and normalizes it by its prior probability p_{prior}, obtained from a neutral template ("`[MASK] is a [MASK]`"). The resulting score captures asymmetries between social groups as:

$$LPBS(S) = \log \frac{p_{a_i}/p_{\text{prior}_i}}{p_{a_j}/p_{\text{prior}_j}},$$

where a_i and a_j denote group-specific tokens (e.g., *he* vs. *she*). A positive value indicates a stronger association between a neutral attribute and one social group over another. This metric has been widely adopted to measure occupation- or adjective-based stereotypes encoded in masked language models.

Expanding on LPBS, Ahn and Oh [74] proposed the *Categorical Bias Score (CB-Score)*, which generalizes the formulation to non-binary protected attributes. Instead of comparing only two social groups, it measures the variance of normalized log-probabilities across multiple categories. Given a set of templates T, attributes A, and predicted tokens N, CBScore is defined as:

$$CB_{score} = \frac{1}{|T||A|} \sum_{t \in T} \sum_{a \in A} \text{Var}_{n \in N} \left(\log \frac{p_a}{p_{\text{prior}}} \right).$$

Pseudo-log-likelihood (PLL) methods, in turn, estimate bias by approximating the probability of generating each token conditioned on all remaining words in a sentence. For a given sentence S, the pseudo-log-likelihood is defined as:

$$PLL(S) = \sum_{s \in S} \log P(s \mid S_{\setminus s}; \theta),$$

where $S_{\setminus s}$ denotes the sentence with the token s masked. This formulation enables evaluation of autoregressive models without explicit masking objectives. Based on PLL, CrowS-Pairs [75], CAT [76], and AUL/AULA [77] compare stereotypical and anti-stereotypical sentences, using relative likelihoods or perplexities to quantify bias.

Finally, pseudo-log-likelihood metrics are closely related to perplexity, as both reflect a model's confidence in generating or reconstructing text. Building on this relationship, Barikeri et al. [78] introduced the Language Model Bias (LMB) metric, which directly compares the mean perplexity PP$(\cdot)$ between biased statements S_1 and their counterfactual counterparts S_2 representing alternative social groups. After excluding outlier pairs with extremely high or low perplexity, LMB computes the t-value of a two-tailed Student's t-test between PP(S_1) and PP(S_2).

Generated-text-based Metrics. Generated-text-based metrics assess bias at the level of model outputs, focusing on linguistic behavior and potential social harms manifested in generated content. Three main families of such metrics are typically distinguished [53]: *distribution-based metrics*, *classifier-based metrics*, and *lexicon-based metrics*. These methods capture *behavioral bias*—bias revealed through generation rather than internal model representations—and are frequently used for extrinsic evaluations of generative systems. However, they often rely on external resources, such as classifiers or lexicons, whose own fairness and coverage can affect reliability.

Distribution-based metrics quantify bias by comparing the distribution of generated tokens across social groups. One popular example, is the *Co-Occurrence Bias Score* [65] that measures the asymmetry in co-occurrence probabilities of a token w with gendered word sets A_i and A_j:

$$\text{Co-Occurrence Bias Score}(w) = \log \frac{P(w \mid A_i)}{P(w \mid A_j)},$$

where a score near zero denotes parity between feminine and masculine contexts. Expanding beyond pairwise co-occurrence, Demographic Representation (DR) [79] measures the aggregate frequency of social group mentions in generated outputs, comparing these frequencies to a reference or expected distribution. Building on this idea, Cheng et al. [80] introduced the *Marked Persons* framework, which leverages linguistic theories of markedness to identify biases in descriptions of marginalized groups. After prompting a model to generate personas of specified identities, their method identifies lexical markers that statistically distinguish marked from unmarked group descriptions.

Classifier-based metrics quantify bias using external models that evaluate generated outputs for attributes such as sentiment, toxicity, or regard. Bias is detected when outputs generated from semantically equivalent prompts, differing only by social group references, receive systematically different classifier scores [53]. Examples include *Perspective API*-based toxicity evaluations, *Score Parity*, and *Regard Score*, which quantify differences in perceived offensiveness or respectfulness across demographic groups. While these methods provide interpretable behavioral indicators, their results can be affected by the biases inherent in the classifiers themselves.

Lexicon-based metrics, in contrast, operate at the word level, comparing generated tokens to curated lexical resources that encode prior knowledge about bias, toxicity, or sentiment. The *HONEST* metric [81] quantifies the proportion of harmful or offensive terms produced by a model based on a lexicon of identity-related slurs. Similarly, *Gender Polarity* [82] computes polarity scores over gendered words to measure asymmetries in word choice and sentiment toward different gender groups. These approaches are simple to compute and interpretable but depend heavily on the completeness and cultural coverage of the underlying lexicons.

Overall, embedding-, probability-, and generation-based methods provide complementary perspectives on gender bias in natural language systems. Corpus-based analyses reveal representational asymmetries, embedding and probability metrics trace how these asymmetries are encoded and amplified, and generation-based methods capture real-world behavioral manifestations. Integrating these approaches allows for a multi-layered understanding of bias propagation, bridging linguistic evidence and computational evaluation, and informing mitigation strategies across model architectures and applications.

Although a wide range of bias measurement methods has been proposed in prior work, not all of them are equally applicable to literary texts or to the research goals addressed in this dissertation. In particular, several methods rely on controlled templates, task-specific prompts, or external resources that are poorly suited to long-form, stylistically diverse literary narratives. For this reason, the analyses developed in the following chapters focus on a subset of linguistically grounded and interpretable metrics that can be systematically applied to literary corpora, while remaining compatible with computational modeling.

2.3 Gender Bias in Literary Texts

Following the discussion of how gender bias has been conceptualized and measured in NLP, this section turns to its manifestation within literary contexts. To provide a systematic overview of existing research in this area, we conduct a Systematic Mapping Study (SMS) —see Appendix A for full results, by collecting and categorizing studies that address gender bias in literary corpora across languages, genres, and historical periods. The SMS is designed to identify how gender bias has been conceptualized, what methodological approaches have been employed to detect or measure it, and which dimensions of this phenomenon remain underexplored.

Using the PRISMA-ScR guidelines [86], the study involves defining research questions, implementing a keyword-based search across multiple databases (SCOPUS,[2] DBLP,[3] SciELO[4]), applying inclusion and exclusion criteria, and systematically screening and charting relevant studies. The search window for this mapping spans publications from 1990 through June 30, 2024. This process resulted in a curated set of 56 papers, which are analyzed to uncover trends in conceptualization, methodological practice, and research gaps. The complete protocol, including search strings, selection criteria, and data extraction procedures, is detailed in Appendix A.1. Next, we summarize these findings to highlight remaining gaps and opportunities, which are discussed at the end of this chapter.

2.3.1 Descriptive Mapping of Studies

The analysis of gender bias has received increasing attention across various domains, such as film, music, and social media, frequently addressing stereotypes, character roles, and other forms of representational inequality [16, 87–94]. In literature, interest in this topic has intensified since the 1990s, with studies focusing on textual features, author gender, and character representation, as detailed in Table A.2 [10, 32, 95–98]. Research trends on

[2]SCOPUS: https://www.scopus.com/
[3]DBLP: https://dblp.org/
[4]SciELO: https://www.scielo.br/

gender in literature can be analyzed across five dimensions: temporal, linguistic, literary genres, gender definition, and analytical methods.

Temporal Trends. Research interest in gender bias within literary studies has grown steadily over the years, particularly from the early 2000s onward, driven by increasing societal awareness of gender issues, the emergence of digital humanities, and the development of advanced computational tools for large-scale literary analysis [99, 100]. Early studies relied on small, manually curated datasets and close reading, which limited the scale of analysis. With the expansion of digital literary collections and the adoption of automated text analysis methods, researchers have been able to study much larger corpora, yielding statistically robust insights [17]. This growing incorporation of distant reading approaches, alongside the continued predominance of close reading in literary studies, is reflected in the increasing average size of analyzed corpora and the more frequent use of NLP techniques, enabling discoveries that were previously infeasible through manual analysis alone (see Figures A.3 and A.4).

Language. Most studies on gender bias in literature focus on English-language works (approximately 80%), leaving significant gaps in non-English contexts, including Portuguese [32, 50]. Portuguese-language literature offers a unique perspective shaped by distinct historical, social, and cultural contexts. Research in this domain, although limited, shows that authors are predominantly white men with higher education levels, while female characters often reflect traditional stereotypes or, in contemporary works, challenge societal norms in nuanced ways [10, 32, 96–98].

Due to the historical scarcity of NLP tools and annotated corpora for Portuguese, earlier studies relied primarily on manual close reading of small datasets, which provided detailed but labor-intensive analyses [97, 98]. Recent advances in digital humanities and NLP, along with the development of annotated Portuguese corpora [10, 32], enable automated large-scale analysis, revealing patterns in how male and female characters are described across social, emotional, physical, and character traits.

Literary Genres. Research on gender bias in literature mostly focuses on fiction (93%), particularly children's literature and novels, while drama and poetry remain marginally explored—5% and 2%, respectively (see Figure A.5). The emphasis on fiction reflects its well-defined narrative structures—characters, plots, and dialogue—which provide rich material for analyzing representational patterns. Within fiction, children's literature has been the most studied sub-genre [95, 101–111], highlighting its formative role in shaping early perceptions of gender roles, while novels [14, 18, 45, 66, 96–98, 112–118], short stories [12, 114, 119, 120], and fairy tales [121–123] have also received attention, revealing persistent disparities in character representation and gendered narrative patterns.

In contrast, drama and poetry are underrepresented in gender bias research, likely due to their structural and linguistic complexities, which pose challenges for both manual

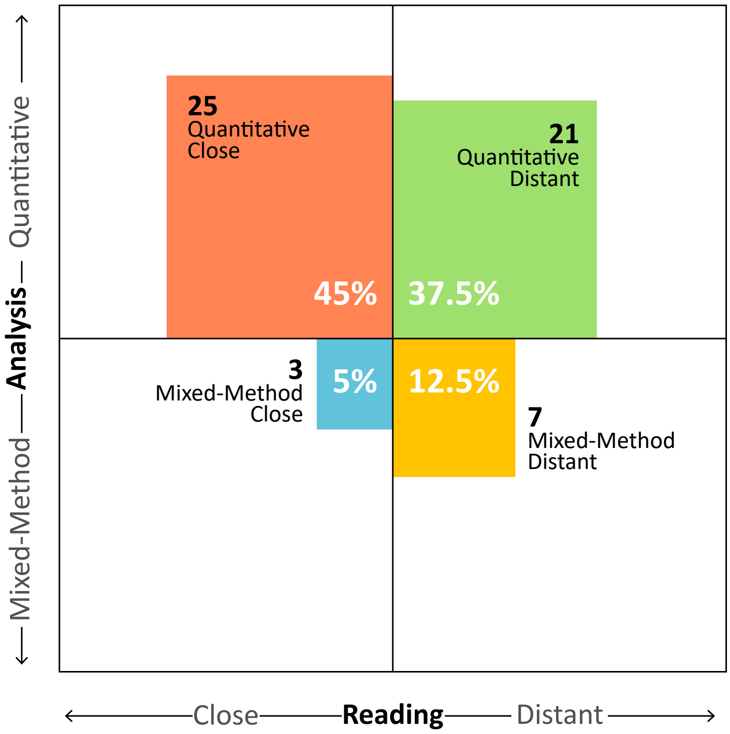


Figure 2.2: Distribution of the mapped studies based on reading strategy (close vs. distant) and type of analysis (quantitative vs. mixed-method).

and computational analysis [124–127]. Overall, this genre-specific distribution illustrates both the progress in understanding gender bias in widely studied forms and the need for broader, genre-diverse investigations. Expanding research beyond fiction could uncover nuanced patterns of representation, leading to a more comprehensive and intersectional understanding of gender in literature.

Gender Definition. Most studies adopt a binary framework for gender, classifying characters strictly as male or female (91%) [23, 97, 113]. While this approach facilitates data collection and aligns with traditional literature, which often reinforces binary gender norms, it oversimplifies the complexities of gender identity and expression, overlooking non-binary, genderqueer, and transgender characters. Only a small fraction of studies (9%) employ more inclusive definitions, introducing categories such as "other", neutral, or plural [45, 104, 107], allowing for broader analyses of gender representation. The predominance of binary definitions highlights a limitation in the field and underscores the need for future research to adopt more diverse and intersectional frameworks that better reflect contemporary understandings of gender.

Methodological Strategies. The mapped studies employ diverse methodological strategies, which we classified along two axes, as illustrated in Figure 2.2: reading approach (close vs. distant) and type of analysis (quantitative vs. mixed-method). *Quantitative-Close Reading* (manual coding and statistical analysis of selected texts) represents the largest group (45%) [11, 95, 97, 98, 101, 102, 104–107, 112–115, 119, 121, 122, 128–135].

This approach provides fine-grained insights into textual features and gendered patterns, though it is limited in scale due to labor-intensive coding. *Quantitative-Distant Reading* relies on NLP, text mining, and statistical or machine learning models across large corpora, accounts for 38% of studies [12, 14, 15, 18, 19, 23, 32, 45, 46, 66, 96, 110, 111, 116–118, 120, 124–127], uncovering systemic trends in gender representation that are difficult to detect through manual analysis. Mixed-method approaches combine computational techniques with qualitative interpretation, allowing researchers to contextualize quantitative findings. *Mixed-Method-Close Reading* (5%) offers detailed insights on small corpora [103, 136, 137], while *Mixed-Method-Distant Reading* (13%) integrates large-scale analysis with interpretive depth [10, 20, 24, 108, 109, 123, 138]. Together, these methodological strategies reflect a spectrum from in-depth textual interpretation to broad computational analysis, highlighting complementary strengths and limitations in investigating gender bias in literature.

Overview. In summary, the descriptive mapping highlights the evolution of gender bias studies in literature, emphasizing temporal growth, English-language predominance, concentration on fiction, widespread adoption of binary gender definitions, and methodological diversity. This overview provides a solid foundation for future research integrating larger corpora, underexplored literary genres, and more inclusive approaches to gender.

2.3.2 Analytical Mapping of Measurement Methods

Assessing gender bias in literary texts involves a combination of literary interpretation, linguistic analysis, and computational modeling. Despite the methodological variability across prior studies, it is possible to identify a coherent set of analytical stages that underpin most approaches. Based on the systematic mapping conducted in this chapter, we organize these stages into a unified analytical structure comprising four core components (Figure 2.3): (i) *character identification*, (ii) *gender classification*, (iii) *dependency analysis*, and (iv) *gender bias assessment*. This organization provides an overview of the methodological landscape, allowing heterogeneous studies to be compared within a common structure.

By mapping these stages, this dissertation clarifies the scope of existing approaches and establishes a conceptual foundation for the framework proposed here, which follows this general structure while introducing adaptations specific to Portuguese-language literary data. The following sections discuss each of the four components in detail, outlining their goals, typical techniques, methodological challenges, and their implications for gender bias assessment in literary texts.

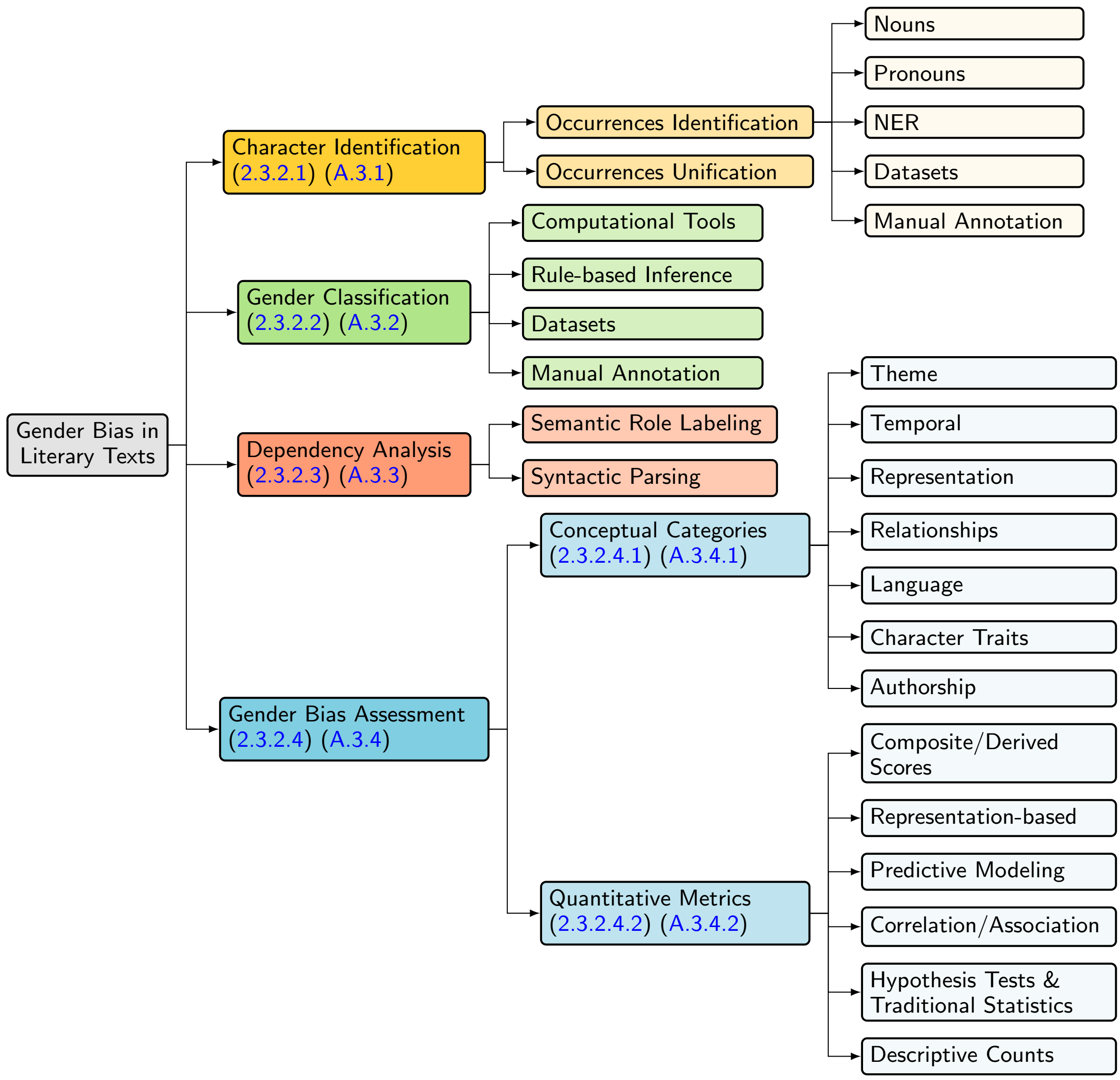


Figure 2.3: Conceptual map of gender bias in Literary Texts – sections that explain each category are presented between parentheses, including those in the appendix.

2.3.2.1 Character Identification

Character identification is the foundational step in measuring gender bias in literary texts, as it defines the entities whose representation will be analyzed and sets the scope of subsequent analyses. The definition of what constitutes a "character" is not merely technical but methodological [48]: a broad definition that includes all textual references to entities—real, fictional, or personified—captures peripheral voices and marginal presences, while a narrower definition focusing on central plot agents highlights dominant narrative patterns [50–52].

The conceptual breadth chosen has practical consequences. A comprehensive approach may reveal systemic gender imbalances among minor or background characters, whereas a restrictive focus emphasizes the gendered structure of the core narrative. Regardless of scope, character identification involves two subtasks: (i) *occurrences iden-*

tification, which detects each mention of a character in the text, and (ii) *occurrences unification*, which links different mentions that refer to the same underlying character [51]. Errors at this stage propagate through gender classification, dependency analysis, and bias measurement, making accuracy crucial.

Occurrences Identification. This subtask consists of detecting every instance in which a character appears in the text. Approaches vary from manual annotation to the use of pre-existing datasets and automated methods (Table A.4). *Manual annotation* is common in close reading studies, where annotators meticulously mark character mentions. It ensures high accuracy and allows nuanced interpretations, but is labor-intensive and thus less scalable. Multiple annotators are often used to reduce subjective bias and improve reliability [11, 95–97, 101–107, 113, 119, 121, 127–133, 135, 137]. Single-annotator approaches exist but may introduce subjectivity [98, 112, 115, 122, 134, 136].

Datasets provide annotated character mentions, facilitating larger-scale analyses. Examples include the Black Drama Database [124], Project Gutenberg texts [125], Goodreads lists [138], the Chapitres corpus [45], and OBras [10]. These resources accelerate identification but may be limited by genre, language, or annotation granularity. *Automated methods* offer scalability and efficiency. Named Entity Recognition models identify character names based on linguistic patterns and context [14, 15, 23, 24, 110, 116, 117, 120], often using tools like BookNLP [139]. Pronoun-based detection leverages gendered pronouns [18, 46, 108, 109, 118], while noun-based methods detect descriptive or gendered nouns such as "woman" or "child" [19, 66, 111]. These automated approaches may require post-processing to resolve ambiguities or unusual textual references.

Occurrences Unification. Once mentions are identified, they must be linked to the same underlying character, a task often referred to as *coreference resolution* [12]. Characters may be referenced by names, titles, pronouns, or descriptive phrases (e.g., "John Smith", "he", "the brave knight"). Accurate unification ensures coherence for subsequent analyses of character dynamics and narrative roles [51].

Approaches range from manual linking to automated models that exploit syntactic and semantic features [12, 117, 120, 140–142]. While optional in simpler analyses, unification is essential for in-depth studies, particularly those exploring character interactions, development, or structural narrative patterns. Effective occurrence unification increases analytical robustness by ensuring that all references to a character are consistently captured throughout the text.

2.3.2.2 Gender Classification

After characters have been identified and their mentions unified, the next critical step is gender classification. Assigning gender labels to characters is essential for analyzing gender dynamics in literary texts, as it enables the study of representation, interactions, and narrative bias. The choice of classification method can affect both accuracy and interpretability. Table A.5 summarizes the main approaches identified in the literature, grouped into manual annotation, pre-annotated datasets, rule-based inference, and computational tools, which are briefed next.

Manual Annotation. Manual annotation remains a widely used method, especially in studies that prioritize interpretive depth. Annotators verify contextual cues—such as pronouns, proper names, titles, and narrative descriptions—to assign gender labels to characters [98, 112, 114, 115, 122, 134, 136]. This approach can capture complex or implicit gender representations, including non-binary or fluid identities, which automated methods may overlook. However, manual annotation is labor-intensive and time-consuming, limiting scalability. Multiple annotators are typically employed to mitigate individual bias, with consensus procedures used to resolve discrepancies [11, 45, 95–97, 101–107, 113, 119, 121, 127–133, 135, 137].

Pre-Annotated Datasets. Pre-existing datasets containing gender labels provide a reference for automated or semi-automated classification. For example, the Black Drama Database [124] includes detailed gender annotations for characters, facilitating large-scale analyses without requiring full manual labeling. While datasets streamline classification, their applicability may be limited by factors such as language, genre, or corpus coverage.

Rule-Based Inference. Rule-based approaches classify gender using predefined linguistic cues, such as pronouns, nouns, titles, or proper names [12, 14, 18–20, 46, 66, 108–111, 118, 120, 123, 126, 138]. These methods are widely used because they are transparent, relatively easy to implement, and can scale efficiently to large corpora. However, they rely on the assumption that linguistic markers consistently convey gender, which may not hold in all contexts or for non-binary identities.

- **Name-based inference.** This approach uses databases mapping first names to gender distributions. For example, U.S. Social Security Administration data has been employed to associate given names with likely gender [110]. Name-based inference is effective for characters whose names carry strong cultural or historical gender associations, and it scales easily to large datasets. Nevertheless, it can fail with ambiguous names, nicknames, or culturally diverse names not covered in the database. Additionally, it cannot capture characters with non-binary or fluid gender identities. Name-based inference is particularly useful when combined with other cues to increase robustness.

- **Honorifics and titles.** Titles such as “Mr.”, “Mrs.”, “Miss”, or their equivalents in other languages provide explicit gender information [14]. This method is straightforward and often highly accurate in formal or traditional texts, where honorifics are consistently applied. Its limitations emerge in narratives with informal language, historical variation, or texts that omit titles. Furthermore, like name-based inference, it presumes a binary gender system and cannot represent non-binary identities.

- **Pronoun-based inference.** This method relies on gendered pronouns (e.g., “he/him”, “she/her”) to classify characters, exploiting their syntactic and semantic roles in the text [12]. Pronoun-based inference is effective in languages where pronouns are explicitly gendered, and it captures references that may not appear in proper names or titles. However, it can be challenged by ambiguous pronouns, dropped pronouns (as in some Romance languages), plural pronouns, or texts with neutral or non-binary pronouns. It is also sensitive to coreference resolution errors, as pronouns must be correctly linked to the intended character.

- **Gendered nouns and lexical cues.** Words indicating social roles, relationships, or attributes (e.g., “woman”, “girl”, “man”, “boy”, “mother”, “father”) provide additional clues for gender classification [60, 66, 138]. These cues are particularly useful when characters are referred to descriptively rather than by name. While they enrich the classification, they may introduce bias by reinforcing stereotypical associations (e.g., associating certain traits with a specific gender). Moreover, their applicability is limited in texts that deliberately avoid gendered terms or employ abstract language.

In practice, effective rule-based inference often combines multiple cues (names, pronouns, titles, and lexical markers) to improve coverage and accuracy. Researchers must carefully consider the textual context, the historical and cultural background of the work, and potential biases inherent in the linguistic markers. While rule-based methods offer transparency and interpretability, they should ideally be complemented with manual verification or computational approaches to ensure robust gender classification, especially when analyzing subtle or complex gender dynamics.

Computational Tools. Computational tools leverage natural language processing to infer gender at scale. **BookNLP** [10, 15, 23, 116] identifies characters and assigns gender using pronouns and contextual cues in English texts. For Portuguese, **PALAVRAS** [32, 143, 144] applies syntactic parsing and linguistic rules to detect gendered references. Other resources include the *gender* R package [145], NLTK-based methods [125], and Python libraries such as *Gender_Detector* [24]. Computational tools combine scalability with the ability to process complex linguistic patterns, making them suitable for large corpora and automated pipelines.

2.3.2.3 Dependency Analysis

Dependency analysis explores grammatical relationships within sentences, mapping how words such as verbs, subjects, objects, and modifiers are syntactically connected. This approach can reveal patterns in narrative agency, for example, showing whether characters act as agents or are positioned as recipients of actions [18]. Its adoption across literary gender studies, however, is uneven. It is often absent in research focused solely on metadata (e.g., authorship) or in studies relying on manual annotation, where syntactic patterns are analyzed qualitatively [95, 101, 102, 104, 107, 112, 116, 124, 128, 130–134, 136].

When applied, dependency analysis can allow scalable, detailed investigation of structural asymmetries in the distribution of narrative agency and descriptive attributes across genders. Key techniques include syntactic parsing, Semantic Role Labeling (SRL), and collocation analysis. Syntactic parsers, such as Stanford Dependency Parser, SpaCy, PALAVRAS, and BookNLP, construct dependency trees that identify who performs which actions [10, 15, 23]. SRL assigns semantic roles, including agent, patient, and experiencer, highlighting active versus passive portrayals [14]. Collocation analysis, in turn, verifies co-occurrences of gendered nouns or pronouns with verbs and adjectives, revealing stereotypical lexical associations [66, 138].

Although the accuracy of these methods depends on parser performance and may be limited in texts with complex syntax or non-binary gender representations, dependency analysis provides a robust, data-driven foundation for assessing narrative gender bias.

2.3.2.4 Gender Bias Assessment

Gender bias assessment evaluates how gendered dynamics are represented in literary texts. It comprises two complementary dimensions: (i) *conceptual categories*, which capture *what* aspects of gender bias are investigated (e.g., authorship, character traits, representation), and (ii) *quantitative metrics*, which capture *how* such biases are measured. Together, these dimensions reveal both the conceptual framing and the methodological strategies used to assess and interpret bias in literary corpora.

2.3.2.4.1 Conceptual Categories Studies conceptualize gender bias in multiple ways, reflecting the interdisciplinary nature of the field and the varied theoretical lenses applied to literary texts. The categories identified in the SMS are inductively derived from the mapped corpus and represent recurring analytical perspectives on how gender is constructed, represented, and evaluated in literature. Each category captures a particular facet of gendered representation, ranging from authorship and character portrayal to linguistic patterns, relational structures, and thematic framing. Table A.6 (see Appendix A) summarizes these categories and their representative studies.

Authorship. This category investigates how the gender of the author influences character portrayal [14, 15, 23, 24, 32, 46, 66, 102, 103, 106, 116, 124, 134, 135, 138]. Results often show systematic differences: female authors tend to depict genders more equally, while male authors disproportionately foreground male characters. For instance, Qian

[66] found that male authors referenced male characters three times more frequently than female characters, while female authors distributed references more evenly.

Character Traits. Analyses of character traits focus on attributes and qualities associated with male and female characters, such as personality traits and physical appearance. Regarding personality traits, studies often investigate whether characteristics such as assertiveness, empathy, and resilience are assigned differently based on gender, reflecting or challenging societal expectations [102, 105, 113, 128]. For example, Diekman and Murnen [113] used an adapted version of the Bem Sex Role Inventory [146] to assess whether characters expressed traditionally gender-stereotyped traits, finding that sexist texts portrayed more conventional gender roles compared to nonsexist texts. These analyses frequently reveal that male characters are often attributed assertive traits, while female characters are assigned more communal traits like empathy and gentleness.

Physical appearance analysis explores how gendered descriptions can challenge stereotypes, often revealing gender bias that aligns with traditional views of masculinity and femininity [12, 14, 32, 108, 109, 111, 117–119, 121, 122]. Female characters, for instance, are frequently depicted with an emphasis on beauty or physical fragility, highlighting delicate or nurturing attributes, whereas male characters are typically portrayed as strong, imposing, or physically powerful [121, 122]. This pattern reinforces a visual bias where physical descriptions of female characters emphasize attractiveness or fragility, whereas male characters' descriptions emphasize physical strength and dominance. Moreover, some studies highlight a notable scarcity of emotional descriptors in female character portrayals; instead, descriptions focus predominantly on physical attributes, reinforcing an appearance-based portrayal over one that explores inner qualities [32].

Language. Language-based analyses examine broader linguistic patterns across the narrative, focusing on how words, expressions, and semantic associations reflect gender bias. Adjectives, in particular, offer insight into how specific attributes are assigned to male and female characters, often reinforcing traditional gender stereotypes [10, 11, 19, 20, 23, 32, 45, 46, 98, 103, 111, 122, 125]. For instance, Fast et al. [46] found that adjectives associated with male characters commonly reinforced stereotypes like "strong", "arrogant", and "dominant", while adjectives associated with female characters included terms like "submissive", "dependent", and "hysterical". Similarly, Schulz and Štěpán Bahník [19] observed that men were more likely to be described as "wise", "honorable", and "able", and less likely to be described as "fashionable" and "warm". In contrast, the authors uncovered that women were more likely to be described as "charming", "vulgar", "independent", and "foolish".

In addition to adjectives, verbs and predicates are also crucial elements in analyzing gender bias within language [10, 11, 15, 18, 20, 23, 45, 46, 111, 114, 118, 125, 137]. Studies have shown that male characters are more often depicted with action-oriented verbs, suggesting agency and control, while female characters are more frequently associated with passive verbs or verbs related to emotions and social interactions, reinforcing traditional gender stereotypes [45]. For instance, Freitas and Martins [10] identified specific verbs associated with female characters in their study, such as “amamentar” (to breastfeed) and “alimentar” (to nourish), which emphasize maternal roles. Other verbs like “menear” (to sway), “verter” (to shed), and “definhar” (to wither) convey a sense of vulnerability or delicateness. Additionally, actions like “ostentar” (to flaunt) and “enfeitar” (to adorn) align with vanity or superficiality, echoing societal expectations of femininity. In contrast, male characters are often portrayed with verbs suggesting dominance and resilience, reinforcing masculinity archetypes that prioritize strength and assertiveness.

Some studies leverage lexicons to explore how specific vocabulary choices reflect and perpetuate gender biases in literary texts [10–12, 32, 46, 66, 120, 123, 138]. Lexicon-based analyses typically involve compiling lists of gendered or stereotype-laden terms, which can then be used to identify and quantify patterns in how male and female characters are described. For example, Fast et al. [46] constructed a stereotype lexicon based on 2,000 common verbs and adjectives, associating specific descriptors with either male or female characters. Categories linked to male characters included terms like “violent”, “dominant”, “strong”, “arrogant”, “sexual”, “angry”, and “active”, while female characters were predominantly associated with categories such as “domestic”, “hysterical”, “childish”, “afraid”, “dependent”, “emotional”, “beautiful”, and “submissive”.

Finally, some studies analyze additional lexical features to uncover gender biases, including word embeddings [123, 125–127], co-occurrence patterns [96, 108, 109, 111, 118, 122, 125], valence [12, 19, 20, 125], and utterance-final expressions [115]. Embedding-based analyses, for instance, use computational methods to analyze the semantic associations between words and reveal implicit gender bias. By mapping words to vector spaces, researchers can quantify how closely certain words (such as those describing characters or actions) are associated with gendered terms. Zhang and Wu [126] used word embeddings to analyze gender representation in popular fiction, assigning a *gender bias score* to quantify and highlight bias in how male and female characters are portrayed.

Co-occurrence analyses focus on the frequency and patterns of word associations within a corpus, revealing underlying stereotypes in character interactions. For example, Poynter [111] investigated collocations of terms like GIRL, BOY, WOMAN, and MAN, noting both distinctions and similarities between different genres. While some gender stereotyping emerges in reporting verbs and physical descriptions, the study finds no clear polarization of gender roles in the collocations. In a different approach, Čermáková and Mahlberg [108, 109] explored Gendered Body Part Clusters (GBPCs), which consist of clusters with at least one body part noun and at least one third-person singular possessive determiner or pronoun.

Valence analysis investigates the emotional tone or positivity associated with descriptors, offering another layer of gender-based analysis. Research has shown that, historically, men have been described more positively than women, with girls receiving more favorable portrayals than boys at the beginning of the twentieth century [19]. Furthermore, Hoyle et al. [20] found that positive adjectives used to describe women often focus on their physical attributes, while those used for men tend to emphasize behavioral traits. This distinction not only reflects existing gender stereotypes but also underscores the different standards applied to male and female characters.

Relationships. Analyzing relationships within narratives reveals significant patterns in how male and female characters interact, including antagonistic, nurturing, and sexual or social dynamics. Studies often explore co-occurrence to determine which characters appear together, shedding light on the social and narrative connections across genders [15, 116]. For instance, studies have found that antagonistic relationships—where characters are in opposition—often reflect stereotypical power dynamics, with male characters more frequently portrayed in competitive or confrontational roles compared to their female counterparts [116]. In contrast, female characters are often positioned in supportive or nurturing roles, which may perpetuate traditional gender norms that associate femininity with care and cooperation [98, 105, 106, 129].

Additionally, some studies delve into more nuanced social aspects, such as marital status and sexuality. For instance, Das and Das [119] observed that heroines in popular magazine short stories in India predominantly navigate marital and interpersonal issues, reflecting societal expectations regarding women's roles in relationships. Similarly, Dalcastagnè [97] found that contemporary Brazilian novels emphasize romantic and family relationships, a trend that is particularly pronounced for female characters, suggesting that their narratives are often centered around domesticity and emotional connections. In contrast, Zolin [98] highlights a shift in the portrayal of women in contemporary Brazilian novels written by female authors. These narratives reveal that the female characters explore their sexuality in ways that transcend traditional gendered norms.

Representation. Representation analyses explore the balance of gender roles assigned to characters, considering various demographic and narrative factors. This category includes the overall ratio of male to female characters [11, 14, 15, 45, 66, 98, 101, 105, 110, 113, 119, 120, 124, 128–133, 135, 136], their ages [19, 97, 98, 103, 107, 110, 111, 119, 121, 122, 129, 135], education levels [98, 119], settings and locations [98, 102, 106, 119, 132], professional and leisure occupations [66, 97, 98, 106, 113, 119, 129, 132], portrayed roles in the narrative [12, 15, 95, 97, 98, 101, 103, 105, 106, 112, 113, 119, 128, 131, 133, 136], race [97, 103, 110, 124, 129], religion [129], sexual orientation [97], and social status [97, 105, 113, 119].

Overall, studies on gender representation consistently find a predominance of male characters, who not only outnumber female characters but are also more likely to occupy central roles, while female characters are frequently relegated to secondary or supportive positions [23, 24, 32, 95, 97, 102–104, 106, 107, 116]. Such gender imbalance is widespread across genres, from children's books to popular adult fiction, suggesting that literature continues to mirror and reinforce societal biases. While female representation has improved over time, the nature of this inclusion remains uneven. Adukia et al. [110], for example, found that females are often more visible in images than in textual mentions, suggesting that visual representation may offer a more symbolic inclusion, whereas the narrative content still prioritizes male-centric storylines.

Other demographic aspects, such as age and race, further illustrate how language reflects and reinforces gender and social bias in literary texts. For instance, in twentieth-century English-language literature, Turner-Bowker [103] found that adjectives used to describe children were generally less associated with traditionally masculine qualities than adjectives used for adults. Regarding race, Argamon et al. [124] revealed striking differences in language used by male and female authors to describe characters of different racial backgrounds, revealing how intersectional factors of race and gender can influence character portrayals. These analyses underscore that representation in literature is not just about gender but may involve a nuanced web of social characteristics.

Analyses of occupation and locations also reveal how gender stereotypes are embedded in the environments and professional roles assigned to characters. For instance, studies indicate that female main characters are more frequently depicted in indoor settings compared to outdoor ones, which can reinforce traditional views of women as being associated with domestic or private spheres rather than public or adventurous spaces [102, 106, 119, 132]. In terms of occupations, studies have also found that female characters, especially in children's and young adult literature, are more likely to have no paid occupation at all, reinforcing notions of dependency or domesticity [106].

Temporal. Temporal analyses explore how gender bias has shifted over time in literary texts. By exploring works from different periods, researchers can trace changes in the portrayal of gender roles and the prevalence of gender bias, shedding light on the cultural and historical influences that shape these representations [15, 23, 24, 101, 106, 107, 109, 110, 117, 128, 131–133, 135]. For instance, McCabe et al. [107] found that the period between the 1930s and 1960s exhibited the most significant disparity in representation between male and female characters in children's books, coinciding with the aftermath of the first-wave women's movement. This suggests that societal changes and movements can significantly impact literary portrayals of gender, reflecting broader cultural dynamics.

Theme. Finally, analyses of themes address the broader topics [120, 130, 138] and literary genres [32, 135, 138] associated with male and female characters. Specific themes, such as adventure, war, or leadership, may appear more frequently in stories centered on male characters, while themes like romance, family, or domestic life may be more commonly associated with female characters. This thematic analysis illustrates how gender expectations influence not only individual character roles but also the larger narrative structures and themes of literary works, providing a comprehensive view of gender representation within various literary genres [138].

2.3.2.4.2 Quantitative Metrics While conceptual categories capture the thematic dimensions of gender bias, quantitative metrics operationalize these dimensions into measurable variables. Studies employ diverse computational techniques to quantify bias, ranging from simple frequency analyses to complex network and embedding-based models. These metrics can be broadly organized into six main categories: descriptive counts, hypothesis tests and traditional statistics, correlation or association, predictive modeling, representation-based methods, and composite or derived scores. Table A.7 (see Appendix A) summarizes these categories and their representative studies.

Descriptive Counts. Descriptive counts provide a baseline quantification of gendered patterns by reporting raw frequencies, proportions, or percentages of characters, traits, or lexical features [11, 15, 18–20, 24, 32, 45, 46, 66, 95–98, 101–119, 121–123, 128–135, 137]. Although simple and interpretable, these counts do not establish statistical significance or causality. Nonetheless, they are essential for identifying general tendencies, such as the proportion of male versus female characters or the frequency of gendered adjectives. For instance, Filipović [136] used quantitative counts to examine male-to-female ratios in both text and illustrations of children's books, highlighting discrepancies between visual and textual representation. Similarly, Freitas and Martins [10] counted occurrences of predicates and action verbs associated with male and female characters.

Hypothesis Tests & Traditional Statistics. Inferential statistics are applied to compare distributions or test differences between groups [12, 14, 19, 20, 46, 66, 95, 96, 101–107, 111, 113, 116, 119–121, 128, 131–135]. Methods include chi-square tests, t-tests and ANOVAs for comparing group means, and non-parametric alternatives such as Wilcoxon tests. These approaches provide evidence of whether observed gender differences in representation or language use are statistically significant. For example, Kejriwal and Nagaraj [24] used an extensive set of statistical tests to demonstrate a significant disparity between the prevalence of female and male characters in pre-modern literature. Similarly, Pownall and Heflick [11] used ANOVAs to test for differences in direct speech and peripheral character gender across books, and applied Chi-squared tests to examine gender differences in ratings of characters' agency, passivity, and stereotypes. Such inferential methods move beyond simple counts to statistically validate gendered patterns.

Correlation/Association. Correlation and association metrics quantify the strength and direction of relationships between gender and textual features [11, 12, 18–20, 45, 116–119, 121, 123, 125, 134]. Common measures include Pearson or Spearman correlations, log-likelihood ratios, and assortativity coefficients. These metrics reveal whether certain words, traits, or social connections systematically co-occur with male or female characters. For example, Fast et al. [46] analyzed stereotype categories and found that men were more often described as strong and active, while women were characterized as weak and submissive. Similarly, Weingart and Jorgensen [122] investigated correlations between abstract nouns and gender; for instance, "beauty" appeared more frequently in reference to women than "handsome" did for men. Correlation and association metrics allow precise quantification of these relationships, exposing subtle patterns of bias in literary texts.

Predictive Modeling. Predictive modeling is employed both to measure gender bias explicitly and to infer character gender from lexical features. Some studies develop supervised classifiers for bias detection [12, 46, 124, 127], while others use the models to identify which features most strongly signal gender [18, 23, 45, 117]. Logistic regression and other supervised classifiers are common, and metrics such as accuracy, precision, recall, and F1-score indicate how strongly gendered patterns are encoded. These models provide insights into both predictive performance and the discriminative power of textual features in shaping gendered portrayals.

Representation-based. Vector-based methods leverage distributed semantic representations to assess implicit bias [12, 120, 123, 125]. Metrics such as cosine similarity or the WEAT [67] are used to capture latent associations between gendered terms and semantic domains. For instance, Zhang and Wu [126] applied word embeddings to analyze and quantify gender bias in popular fiction, computing the Euclidean distance between the embedding of a given text and the average embeddings of gendered word sets (male vs. female). Similarly, Luo et al. [14] employed WEAT to compare the similarity of "male" and "female" word sets with semantic categories such as "appearance". Their analysis revealed that, in corpora where women were depicted primarily as objects of visual pleasure, female terms tended to cluster closer to appearance-related words than male terms in the embedding space.

Composite/Derived Scores. Some studies proposed composite or derived scores that integrate multiple textual features into measures of gender bias [14, 15, 19, 45, 66, 138]. These scores are usually designed to capture higher-level patterns that cannot be observed from single features alone, operationalizing constructs such as agency, stereotypicality, genderedness, or associations between words and gender. We describe some representative examples as follows, highlighting both their construction and interpretability.

Composite Score (*CS*). Schulz and Štěpán Bahník [19] proposed the *composite score* to quantify bias in adjective usage across specific dimensions (e.g., positivity, masculinity) for a given noun (e.g., "man", "woman", "boy", "girl") and year. The composite score, CS, is computed as a weighted average of adjective ratings, multiplying the rating of each adjective on a dimension by its proportion in adjective-noun bigrams, summing these products, and dividing by the total frequency of the most frequent adjectives:

$$CS_{y,d,n} = \frac{\sum_{a \in A_n} f_{a,y,d,n} \times AR_{a,d}}{\sum_{a \in A_n} f_{a,y,d,n}},$$

where y is the year, d the dimension, n the noun, $f_{a,y,d,n}$ the frequency of adjective a, A_n the set of most frequent adjectives, and $AR_{a,d}$ the average rating of adjective a on dimension d.

Stereotype Scores. As previously discussed in Section 2.2.2.1, Qian [66] introduced three stereotype scores: gender, occupation, and emotion. The gender stereotype of a word w is defined as:

$$b(w) = \left| log \frac{c(w,m)}{c(w,f)} \right|,$$

where $c(w,g)$ counts co-occurrences of a gender-neutral word w with gendered words $g \in m, f$ (e.g., $f =$ she, girl, woman, $m =$ he, actor, father). A word is neutral if $b(w) = 0$. The overall gender stereotype score for a text, T_b, is the mean of all relevant words:

$$T_b = \frac{1}{N} \sum_{w \in N} b(w).$$

where N is the set of words with sufficient co-occurrences. Occupation (O_b) and emotion (E_b) stereotypes are computed similarly using predefined word lists O and E:

$$O_b = \frac{1}{|O|} \sum_{w \in O} b(w), \quad E_b = \frac{1}{|E|} \sum_{w \in E} b(w).$$

Genderedness Score. Gala et al. [138] defined a *genderedness score* g_i for narrative tropes, based on the proportion of female and male gendered terms in a trope relative to the corpus:

$$d_i = \frac{f(X_i)}{f(X_i) + m(X_i)} \Big/ \frac{f(\mathrm{T})}{f(\mathrm{T}) + m(\mathrm{T})},$$

where $f(X_i)$ and $m(X_i)$ are counts of female and male tokens in trope i, and $f(\mathrm{T})$ and $m(\mathrm{T})$ are totals for the corpus. The final score g_i is z-score normalized, with values outside [-1,1] considered highly gendered.

Association Score. Barré and Dupont [45] introduced an *association score* to capture the link between words (verbs/adjectives) and gender:

$$(masc - fem)/(masc + fem),$$

where "masc" and "fem" are the proportions of masculine and feminine characters associated with a word. Positive scores indicate a stronger male association, negative scores stronger female association.

Appearance Bias Score. Luo et al. [14] proposed an *appearance bias score*, which quantifies gender bias by comparing the closeness of female versus male words to appearance-related words in a text's embedding space. Using WEAT:

$$WEAT(\mathbb{W}') - WEAT(\mathbb{W}),$$

where $\mathbb{W}$ is a pre-trained embedding space and $\mathbb{W}'$ is fine-tuned on the text. A positive appearance bias score indicates that female words are more closely related to appearance words than male words, suggesting the text portrays women as objects of visual pleasure. Conversely, a negative score would imply male objectification, while a score of zero indicates the absence of this specific bias.

Agency Score. Several studies have explored the concept of agency to capture characters' capacity to act intentionally and autonomously within narratives. Barré and Dupont [45] introduced the *agency score*, which quantifies a character's propensity to function as the grammatical subject versus object in sentences. It is calculated as:

$$(nsubj - obj)/(nsubj + obj),$$

where "nsubj" is the number of times a character appears as the subject of a sentence, and "obj" is the number of times they appear as the object. The score ranges from -1 to 1; higher positive values indicate that a character more frequently occupies the subject role, thus taking a more active role in advancing the story, whereas lower or negative values indicate a more passive or reactive role.

Similarly, Luo et al. [14] measured gender bias by comparing how often male versus female entities act as grammatical agents, where positive scores indicate greater male agency and negative scores greater female agency. Extending this idea, Stuhler [15] introduced male agency surplus for dyadic interactions, defined as the percentage-point difference in actions initiated by male and female characters (ranging from –100 to 100), enabling a fine-grained analysis of active versus passive roles in cross-gender interactions.

Despite their diversity, these composite metrics share a common goal: translating abstract narrative constructs into interpretable quantitative signals, allowing gender bias to be analyzed systematically without fully abstracting away literary meaning.

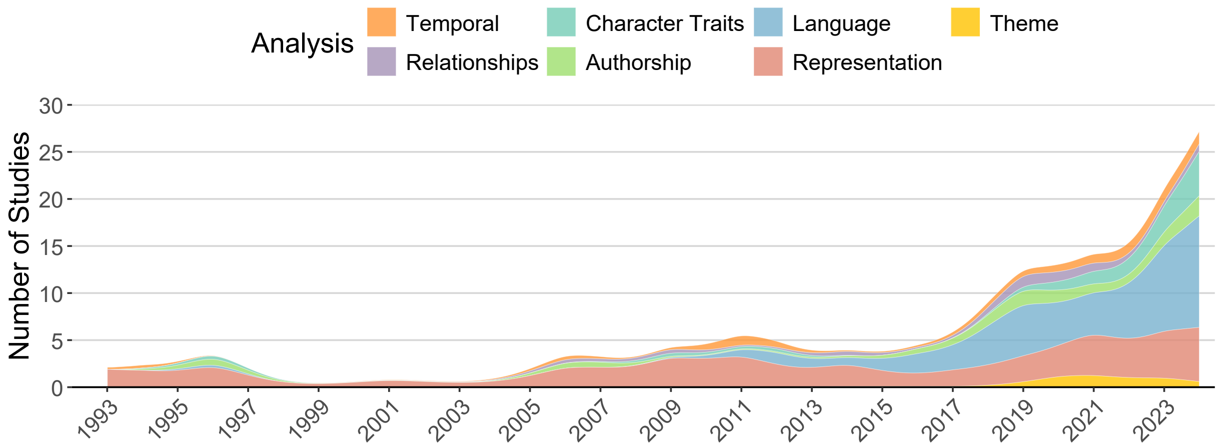


Figure 2.4: Evolving distribution of gender bias analysis conceptual categories.

2.4 Gaps and Opportunities

The study of gender bias in literature presents rich opportunities for exploration, revealing patterns of representation that inform discussions on gender equality and cultural perceptions. Despite substantial progress, several persistent gaps remain, spanning both Digital Humanities and Computer Science domains. Addressing these gaps opens several avenues for future research, as well as motivates the development of the framework proposed in this dissertation. In this section, we first discuss conceptual and methodological limitations within the Digital Humanities literature (Section 2.4.1), then turn to the Computer Science domain, highlighting computational challenges in modeling, measurement, and evaluation (Section 2.4.2).

2.4.1 Limitations in Digital Humanities

As discussed in Section 2.3.2.4.1, research on gender bias in literature has evolved through multiple conceptual categories, including authorship, character traits, language, relationships, representation, temporal shifts, and thematic focus. Figure 2.4 illustrates the evolving distribution of these analytical dimensions, highlighting the diversification of research interests over time.

Early studies largely relied on broad quantitative indicators, such as male-to-female character ratios or differences in portrayal between male and female authors. Over time, research has shifted toward more nuanced analyses that explore linguistic patterns, character traits, themes, and inter-character relationships. This shift reflects a growing recognition that gender bias in literature operates across multiple levels, from straightforward quantitative indicators to deeper linguistic and narrative structures.

Indeed, analyzing multiple dimensions of gender bias simultaneously is critical to understanding how narrative elements intersect. For instance, combining analyses of language and character traits may reveal how descriptive choices or verb usage reinforce traditional gender roles, while the joint study of relationships and themes can expose relational hierarchies that shape character agency and narrative direction. Future research should therefore move toward integrated frameworks capable of capturing both explicit and implicit manifestations of gender bias in literary texts.

Another limitation concerns the scope and inclusiveness of existing corpora. Most prior studies rely on restricted or homogeneous literary datasets, often focused on canonical or Western authors [24, 32, 118, 127, 138]. Expanding corpora to include diverse genres, historical periods, and underrepresented authors is crucial for achieving more generalizable and culturally representative insights. Larger and more heterogeneous datasets also enable the use of computational models capable of identifying linguistic and narrative patterns otherwise hidden in smaller corpora.

Corpus expansion also opens avenues for longitudinal research, enabling the analysis of temporal dynamics in gender representation [24, 45]. Such analyses can trace how literary depictions of femininity and masculinity evolve alongside societal transformations, feminist movements, and changing gender norms, revealing both enduring stereotypes and emerging shifts toward equality [92, 107, 135].

A further conceptual limitation lies in the persistent reliance on binary gender frameworks. Traditional categorizations that assume a strict male–female dichotomy fail to account for the diversity of gender identities represented in literature [12, 14, 24, 116, 117]. As social understandings of gender become increasingly fluid, digital literary studies need to incorporate non-binary, transgender, and other diverse identities into their analyses. Doing so would enable a more inclusive and accurate understanding of how gender operates as a dynamic and socially constructed category within literary discourse.

Moreover, while gender remains the predominant focus of most studies, intersectional perspectives are still rare. Integrating additional social dimensions, such as race, age, class, and sexuality, offers the potential to uncover more complex and layered forms of bias [15, 18]. For example, exploring how race and gender interact in character portrayal can reveal compounding stereotypes that marginalize particular groups, while considering class or age may illuminate disparities in power and narrative centrality.

Finally, some methodological and linguistic limitations still exist in the field. Most studies are conducted in English, with about 80% of the reviewed works focusing on English texts, while only a few analyze Portuguese or other languages, which limits cross-linguistic and cultural generalizability. Additionally, nearly half of the studies fall into the *Quantitative-Close Reading* category, relying on manual annotation and coding, which can be labor-intensive, susceptible to human bias, and hard to scale to larger datasets. Addressing these methodological limitations by expanding language coverage, adopting automated text analysis, and using scalable computational methods would enhance the rigor and inclusivity of gender bias research in literature.

2.4.2 Limitations in Computer Science

While the studies reviewed in Digital Humanities differ in scope and methodology from those in Computer Science, both domains reveal similar underlying limitations in how gender bias is conceptualized and operationalized. Computational approaches to measuring and analyzing gender bias in Natural Language Processing have advanced rapidly, yet persistent challenges remain. These challenges include the treatment of gender, linguistic diversity, and methodological rigor [44, 53], reflecting broader gaps observed across both research domains.

A key limitation is the prevalent binary treatment of gender, a challenge also noted in Digital Humanities (Section 2.4.1). Most NLP research simplifies gender as a dichotomy (male versus female), overlooking its fluidity and spectrum [44, 53, 58, 84, 147]. Such simplifications can perpetuate harms, including misgendering and the erasure of non-binary identities [46]. In many studies, gender labels are assigned without clear justification or transparency, raising ethical concerns and affecting model reliability. Moreover, NLP models frequently achieve lower performance for non-binary pronouns and references compared to binary ones, highlighting persistent gaps in inclusivity and fairness [44].

Another significant challenge, also mirrored in Digital Humanities research, is the focus on high-resource, monolingual settings, particularly English [148–150]. This narrow linguistic scope restricts understanding of how gender bias manifests across different linguistic and cultural contexts. Languages with grammatical gender agreement, such as Portuguese, exhibit distinct patterns of gendered expression that cannot be adequately captured by models trained exclusively on English [31]. Expanding research to include low-resource and morphologically rich languages is, therefore, essential for developing more inclusive and generalizable insights.

Methodological limitations also persist. Current approaches often rely on narrow or inconsistent definitions of gender bias, frequently focusing on a single type of bias, such as occupational stereotypes, while ignoring other dimensions of harm [55]. Evaluation metrics are similarly inconsistent, with many studies relying on a single test or score, which reduces robustness and comparability [44]. Moreover, while existing gender bias measures have led to an awareness of how models integrate and emphasize existing biases in the data, a definitive bias measure that works reliably does not yet exist [84]. Finally, the field lacks standardized baselines and benchmark datasets, making cross-study comparison and reproducibility challenging [53, 84].

2.4.3 Overall Considerations

As discussed throughout this chapter, both Digital Humanities and Computer Science approaches to gender bias in literature face overlapping challenges. While the former often relies on limited corpora, manual annotation, and binary conceptualizations of gender, the latter encounters comparable issues related to modeling assumptions, linguistic coverage, and evaluation consistency. Despite differences in epistemological orientation and methodological focus, both fields share a need for more inclusive, generalizable, and interdisciplinary research on gender bias in literature.

Bridging these domains requires combining the interpretative depth characteristic of the Humanities with the methodological precision of NLP. Yet, existing studies rarely achieve such integration. Digital Humanities research often prioritizes contextual interpretation, sometimes limiting scalability, whereas computational work tends to emphasize quantitative assessment without sufficient engagement with literary or social theory [55]. Addressing these gaps may enable more comprehensive and context-sensitive analyses of gender bias in literary texts.

The framework proposed in this dissertation is designed to respond to the limitations identified across both domains. It provides a flexible and modular structure capable of handling diverse literary corpora and supporting analyses that are more scalable and systematic than traditional manual approaches. By explicitly structuring gender bias analysis into modular analytical stages and grounding each stage in both linguistic theory and computational practice, the proposed framework advances the state of the art beyond ad-hoc or task-specific approaches commonly found in the literature.

Finally, the framework integrates resources adapted for Portuguese, an underrepresented language, addressing the predominance of English-language studies and the limited availability of NLP resources for literary texts in other languages. Rather than aiming to exhaustively capture all dimensions of gender and social bias, the proposed framework is intended as a methodological baseline that makes analytical assumptions explicit and supports extensibility as new theories, resources, and models emerge.

Chapter 3

Computational Framework for Gender Bias Measuring and Analysis

Gender bias in literary texts can manifest across multiple dimensions, including character representation, linguistic expression, relational dynamics, and broader thematic patterns. Although previous research in Digital Humanities and Computational Linguistics has explored these phenomena, existing approaches remain constrained by several persistent challenges: the reliance on small or homogeneous corpora, a predominance of manual and often non-replicable annotation procedures, the lack of standardized and comparable metrics, and a concentration of methodological development in English-language contexts.

These limitations are particularly evident in the context of Portuguese, a morphologically rich language whose grammatical gender system, pronominal dependencies, and flexible word order introduce analytical challenges not addressed by methods originally developed for English. To address these gaps, this chapter proposes a computational framework specifically designed for the large-scale and linguistically grounded analysis of gender bias in Portuguese-language literary corpora.

Rather than merely aggregating existing techniques, the framework formalizes a structured analytical pipeline that integrates domain-adapted NLP models with linguistically informed heuristics, context-sensitive segmentation procedures, and metrics grounded in syntactic and semantic structure. Its design emphasizes reproducibility by specifying each processing stage within a unified workflow that can be re-executed, extended, and audited. This abstraction enables the framework to accommodate different modeling choices, corpus configurations, and analytical goals, thereby supporting both replicable experimentation and comparative literary analysis.

The framework consists of three core components: (i) *character identification*, which identifies and extracts characters as entities within the text; (ii) *gender inference*, which infers character gender through linguistic and contextual cues; and (iii) *gender bias measuring*, which quantifies gendered asymmetries across linguistic patterns and computational representations. While these components can operate as a sequential, fully automated pipeline, we adopt the term *"framework"* to highlight its broader conceptual and methodological scope. In particular, the framework does not prescribe a mandatory execution order nor assume full automation; instead, it defines analytical components and interfaces that can be selectively combined, replaced, or partially instantiated depending on the research context.

This chapter directly addresses our first research goal, **RG1** (Section 1.2), which aims to develop a computational framework for measuring and analyzing gender bias in Portuguese-language literary texts. The framework presented here forms the conceptual foundation of this dissertation, providing the methodological structure that supports all subsequent analyses. Its modular organization establishes the analytical dimensions through which gendered patterns are systematically investigated in the chapters that follow. The initial formulation and evaluation of this framework resulted in the conference paper [33]; the current chapter expands that contribution by offering a unified theoretical and computational account of the framework, clarifying its design principles, scope, and applicability for large-scale literary analysis.

The remainder of this chapter is organized as follows. Section 3.1 formally defines the problem and the scope of gender bias analysis addressed in this dissertation. Sections 3.2, 3.3, and 3.4 detail the three main components of the framework, outlining their motivations and methodological choices. Section 3.5 integrates these components into a unified workflow, illustrating how they interact within the complete framework.

3.1 Problem Definition

This dissertation addresses the problem of computationally detecting and analyzing gender bias in Portuguese-language literary texts. In such texts, gender bias manifests as systematic asymmetries in how male, female, and gender-diverse entities[1] are represented, described, and related throughout narratives. These asymmetries may emerge across lexical, syntactic, and semantic levels,[2] reflecting broader social norms, cultural expectations, and ideological patterns encoded in language.

From a computational perspective, the central challenge lies in translating this socially grounded and interpretive phenomenon into measurable linguistic patterns. As summarized in the analytical mapping presented in Section 2.3.2, existing approaches to gender bias detection in textual data tend to converge on a three-stage workflow:

1. **Character Identification** – identifying textual entities (e.g., characters and named references) that anchor gender representation within the narrative.

[1] As explained in Section 2.1, a *character* can be defined as any entity—individual or collective, human or human-like—participating in the storyworld [49]; in this dissertation, we use entity as the general term following Computer Science traditional nomenclature (e.g., Named Entity Recognition).

[2] Linguistic levels refer to: (i) the lexical level, concerning word choice; (ii) the syntactic level, involving grammatical structures and dependency relations; and (iii) the semantic level, addressing meanings and conceptual associations.

2. **Gender Inference** – inferring the gender of identified entities, a task complicated in literary texts by ambiguity, metaphorical language, and stylistic variation.

3. **Gender Bias Measuring** – quantifying gendered asymmetries through linguistic, relational, and representational features, enabling comparative analysis across texts, authors, and temporal contexts.

Formally, let a literary corpus be defined as $C = \{t_1, t_2, \ldots, t_n\}$, where each t_i denotes a literary text written in Portuguese. The goal is to define a function:

$$B : C \to \mathbb{R}^k$$

that maps each text $t_i \in C$ to a multidimensional vector of bias indicators $b_i = \{b_{i1}, \ldots, b_{ik}\}$, where each component b_{ij} captures a distinct dimension of gender bias (e.g., distribution of male vs. female entities, adjective co-occurrence patterns, verb-based agency). The dimensionality k is not fixed a priori. Instead, it depends on the set of linguistic and representational features instantiated in a given analysis, such as the number of lexical categories, syntactic relations, or bias metrics considered. In this sense, the framework specifies the *structure* of bias representation rather than enforcing a single, fixed-dimensional vector.

To illustrate, consider a toy example in which a literary text t is analyzed using three bias indicators: (i) the proportion of male versus female characters, (ii) the relative frequency of adjectives associated with female characters, and (iii) an agency score derived from subject–object asymmetries. In this case, $B(t)$ would yield a three-dimensional vector, such as:

$$b = [0.65,\ 0.72,\ -0.31],$$

where each value corresponds to one operationalized bias dimension. More comprehensive analyses may include additional features (e.g., specific adjective classes, verb categories, embedding-based association scores), resulting in higher-dimensional vectors. Importantly, this variability is a design choice: by allowing the feature space to expand or contract according to the analytical goals, the framework supports extensibility, comparative analysis, and adaptation to different corpora or research questions.

It is important to note that this dissertation does not assume that gender bias can be reduced to a single numerical score. Instead, it treats bias as a multifaceted construct, observable through a combination of linguistic features. Accordingly, the proposed framework does not aim to issue normative judgments or causal explanations, but to provide a structured, transparent, and reproducible methodology for uncovering, describing, and comparing gendered patterns in Portuguese-language literary works.

3.2 Character Identification

Identifying literary characters is the first and most fundamental step in the proposed framework for measuring gender bias. Characters constitute the primary narrative agents through which actions, descriptions, and relationships are articulated; thus, gender representations in literature emerge largely through how characters are linguistically constructed. From a computational perspective, this stage can be modeled as the task of NER, which identifies and extracts named entities from text [151, 152]. Although the literary notion of character may extend beyond human individuals, this framework operationalizes character identification through the detection of `PERSON` entities, which serve as a computational proxy for narrative agents in the subsequent analyses.

Although NER is a well-established task in Natural Language Processing, its application to literary texts, particularly in Portuguese, poses distinct challenges. Literary texts often employ complex, stylistically rich language and reflect diverse cultural and historical contexts [153]. Characters may appear under multiple surface forms (e.g., first names, surnames, diminutives, titles of address), and personified entities or symbolic figures may be mentioned in ways that blur the boundary between literal and metaphorical reference. Moreover, Portuguese poses additional difficulties due to morphological richness, free or flexible word order, and extensive use of honorifics and epithets, all of which affect the reliability of token-level predictions.

Recent research emphasizes the effectiveness of domain-adaptive pre-training and fine-tuning strategies for improving NER performance on specialized textual domains. Pre-training language models on in-domain corpora enables them to capture domain-specific linguistic nuances and terminologies, leading to significant gains in accuracy [154]. For instance, Bamman et al. [139] introduced a corpus of 100 English literary texts annotated for named entities, showing that domain-adapted models improved F1 scores by over 20 absolute points. These findings highlight the importance of tailoring language models to literary data before applying them to tasks such as character identification.

The European Literary Text Collection (ELTeC) [155] has further contributed to advances in literary NER. This multilingual corpus contains annotated literary texts in more than ten European languages, including Portuguese. Santos et al. [152] describe the creation of the Portuguese subset (ELTeC-por) and introduce PALAVRAS-NER [143, 156], a rule-based NER system grounded in linguistic grammar rules. PALAVRAS-NER shows the feasibility of rule-based character extraction for Portuguese, but its deterministic architecture and reliance on handcrafted rules limit its scalability and adaptability across literary corpora [153]. These limitations motivate the exploration of data-driven approaches, particularly transformer-based models such as BERT, which can learn domain-specific patterns directly from annotated literary corpora.

Building upon these foundations, this dissertation develops and evaluates NER models adapted to the Portuguese-language literary domain. By fine-tuning BERT-based architectures on literary corpora, our approach aims to capture implicit, context-dependent, and stylistically variable character references that rule-based systems often fail to detect. The resulting character identification module serves as the foundation for the subsequent *Gender Inference* step, which infers the gender of each identified entity.

3.3 Gender Inference

Once literary characters have been identified, the next stage in the framework involves defining their gender marking.[3] This task, referred to as gender inference, is essential for analyzing how gendered patterns manifest across narrative discourse. Accurate gender inference enables the measurement of representational asymmetries, such as the proportion of male versus female characters, the distribution of actions and attributes, or linguistic patterns of association. In literary analysis, where gender roles and identities are frequently implicit rather than explicitly stated, this step provides the conceptual and computational bridge between character recognition and the quantification of gender bias.

Indeed, automatic gender detection has received growing attention in recent years. However, most research has focused on author profiling, which aims to infer the gender of a text's author based on linguistic patterns, writing style, or word choices [157, 158]. Although relevant for social media analysis and behavioral studies [159], author profiling differs fundamentally from the task addressed here. Our interest lies in entity-based gender detection, which seeks to infer the gender of individuals mentioned within the text rather than the author. This distinction is critical: whereas author profiling relies on features distributed across an entire text unit, entity-based gender inference requires high-resolution, context-sensitive analysis of specific textual mentions.

Entity-based gender detection is often integrated with NER tasks, where `PERSON` entities are classified by gender. Most research in this area, however, has concentrated on English-language datasets, which benefit from large-scale annotated resources and pre-trained models for gendered entity recognition [160, 161]. For example, Das and Paik [160] introduced four datasets for automatic gender identification and proposed a supervised transformer-based approach for identifying the gender of named entities. Then, they developed a two-stage model that infers gender without requiring explicit name-gender annotations, leveraging coreference resolution to capture context [161].

Although these methods perform well in English, they cannot be directly transferred to languages with distinct grammatical gender markers, agreement rules, and cultural conventions, such as Portuguese. In Portuguese, most research has addressed author profiling [157, 158] or relied solely on name-based gender prediction [162]. While effective

[3] In this dissertation, gender inference refers to identifying how characters are linguistically marked in terms of gender within the text, rather than inferring their personal gender identity.

in specific contexts, these approaches are limited for literary texts, where character gender is often implicit and inferred from context, honorifics, pronouns, or linguistic markers rather than explicitly stated.

To address these limitations, rule-based heuristics provide a promising alternative. These methods exploit linguistic cues, including pronouns [12, 14, 108, 109, 120] and gendered nouns [111, 123, 126, 138], enabling context-sensitive gender inference without requiring large annotated corpora. Building on this paradigm, our framework introduces a hybrid approach that integrates name-based prediction with context-based inference components. This design supports context-aware gender inference, enabling scalable and interpretable analysis of gender representation in Portuguese-language literary corpora.

3.4 Gender Bias Measuring

Once characters are identified and their genders inferred, the final stage of the proposed framework focuses on measuring and analyzing gender bias. As discussed in Section 2.3.2.4, prior work in literary studies has conceptualized gender bias through multiple analytical categories that capture different dimensions of representation, including *Authorship*, *Character Traits*, *Language*, *Relationships*, *Representation*, *Temporal*, and *Theme*. These categories span a wide range of phenomena, from authorship demographics and publication context to narrative roles, character networks, and thematic framing.

However, from a computational perspective, not all of such dimensions are equally manageable. Several categories depend on external metadata (e.g., author information, publication year, literary genre), manual interpretation (e.g., personality traits or themes), or high-level semantic judgments that are difficult to operationalize reliably at scale, especially in literary corpora with limited annotation. Rather than attempting to approximate all conceptual categories computationally, this dissertation adopts a focused and methodologically grounded strategy.

Therefore, our framework focuses on the *Language* category, which explores how gender is encoded directly in linguistic structure. This category captures systematic asymmetries that are observable in the text itself—independently of external metadata—and can be measured through reproducible computational procedures. By focusing on language, the framework prioritizes analytical dimensions that are (i) empirically observable, (ii) linguistically interpretable, and (iii) scalable across large literary corpora. This choice does not imply that other analytical categories are less relevant for understanding gender bias in literature, but rather reflects a methodological decision to prioritize dimensions that can be robustly operationalized and compared at scale.

From a computational perspective, gender bias measurement can be addressed along two complementary dimensions [44]: (i) *Gender Bias in Language*, which investigates asymmetries in how male and female characters are represented and described within the texts themselves; and (ii) *Gender Bias in Methods*, which probes whether computational models and embedding spaces reproduce or amplify gender stereotypes in their learned representations. Accordingly, this stage of the framework evaluates bias in two complementary domains: within the language of the texts (Section 3.4.1), and within the computational methods (Section 3.4.2).

3.4.1 Gender Bias in Language

Analyses of gender bias in the language aim to uncover how linguistic structure encodes asymmetries in representation, description, and agency across male and female characters. As established in the related work (Section 2.3.2.4), language-based analyses constitute one of the most extensively explored approaches to gender bias in literature. It captures the lexical, syntactic, and semantic patterns through which narratives represent gender and reproduce or challenge social stereotypes.

Previous studies have primarily relied on lexical features, such as adjectives and verbs, to investigate gendered patterns. Adjectives, for instance, reveal how descriptive traits are unevenly distributed across genders, often reflecting culturally entrenched stereotypes [10, 11, 19, 20, 23, 32, 45, 46, 103, 111, 122, 125]. Verbs and predicates, in turn, have been shown to capture differences in narrative agency, with male characters more frequently associated with action-oriented processes and female characters with passive or relational ones [10, 18, 45].

While informative, analyses based solely on surface lexical counts are limited: they capture which words co-occur with characters, but not how those words function syntactically in the sentence. In literary texts—where word order is flexible, modifiers may be distant from their heads, and grammatical relations carry narrative meaning—syntactic structure is essential for interpretation [10, 15, 32].

To address this limitation, the framework relies on dependency parsing rather than part-of-speech (POS) tagging alone. POS tagging assigns a grammatical category to individual tokens (e.g., noun, verb, adjective), but it does not encode the relational structure between words [163]. Dependency parsing, in contrast, models directed grammatical relations between tokens, explicitly representing who acts upon whom, who is described by which attribute, and how entities are syntactically positioned within events.

This distinction is crucial for gender bias analysis. Two characters may co-occur with the same verb, but occupy different syntactic roles (e.g., agent vs. patient), leading to fundamentally different narrative interpretations [14, 15, 45]. Dependency relations, therefore, enable a structural, rather than purely lexical, characterization of gendered patterns. Within this framework, gender bias in language is operationalized through a set of lexically grounded features extracted via dependency relations. Specifically, four syntactic relations are considered:

- `nsubj` (nominal subject), which captures characters occupying agentive positions in events and serves as a proxy for narrative agency;
- `obj` (direct object), which identifies characters that are acted upon, supporting analyses of passivity, exposure, or marginalization;
- `amod` (adjectival modifier), which links characters to descriptive properties and is central to the study of stereotyping; and
- `nmod` (noun modifier), which captures relational phrases (e.g., *filha de*, *senhor de*, *amigo de*) and role-denoting nouns (*médico*, *governanta*), situating characters socially and symbolically.

These relations are selected because they are (i) cross-linguistically stable within Universal Dependencies;[4] (ii) theoretically grounded in literary and linguistic analyses of gender; and (iii) consistent with prior English-focused work [18, 23, 46], allowing conceptual comparability while also accommodating the morphological characteristics of Portuguese. From these dependency relations, three classes of lexical features are extracted:

- **Adjectives**, which encode evaluative and descriptive attributes and are central to analyses of stereotyping;
- **Verbs**, which capture actions and processes associated with characters, enabling the study of agency and role asymmetries; and
- **Nouns**, which denote social roles, relational identities, and symbolic attributes, complementing adjectival and verbal analyses.

Although prior studies have explored gender bias in Portuguese-language literary texts using lexical features, these analyses are often predominantly manual, qualitative, or based on simple frequency counts [10, 32, 50, 96–98]. The contribution of this framework is therefore not merely in adapting existing methods to Portuguese, but in formalizing and scaling gender bias analysis for this linguistic context.

By combining dependency parsing with gender annotations, the framework measures the distribution of these lexical features across male and female entities. The resulting analyses allow for both aggregate comparisons (e.g., frequency distributions and co-occurrence patterns) and more nuanced interpretations of linguistic framing and character portrayal. In doing so, it advances the state of the art by transforming elements that were previously studied manually into a systematic computational procedure, enabling cross-work and cross-period comparability and positioning Portuguese within methodological debates that have been shaped primarily by English-centric research.

[4] https://universaldependencies.org/

3.4.2 Gender Bias in Methods

While gender bias in language reflects how social stereotypes are linguistically encoded in texts, gender bias in methods concerns how such associations are internalized, amplified, or reshaped through computational representations [44]. This dimension thus shifts the analytical focus from the textual surface to the representational behavior of computational models, probing how they learn, encode, and reproduce gendered semantics.

Research on this topic has developed a broad set of metrics for evaluating representational disparities in language models. At the representational level, embedding-based metrics examine geometric relations between gendered and gender-neutral terms to reveal latent structures that align with gender subspaces [26, 67]. WEAT [67] remains one of the most influential approaches, quantifying associative strengths between sets of gendered and attribute terms using cosine similarity. Subsequent extensions, such as SEAT [69] and CEAT [70], adapted it to sentence-level embeddings and contextual encoders.

Beyond embeddings, probability-based metrics directly probe a model's likelihood distributions to measure asymmetries in the probability assigned to gendered alternatives under identical contexts [53, 73]. Masked-token methods estimate the probability of filling masked words with gendered terms, while pseudo-log-likelihood approaches compare the perplexity of stereotypical and anti-stereotypical sentences in autoregressive models. Together, these metrics expose disparities in how models assign linguistic expectations across gendered contexts.

A complementary perspective comes from generated-text-based metrics, which assess bias at the level of model outputs [81, 82]. These methods evaluate behavioral manifestations of bias in generated text, such as differences in sentiment, toxicity, or lexical choice, revealing how stereotypes can emerge during text generation. However, such methods depend heavily on external classifiers or lexicons and are less interpretable in controlled literary settings.

Building upon these insights, the proposed framework assesses gender bias in computational methods across two complementary analyses: (i) representational bias in contextual embeddings, capturing learned semantic associations between gendered entities and descriptive attributes, and (ii) behavioral bias in generated text, assessing how models reproduce or transform gendered patterns when generating literary or near-literary continuation. By combining embedding- and generation-based analyses, this stage provides a comprehensive view of how gender bias persists and transforms across levels of model representation. It bridges intrinsic and extrinsic evaluation, linking the statistical learning of gendered patterns to their narrative manifestations in generated text.

Although most metrics used in this stage were originally developed for English, applying them to Portuguese introduces challenges and opportunities that have not been systematically addressed in prior work [31, 149], particularly in the literary context. Portuguese models differ from English ones in their treatment of grammatical gender, particularly due to rich morphological marking (e.g., agreement in adjectives, determiners, participles), highly productive gendered derivational patterns, and the presence of null subjects that interact with gender inference. These linguistic characteristics shape how gender is encoded in embeddings and generated text, producing bias patterns that existing English-focused evaluations cannot capture without adaptation.

3.5 Framework Overview

PORTALIA (**POR**tuguese **T**ext **A**na**L**ysis for Gender-b**IA**s) is a modular computational framework designed to measure and analyze gender bias in Portuguese-language literary texts. Rather than prescribing a single fixed processing chain, *PORTALIA* defines a structured set of interoperable components, each with explicit inputs, outputs, and assumptions, which can be instantiated, replaced, or extended depending on the corpus, research question, and computational resources.

The framework integrates three core methodological dimensions introduced in the previous sections: (i) *Character Identification*, which extracts and contextualizes literary entities; (ii) *Gender Inference*, which assigns gender labels based on linguistic evidence; and (iii) *Gender Bias Measuring*, which quantifies asymmetries in both textual language and computational representations.

By integrating these components into a unified framework, *PORTALIA* tackles the central research question of this dissertation (**RG1**) and addresses key methodological limitations identified in prior work, including insufficient contextualization of character mentions, the absence of tools adapted to Portuguese morphology, and the lack of a comprehensive framework bridging linguistic analysis and computational modeling.

3.5.1 Workflow Description

Figure 3.1 presents the high-level workflow, composed of five conceptually ordered steps. The workflow begins with the ingestion of one or more plain-text literary works, followed by a *Preprocessing* stage consisting of text cleaning and sentence segmentation. These steps standardize the input, resolve formatting inconsistencies, and enable reliable sentence-level analysis. Although essential for the workflow, preprocessing is a conventional step in NLP pipelines and does not constitute a novel methodological contribution; therefore, detailed implementation notes are presented in Appendix B.

After *Preprocessing*, the *Character Identification* module extracts all mentions of literary characters identified as `PERSON` entities. Using the preprocessed sentences, the framework identifies both proper names and nominal references that denote literary characters. As discussed in Section 3.2, *PORTALIA* employs a fine-tuned NER model trained

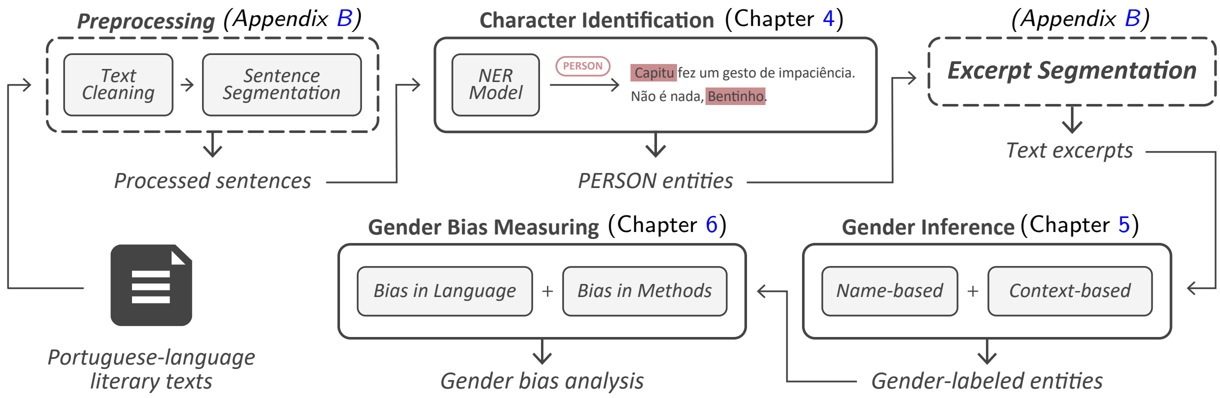


Figure 3.1: High-level overview of the proposed computational framework *PORTALIA* (**POR**tuguese **T**ext **A**na**L**ysis for Gender-b**IA**s). Solid and dashed boxes indicate core and intermediate stages, respectively.

to recognize PERSON entities in literary corpora, enabling large-scale, automated processing. Although the model is trained to recognize multiple entity classes, the present analysis focuses exclusively on the PERSON class to capture character mentions.

The subsequent *Excerpt Segmentation* module, as shown in Figure 3.1, defines local contextual windows around each character mention. These excerpts form the analytical units from which lexical features are extracted and attributed to specific entities. Rather than being a function of text length, excerpt segmentation reflects an explicit choice of *unit of analysis*. In prior work on gender bias and character representation, this choice is often implicit: most studies operate at the word or token level [20, 32, 110], others aggregate features at the sentence level [12, 14, 18, 46]. Although these approaches differ in scope, they all implicitly define a contextual boundary within which linguistic evidence is associated with characters.

Although essential for structuring later stages of the framework, excerpt segmentation does not constitute a novel methodological contribution. Window-based contextualization strategies are well established in corpus linguistics and computational text analysis. Our contribution lies in making this step explicit, formally defined, and reproducible within a unified framework for analyzing literary gender bias. Appendix B documents the full procedure to support transparency and reproducibility.

Once excerpts are defined, the *Gender Inference* module assigns gender labels to the identified entities based on a heuristic that combines two complementary components. The first component relies on proper names to infer the most likely gender associated with each character name. The second component uses syntactic cues extracted from the excerpt, including agreement patterns of determiners, adjectives, and participles that co-occur with the entity. These morphosyntactic signals capture gender information even in cases where names are ambiguous, rare, or absent. This approach combines lexical and grammatical evidence to handle the specificities of Portuguese morphology (Section 3.3).

Finally, the *Gender Bias Measuring* module (Section 3.4) performs gender bias measurement across gendered asymmetries across two complementary domains. The first domain, *bias in language*, focuses on how male and female characters are represented linguistically within the literary texts. The second domain, *bias in methods*, probes how computational models encode, amplify, or reshape these gendered patterns through embeddings and language model generations. This includes analyses of both word embeddings and large language models that have been trained on or exposed to literary texts. Together, these two domains provide a comprehensive view of gender bias, capturing its manifestation in both texts and the computational tools used to study them.

Although the components of *PORTALIA* can be executed sequentially, the framework is intentionally defined at a higher level of abstraction than a traditional pipeline. A pipeline typically enforces a fixed order of operations with tightly coupled components, whereas *PORTALIA* specifies: (i) explicit interfaces between components, (ii) well-defined input and output representations, and (iii) minimal dependency constraints between modules. As a result, each component can be evaluated independently, replaced by alternative implementations, or reused in isolation. For example, the *Gender Bias Measuring* module can be applied to externally annotated corpora, and the *Gender Inference* module can operate on character mentions obtained from other NER systems.

The next chapters provide detailed descriptions of the three core components: *Character Identification*, *Gender Inference*, and *Gender Bias Measuring*. Intermediate steps such as *Preprocessing* and *Excerpt Segmentation*, while essential for ensuring consistent and reproducible analyses, are described in Appendix B, as they constitute operational rather than conceptual contributions.

3.5.2 Design Principles and Contributions

The architecture of *PORTALIA* framework is guided by five design principles as follows.

1. **Adaptation to Portuguese-language literary texts**: the framework is grounded in the linguistic characteristics of Portuguese-language literary texts.

2. **Modularity and extensibility**: each component is self-contained and can be replaced, extended, or independently evaluated, supporting both replication and future methodological advances.

3. **Reproducibility**: all stages are documented and publicly available, enabling full replication, transparent auditing, and application to new corpora.

4. **Large-scale automated analysis**: the framework is designed to support distant reading by processing large-scale corpora, enabling systematic analysis that would be infeasible through manual close reading alone.

5. **Comparability**: by standardizing analytical units (entities and excerpts), gender labels, and procedures, the framework allows comparisons across works, authors, periods, and subgenres, facilitating both corpus-level and longitudinal studies.

Based on these principles, the contributions of *PORTALIA* include: (i) a novel computational framework specifically adapted to the linguistic properties of Portuguese and literary texts; (ii) a modular workflow that integrates *Character Identification*, *Gender Inference*, and *Gender Bias Measuring* in a unified and extensible architecture; (iii) NER models fine-tuned on Portuguese-language literary corpora, enabling accurate and context-sensitive character identification; (iv) a hybrid, linguistically grounded approach for gender inference in Portuguese, combining onomastic cues with morphosyntactic signals extracted from dependency structures; (v) a standardized setup for cross-text and diachronic comparisons, allowing reproducible and interpretable studies of gender bias across diverse Portuguese-language literary corpora.

3.5.3 Modularity and Extensibility

The modular structure of *PORTALIA* ensures that each module is both logically independent and computationally interoperable. Dependencies are minimal and transparent: character identification requires only preprocessed text; excerpt segmentation requires sentences and entity spans; gender inference requires excerpts; and bias measurement requires gender-labeled excerpts. This design enables, among others:

- substitution of the NER model by domain-specific alternatives (e.g., historical or contemporary works, children's literature);
- incorporation of new lexical features (e.g., semantic role labeling, coreference resolution); and
- integration of additional bias metrics or LLM-based probing methods.

To support transparency and reuse, the full implementation of the framework is publicly available at: `https://github.com/marianaossilva/gender_pipeline`.

Chapter 4

Identifying Literary Characters

The first module of *PORTALIA* addresses the task of identifying literary characters, a prerequisite for all subsequent analyses of gender inference and gender bias within the proposed framework. In this dissertation, character identification is treated not as an end in itself, but as an enabling operation that anchors narrative entities to which gender labels and linguistic features can be systematically assigned.

From a computational perspective, this task is operationalized as a specialization of Named Entity Recognition, focusing on the extraction of `PERSON` entities from literary texts. While NER is a well-established task in NLP, its application to literary texts poses additional challenges due to stylistic variation, implicit references, and evolving character identities. These challenges are further amplified in Portuguese-language literature, which exhibits rich morphology, frequent pronominal anaphora, and subject ellipsis, increasing the difficulty of reliable character extraction.

Recent advances in transformer-based language models have substantially improved performance in entity recognition tasks. However, models pre-trained on general-domain corpora often fail to capture the stylistic, narrative, and historical specificities of literary prose. This domain mismatch motivates the investigation of domain-adaptive strategies that incorporate literary data during pre-training and fine-tuning, aiming to improve robustness and generalization in literary character identification.

Within *PORTALIA*, after *Preprocessing* steps, *Character Identification* is implemented as a NER-based module that extracts mentions of literary characters from pre-

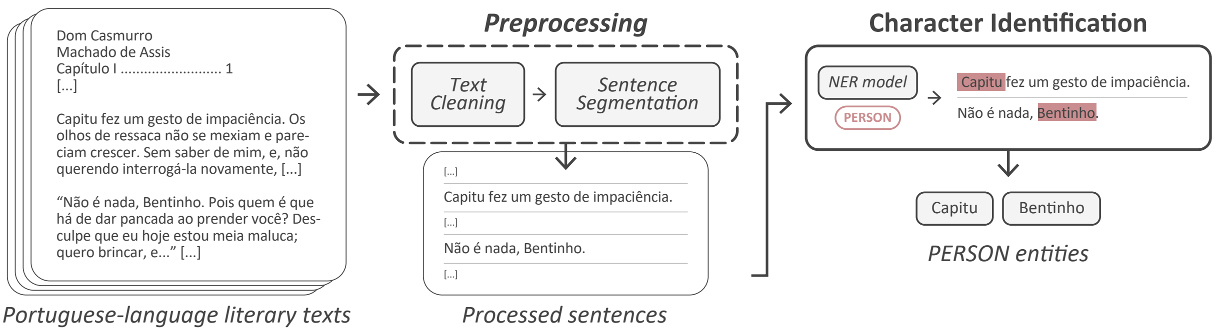


Figure 4.1: Overview of the Character Identification module, which extracts `PERSON` entities from preprocessed sentences using a NER model.

processed sentences, focusing on the `PERSON` entity class, as illustrated in Figure 4.1. Although other entity types are preserved for corpus consistency (e.g., `LOC`, `ORG`, `DATE`), only `PERSON` entities are considered in the analyses presented in this chapter, as they directly support character-centered gender analysis.

Rather than proposing a new NER formulation, this chapter evaluates how different pre-training strategies affect character identification performance in Portuguese-language literary texts. In particular, it assesses the impact of continued pre-training on literary corpora combined with fine-tuning on annotated literary data. After outlining relevant pre-training strategies (Section 4.1), the main contributions of this chapter are:

- the creation of *PPORTAL_ner* (Section 4.3), an annotated corpus of Portuguese-language literary entities;
- the development of two domain-adaptive models, *LitBERT-CRF* and *LitBERTimbau*, which incorporate literary data during pre-training and fine-tuning (Sections 4.2 and 4.4); and
- a comparative evaluation against a general-domain baseline, highlighting the benefits and trade-offs of different adaptation strategies (Section 4.5).

These contributions address the second research goal of this dissertation (**RG2**, Section 1.2) and have resulted in three publications: [34], [35], and [36]. An error analysis is provided in Appendix D, and Section 4.6 summarizes the main findings and their implications for subsequent stages of the *PORTALIA* framework.

4.1 Pre-training Strategies

Language models pre-trained on general-domain corpora often underperform in specialized domains due to domain-specific linguistic, stylistic, and contextual variations. To mitigate this limitation, prior work has explored pre-training strategies that adapt models to target domains by incorporating in-domain data at different stages of training.

Two main strategies are commonly adopted. The first is *training from scratch*, in which a model is initialized with random weights and trained exclusively on a large domain-specific corpus. Although this approach can yield highly specialized representations, it requires substantial computational resources and large-scale data, making it impractical for many low-resource domains, including literary texts.

The second strategy is *domain-adaptive pre-training*, which continues the pre-training of an existing general-purpose language model using unlabeled domain-specific data. This approach is significantly more resource-efficient and has been shown to improve downstream performance across a range of specialized domains while preserving the general linguistic knowledge acquired during initial pre-training [154, 164].

Table 4.1: Overview of the pre-training corpus.

Title	Author	Lang.	Year	Genre	Movement	#Tokens
A Cidade e as Serras	José Maria Eça de Queirós	pt	1901	Novel	Realism	83,280
A Escrava Isaura	Bernardo Guimarães	pt-br	1875	Novel	Romanticism	62,358
A Relíquia	José Maria Eça de Queirós	pt	1887	Novel	Realism	103,808
Cinco Minutos	José de Alencar	pt-br	1856	Novel	Romanticism	17,041
Clepsidra	Camilo Pessanha	pt	1920	Poems	Symbolism	6,033
Do Livro do Desassossego	Fernando Pessoa	pt	1982	Novels	Modernism	7,552
Iaiá Garcia	Machado de Assis	pt-br	1878	Novel	Realism	68,192
Noite na Taverna	Álvares de Azevedo	pt-br	1855	Short stories	Romanticism	22,380
O Guarani	José de Alencar	pt-br	1857	Novel	Romanticism	124,540
Senhora	José de Alencar	pt-br	1875	Novel	Realism	88,604
					Total	583,788

A complementary and widely used approach is *cross-domain transfer learning*, in which a pre-trained model is fine-tuned directly on annotated data from the target task. This strategy is particularly effective when labeled data are scarce, as it allows models to leverage representations learned from large source-domain corpora while adapting to task-specific supervision [165].

In this chapter, we focus on two strategies that balance effectiveness and feasibility for Portuguese-language literary texts: (i) domain-adaptive pre-training on unlabeled literary corpora, and (ii) cross-domain transfer learning via fine-tuning on annotated literary data. These strategies are selected to address the limited availability of labeled resources and the stylistic divergence between literary prose and the general-domain corpora used in standard language model pre-training.

4.2 Domain-Adaptive Pre-training

This section describes the domain-adaptive pre-training procedure adopted to tailor language models to Portuguese-language literary texts. We first present the literary corpus used for continued pre-training (Section 4.2.1), followed by the experimental setup (Section 4.2.2) and the resulting pre-trained models (Section 4.2.3).

4.2.1 Pre-training Corpus

For domain-adaptive pre-training, we curated a literary corpus from a subset of the *PPORTAL* repository [99, 166, 167], comprising 583,788 tokens from ten public-domain literary works. The selected texts span multiple authors, genres, and literary movements, aiming to capture stylistic and thematic diversity while remaining within copyright constraints. Table 4.1 summarizes the corpus composition. All texts are pre-processed to remove non-linguistic artifacts such as email addresses and URLs, while preserving punctuation and hyphenation, which are linguistically meaningful in literary prose.

Table 4.2: Hyperparameters used during pre-training.

Hyperparameter	**Value**
Learning rate	5×10^{-5}
Batch size	16
Max sequence length	512
Epochs	3
Masking probability	15%

The corpus includes novels, poetry, and short stories associated with movements such as Romanticism, Realism, and Modernism. However, due to Brazilian public-domain regulations, the temporal scope is largely restricted to the 19th and early 20th centuries. As a result, the corpus is biased toward canonical authors, formal registers, and historical language use, which may differ substantially from contemporary Portuguese. This limitation is inherent to the available data and is explicitly considered when interpreting downstream results.

4.2.2 Pre-training Setup

All models are continuously pre-trained using the Masked Language Modeling (MLM) task. In this setup, 15% of tokens in each sequence are masked and predicted by the model based on their surrounding context. Pre-training is conducted for three epochs, balancing computational feasibility with sufficient exposure to the domain-specific corpus. To ensure comparability across models, all experiments use the same hyperparameter configuration (Table 4.2). Model checkpoints are not evaluated intrinsically on the MLM task; instead, their effectiveness is assessed exclusively through downstream performance on literary NER, which is the target application of this chapter.

4.2.3 Novel Pre-trained Models

Based on the domain-adaptive pre-training procedure, we introduce two language models specifically adapted for Portuguese-language literary NER. Both models are pre-trained on the curated literary corpus using the MLM objective and are publicly available at `https://huggingface.co/marianaossilva`. Each model is briefly described as follows.

LitBERTimbau. This model builds upon BERTimbau-Base [168], a general-purpose Portuguese language model originally pre-trained on the brWAC corpus [169]. Domain-adaptive pre-training incorporates literary language patterns into the existing representation space, yielding a model better suited to narrative prose.

LitBERT-CRF. This model extends a BERT-CRF architecture [170], combining contextualized embeddings with a Conditional Random Field layer to improve sequence labeling consistency. The base checkpoint was pre-trained on brWAC and fine-tuned on the First HAREM corpus [171] before undergoing literary domain-adaptive pre-training. The CRF layer supports more structured entity predictions, which is particularly beneficial in complex narrative contexts.

4.3 Fine-tuning Corpus

Reliable evaluation and adaptation of NER models to literary texts require annotated corpora that reflect the narrative, stylistic, and referential properties of literature (see Appendix C.1). However, such resources remain scarce, particularly for Portuguese-language literary texts [50, 139]. As a result, most existing approaches rely on models trained and evaluated on general-domain corpora, whose entity distributions and discourse structure differ substantially from those found in literary prose.

Literary texts typically emphasize characters and settings rather than organizations or geopolitical entities, and they frequently employ indirect references, descriptive noun phrases, and stylistic variation [139, 153]. These characteristics challenge standard NER assumptions and limit the applicability of general-domain datasets for both training and evaluation. Without in-domain annotated data, it is difficult to assess the effectiveness of domain-adaptive strategies or to perform meaningful error analysis [172, 173].

These challenges are particularly pronounced for Portuguese-language literature [174, 175]. While similar issues have been documented in English, Portuguese introduces additional linguistic complexities, such as gender and number agreement, rich morphological inflection, and regional variation across Brazil and Portugal [50]. Such characteristics exacerbate the limitations of general-purpose or cross-lingually transferred models, highlighting the need for dedicated resources.

To address this gap, we introduce *PPORTAL_ner*,[1] an annotated corpus of literary entities designed to support research and model development for Portuguese-language literary texts. *PPORTAL_ner* contains 25 literary works, totaling 125,059 tokens and 5,266 annotated entities distributed across five categories: `PERSON`, `LOC`, `GPE`, `ORG`, and `DATE` (Table 4.3). A detailed corpus description, annotation format, and evaluation results are provided in Appendix C.

4.3.1 *PPORTAL_ner* Corpus

The *PPORTAL_ner* corpus is derived from *PPORTAL* [167], an extensive metadata repository comprising over 80,000 public-domain Portuguese-language literary works and more than 9,000 downloadable texts [99, 166]. *PPORTAL* aggregates works from three major digital libraries: Domínio Público,[2] Projecto Adamastor,[3] and the Biblioteca Digital

[1] The corpus is available for download at [176].
[2] https://www.dominiopublico.gov.br
[3] https://projectoadamastor.org

Table 4.3: Statistics of the *PPORTAL_ner* corpus.

Title	**Author**	**Language**	**Year**	**#Tokens**
Menina e Moça	Bernardim Ribeiro	pt	1554	5,004
Os Lusíadas	Luís Vaz de Camões	pt	1572	5,000
Eurico, o Presbítero	Alexandre Herculano	pt	1844	5,000
Memórias de um Sargento de Milícias	Manuel Antônio de Almeida	pt-br	1854	5,000
Amor de Perdição	Camilo Castelo Branco	pt	1861	5,001
Iracema	José de Alencar	pt-br	1865	5,000
As Pupilas do Senhor Reitor	Júlio Dinis	pt	1867	5,000
A Morgadinha dos Canaviais	Júlio Dinis	pt	1868	5,000
Inocência	Visconde de Taunay	pt-br	1872	5,002
Helena	Machado de Assis	pt-br	1876	5,000
O Mandarim	Eça de Queirós	pt	1880	5,000
Memórias Póstumas de Brás Cubas	Machado de Assis	pt-br	1881	5,002
O Alienista	Machado de Assis	pt-br	1882	5,001
Casa de Pensão	Aluísio Azevedo	pt-br	1884	5,000
O Cortiço	Aluísio Azevedo	pt-br	1890	5,000
Quincas Borba	Machado de Assis	pt-br	1891	5,000
Dom Casmurro	Machado de Assis	pt-br	1899	5,003
Os Sertões	Euclides da Cunha	pt-br	1902	5,000
Esaú e Jacó	Machado de Assis	pt-br	1904	5,000
Cartas de Inglaterra	Eça de Queirós	pt	1905	5,000
Memorial de Aires	Machado de Assis	pt-br	1908	5,001
A Confissão de Lúcio	Mário de Sá-Carneiro	pt	1913	5,000
Alves & Companhia	Eça de Queirós	pt	1925	5,000
Capitães da Areia	Jorge Amado	pt-br	1937	5,001
Vidas Secas	Graciliano Ramos	pt-br	1938	5,001
			Total	**125,014**

de Literatura dos Países Lusófonos (BLPL),[4] providing a broad and curated source of Portuguese-language literary data.

The design of *PPORTAL_ner* prioritizes diversity of authorship, genre, and narrative style over strict chronological representativeness. All selected works were published before 1953 to comply with Brazilian public-domain regulations. The final sample spans texts published between 1554 and 1938, covering major literary movements such as Romanticism, Realism, and Modernism, and thus exposing models to heterogeneous stylistic and linguistic conventions.

Each text underwent pre-processing to ensure data quality and consistency, including the removal of special characters (while preserving punctuation and hyphenation, which are semantically relevant in literary prose) and elimination of metadata artifacts. From each work, a segment of approximately 5,000 words was selected, which corresponds, on average, to two chapters. This length was chosen to preserve sufficient narrative continuity to capture recurring entities and stylistic features, while keeping the annotation process manageable and consistent across texts of varying lengths.

4.3.2 Annotation Process

The annotation process for *PPORTAL_ner* is guided by three core narrative questions: *who?*, *where?*, and *when?*. These fundamental questions guide the identification of entities

[4] https://www.literaturabrasileira.ufsc.br

Table 4.4: Distribution of entity categories.

Category	Frequency (%)	Examples
PER	3,609 (68.53%)	"Capitu", "the foreigner", "the youngest son"
LOC	1,126 (21.38%)	"the village", "the town", "under the bridge"
GPE	315 (5.98%)	"Brazil", "Lisbon", "Rio de Janeiro", "Europe"
ORG	115 (2.18%)	"the police", "the Church", "the army"
DATE	101 (1.92%)	"XVIII century", "1847", "the winter"

within the texts and are central to understanding the relationships and dynamics between characters, settings, and temporal markers in literary works. By focusing on these critical aspects, we aim to create an annotation schema that balances the richness of literary language with the structured needs of NER models.

The annotation guidelines for *PPORTAL_ner* are specifically designed to reflect the complexity of literary language. They include both proper names and descriptive nominal expressions used to refer to entities (e.g., "the foreigner", "the village"). This broader approach differs from corpora such as HAREM [171], which focus primarily on proper names, and aligns more closely with the entity definition in ACE [177], which encompasses names, pronouns, and descriptive phrases.

4.3.2.1 Entity Categories

The annotation process focuses on five essential entity categories commonly used in most NER systems. This selection also allows us to evaluate NER models with respect to traditional named entity categories. Table 4.4 provides the distribution of entity categories along with illustrative examples.

PERSON. This category labels entities representing literary characters in the text. These may include proper names, such as "Capitu", as well as nominal references, including anaphoric noun phrases (e.g., "Bentinho's wife"). Furthermore, sets of people, such as family units (e.g., "Capitu's parents"), are included in this category, reflecting the different ways characters are referenced in literary texts.

LOC. This category labels entities that denote locations and spatial references within the narrative. It differs from the GPE category, which is reserved for real-world geographical locations (e.g., London, New York) and nations (e.g., England, the United States). The LOC category can include both named and more general imagined entities in a literary context, such as "the town" or "the village".

GPE. This category labels geopolitical entities, such as countries and cities, which are often central to understanding the narrative's social, cultural, and political contexts. For example, references to specific cities, like "Lisbon" or "Rio de Janeiro", or nations, such as "Brazil" or "Portugal", are marked as GPE.

`ORG`**.** This category labels organizations and institutions mentioned within the texts. In literary data, organizations are relatively rare and defined as formal associations, making this the least frequently occurring entity class in our corpus. Examples include references to "the Church", "the army", or "the police".

`DATE`**.** This category labels time-related entities, such as specific dates, time periods, and temporal expressions. Temporal markers help situate events within the narrative structure, linking them to historical moments. Examples of entities in this category include references such as "in 1920" or "XVIII century".

4.3.2.2 Annotation Guidelines

Our annotation process follows a structured set of guidelines to ensure consistency and accuracy in entity labeling. These guidelines follow a two-step, semi-automated approach, starting with pre-annotation using a pre-trained model, followed by a manual correction and refinement process to ensure high-quality annotations.

Pre-Annotation. The first step involves pre-annotating each of the 25 selected texts using the spaCy model `pt_core_news_lg`, a pre-trained language model designed for Portuguese. This model automatically suggests potential entities based on its existing knowledge of language patterns and common entity types. The pre-annotation provides an initial set of entity suggestions later reviewed and refined by a human annotator.

Correction and Refinement. After pre-annotation, the Prodigy annotation tool is used to refine the annotations. Specifically, the `ner.correct` recipe in Prodigy is employed to create a gold-standard dataset. Using the `--update` argument during the annotation loop, the model is continuously updated based on the received annotations, improving its accuracy over time. The boolean segmentation argument is set to `true` to facilitate the segmentation of lengthy literary texts into manageable sentences, streamlining the annotation process.

4.4 Fine-tuning Setup

For fine-tuning and evaluation, we use the *PPORTAL_ner* corpus introduced in Section 4.3. This corpus comprises 25 public-domain literary works and has no overlap with the data used during domain-adaptive pre-training (Section 4.2), ensuring a clear separation between adaptation and evaluation data.

Model Choice. Three models are fine-tuned on the *PPORTAL_ner* corpus: (i) LitBERTimbau, (ii) LitBERT-CRF, and (iii) BERT-CRF. LitBERTimbau and LitBERT-CRF are first domain-adapted through continued pre-training on a literary corpus (see Section 4.2). In contrast, BERT-CRF is a general-domain model fine-tuned directly on literary

Table 4.5: Overview of each fine-tuned model.

Model	**Strategy**	**Vocab**	$\mathbf{C_1}$	$\mathbf{C_2}$
BERT-CRF	Baseline	General	General	General
FT BERT-CRF	Cross-domain transfer learning	General	General	Literary
LitBERT-CRF	Domain-adaptive pre-training	General	Literary	Literary
LitBERTimbau	Domain-adaptive pre-training	General	Literary	Literary

Vocab: Vocabulary — $\mathbf{C_1}$**:** Pre-training corpus — $\mathbf{C_2}$**:** Fine-tuning corpus

data through cross-domain transfer learning. This configuration allows us to isolate the impact of domain-adaptive pre-training by comparing models that differ only in their exposure to literary language.

We deliberately focus on encoder-based models rather than instruction-tuned Large Language Models (LLMs). Although LLMs have demonstrated strong performance across a wide range of NLP tasks, they are less suitable for the controlled evaluation of sequence labeling tasks such as NER. In particular, instruction-following models typically operate in a generative setting, where outputs are sensitive to prompt design and decoding strategies, jeopardizing reproducibility and systematic comparison. By contrast, encoder-based architectures with token-level supervision provide stable training objectives, transparent error analysis, and well-established evaluation protocols.

Input Length and Text Segmentation. All models employed in this study accept a maximum input length of 512 tokens, following standard BERT-based configurations. Given that literary works are long documents, texts are not processed as whole documents but are segmented into sentence-level units before fine-tuning. Each sentence is treated as an independent training instance, preserving token-level alignment with entity annotations. This sentence-based segmentation avoids truncation of annotated spans and ensures that all entity mentions are fully observed during training and evaluation. While this strategy limits access to long-range discourse context, it is consistent with the formulation of NER as a local sequence labeling task and aligns with standard practice in both general-domain and literary NER research.

Fine-tuned Models. Table 4.5 summarizes the models evaluated in this chapter. The fine-tuned BERT-CRF model is referred to as FT BERT-CRF. For reference, we also report results for a general-domain BERT-CRF model originally fine-tuned on the HAREM corpus [171]. To ensure compatibility with the *PPORTAL_ner* schema, entity categories in HAREM are mapped to the closest corresponding labels (`GPE` merged into `LOC`, and `DATE` mapped to `TIME`).

Training Setup. All models are fine-tuned for ten epochs using identical hyperparameter configurations. This choice prioritizes comparability across models, ensuring that observed differences in performance stem from architectural design and pre-training strategy rather than task-specific optimization. Model selection is based on validation performance, monitored through loss and token-level accuracy.

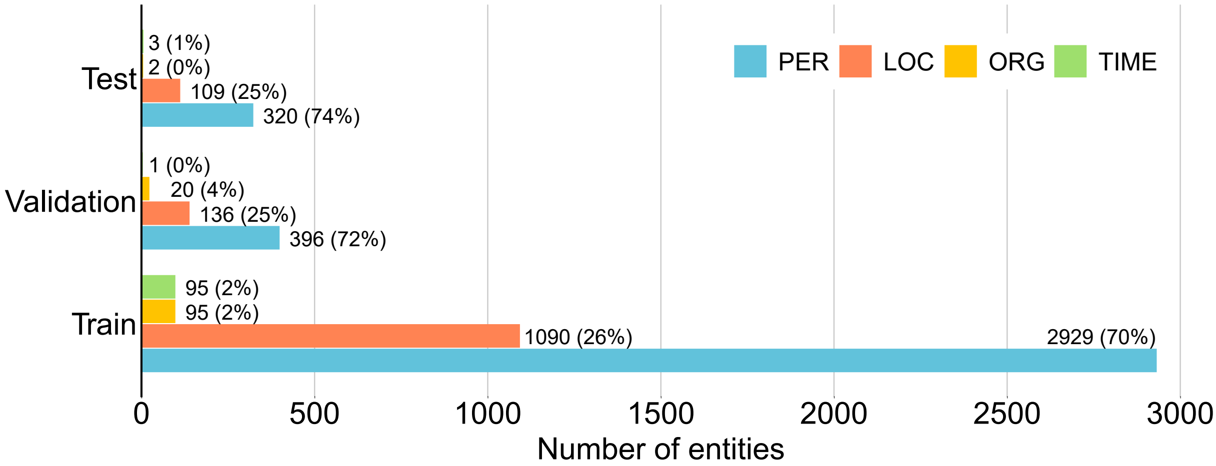


Figure 4.2: Distribution of entity categories across the training, validation, and test splits.

Table 4.6: Error types in NER evaluation.

Error type	**Description**
Correct (C)	True and predicted entities are equal
Incorrect (I)	True and predicted entities do not match
Partial (P)	True and predicted entities are similar
Missing (M)	A true entity that was not predicted
Spurious (S)	A predicted entity that does not exist

The annotated corpus is split into training (80%), validation (10%), and test (10%) sets, corresponding to 5,572, 696, and 697 sentences, respectively. As shown in Figure 4.2, the dataset exhibits class imbalance, with a strong predominance of `PERSON` entities. While this imbalance reflects the narrative focus of literary texts, it may affect performance on less frequent categories. Addressing this issue through resampling or cost-sensitive learning is left for future work.

4.5 Experimental Evaluation

This section presents the experimental evaluation conducted to compare the proposed pre-training strategies for character identification. Section 4.5.1 details the evaluation setup and metrics, while Section 4.5.2 discusses the results.

4.5.1 Evaluation Metrics

In NER evaluation, metrics are often reported at the token level. However, this approach may not always provide a comprehensive assessment, particularly for entities that span multiple tokens. To address this, we adopt the evaluation schema defined by the *SemEval 2013 - 9.1 task* [178], which extends beyond simple token- or tag-based schema. It evaluates multiple scenarios by checking whether all tokens forming a named entity are correctly detected and whether the appropriate entity type is assigned.

Table 4.7: Evaluation scenarios for NER evaluation.

Scenario	**Description**
Strict	Exact boundary and type matching
Type	Correct entity type assignment regardless of exact boundaries
Partial	Partial boundary matching, regardless of the entity type
Exact	Exact boundary matching, regardless of the entity type

Table 4.8: Evaluation results from the five types of errors and four scenarios. The test set used to evaluate the models has 434 annotated entities.

Model	**Strict**				**Type**				**Partial**				**Exact**			
	C	*I*	*M*	*S*	*C*	*I*	*M*	*S*	*C*	*P*	*M*	*S*	*C*	*I*	*M*	*S*
BERT-CRF	119	3	312	29	120	2	312	29	121	1	312	29	121	1	312	29
FT BERT-CRF	335	35	65	65	362	8	65	65	341	29	65	65	341	29	65	65
LitBERT-CRF	336	31	67	62	357	10	67	62	344	23	67	62	344	23	67	62
LitBERTimbau	333	33	68	84	358	8	68	84	341	25	68	84	341	25	68	84

C: Correct — *I:* Incorrect — *M:* Missed — *S:* Spurious — *P:* Partial

The considered schema includes five error types: Correct (C), Incorrect (I), Partial (P), Missing (M), and Spurious (S). Table 4.6 describes each error type. Additionally, four evaluation scenarios (Strict, Type, Partial, and Exact) are used to assess model performance under different conditions. Table 4.7 outlines these scenarios.

For automated evaluation, errors are calculated based on boundary matching by assessing overlaps between the true and predicted entities. Overlaps are determined by computing the intersection between the start and end offsets of true and predicted entities. For example, if the true entity spans tokens 3–7 and the predicted entity spans tokens 5–9, the overlap includes tokens 5–7. This method enables a nuanced evaluation of partial boundary matches without imposing rigid percentage thresholds.

4.5.2 Results and Analysis

Table 4.8 summarizes model performance across error types and evaluation scenarios. As expected, the number of Missing (M) and Spurious (S) entities remains constant across scenarios, since these error types are independent of boundary matching criteria and depend solely on whether an entity is predicted or not. Across all models, the results indicate a persistent trade-off between recall and precision: while fine-tuned models substantially reduce the number of missed entities compared to the baseline, spurious predictions remain a challenge.

Our results also reveal a consistent challenge for all models in accurately identifying entity types. This issue is particularly evident in the **Strict** scenario, where the stringent requirement for exact boundary and type matches leads to a higher Incorrect (I) error rate. In contrast, in the **Type** and **Partial** scenarios, which allow more flexible matching, the number of Correct (C) entities increases, suggesting that models perform better when less rigid restrictions are imposed.

Table 4.9: NER models evaluation results on different training data. The best performance is shown in bold and the second best is underlined.

Model	Strict			Type			Partial			Exact		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
BERT-CRF	**0.788**	0.274	0.407	0.795	0.276	0.410	0.805	0.28	0.415	<u>0.801</u>	0.279	0.414
FT BERT-CRF	0.770	<u>0.770</u>	<u>0.770</u>	**0.832**	**0.832**	**0.832**	<u>0.817</u>	<u>0.817</u>	<u>0.817</u>	0.784	0.784	<u>0.784</u>
LitBERT-CRF	<u>0.783</u>	**0.774**	**0.779**	**0.832**	0.823	<u>0.827</u>	**0.829**	**0.819**	**0.824**	**0.802**	**0.793**	**0.797**
LitBERTimbau	0.740	0.767	0.753	<u>0.796</u>	<u>0.825</u>	0.810	0.786	0.815	0.800	0.758	<u>0.786</u>	0.771

P: Precision — *R:* Recall — *F1:* F1-Score

For the most part, all fine-tuned models show a clear improvement over the general-domain baseline (BERT-CRF) across all evaluation scenarios. The baseline model exhibits limited ability to identify named entities in literary texts, as evidenced by the low number of Correct (C) entities and the high number of Missing (M) errors. In contrast, models exposed to literary data during fine-tuning—either through cross-domain transfer learning or domain-adaptive pre-training—consistently achieve higher recall and substantially higher F1-scores compared to the baseline.

Beyond error counts, Table 4.9 reports precision, recall, and F1-score for each evaluation scenario. Here, precision reflects the percentage of correctly identified entities, recall measures the proportion of entities captured from the gold standard, and F1-Score combines these metrics into a single one, which is especially useful for evaluating performance in unbalanced tasks. Such metrics are calculated under two conditions: requiring an exact match (for **Strict** and **Exact** scenarios) or allowing partial matches (for **Type** and **Partial** scenarios).

Cross-Domain Transfer Learning. The fine-tuned BERT-CRF model (FT BERT-CRF) achieves consistently strong results across all evaluation scenarios. In the **Strict** scenario, it reaches an F1-score of 77%, with balanced precision and recall, indicating effective adaptation from general-domain pre-training to the literary domain. In the **Exact** scenario, which evaluates boundary accuracy independently of entity type, the model attains an F1-score of 78%, suggesting robust boundary detection.

Under more relaxed evaluation conditions (**Type** and **Partial**), FT BERT-CRF achieves F1-scores above 81%, indicating that the model is particularly effective at identifying relevant entity spans even when exact boundaries are difficult to determine. These results suggest that cross-domain transfer learning provides a strong baseline for literary NER, especially when annotated literary data is available for fine-tuning.

Domain-Adaptive Pre-Training. The domain-adaptive models, LitBERT-CRF and LitBERTimbau, achieve performance levels comparable to those of the cross-domain transfer learning approach across all evaluation scenarios in Table 4.9. LitBERT-CRF attains the highest or near-highest F1-scores in most scenarios, particularly under stricter evaluation conditions. However, the absolute differences between LitBERT-CRF and FT BERT-CRF are small, often below one percentage point, indicating that both strategies

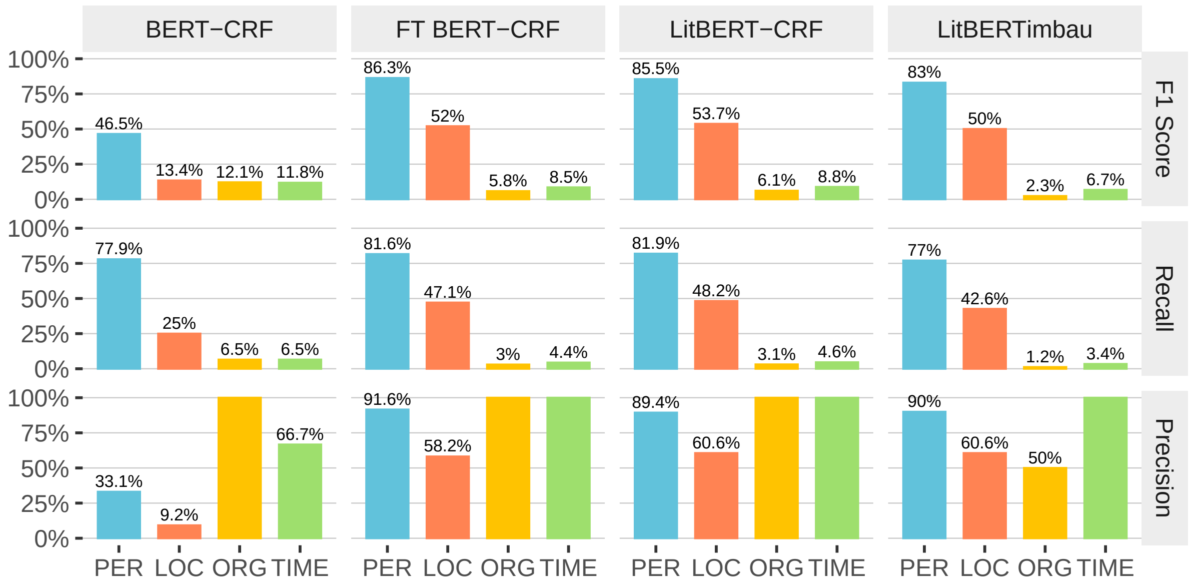


Figure 4.3: Evaluation metrics for each model, considering the **Type** scenario.

Table 4.10: Types of errors made by the evaluated models, by entity type.

Model	Missing				Spurious	Strictly Incorrect				Type Incorrect				Partial			
	P	*L*	*O*	*T*	-	*P*	*L*	*O*	*T*	*P*	*L*	*O*	*T*	*P*	*L*	*O*	*T*
BERT-CRF	213	98	0	1	29	2	1	0	0	1	1	0	0	1	0	0	0
FT BERT-CRF	26	39	0	0	65	15	20	0	0	1	7	0	0	14	15	0	0
LitBERT-CRF	33	34	0	0	62	12	19	0	0	1	9	0	0	11	12	0	0
LitBERTimbau	30	38	0	0	84	13	18	2	0	2	5	1	0	11	13	1	0

P: PER — *L:* LOC — *O:* ORG — *T:* TIME

yield similar effectiveness for the character identification task.

LitBERTimbau also presents competitive results, albeit with slightly lower F1-scores across scenarios. These differences are modest and suggest that architectural choices, such as the inclusion of a CRF layer, as well as differences in pre-training trajectories, may influence performance more than domain adaptation alone. Overall, the results indicate that domain-adaptive pre-training and cross-domain transfer learning offer comparable benefits for literary NER, with neither strategy clearly dominating across all evaluation conditions.

Entity-Level Analysis. Figure 4.3 nd Table 4.10 present a breakdown of performance by entity type under the **Type** scenario. All fine-tuned models perform well on the `PERSON` category, which is expected given its frequency and relative salience in literary texts [179]. While LitBERT-CRF achieves slightly higher precision for `PERSON` entities, FT BERT-CRF attains a marginally higher recall, resulting in very similar F1-scores. These small differences likely reflect trade-offs between precision and recall rather than systematic superiority of one model over another.

For other entity categories, particularly `LOC`, domain-adaptive models tend to reduce the number of missing entities compared to the baseline, suggesting improved sensitivity to location mentions in literary texts. At the same time, these models exhibit a higher number of incorrect or partially matched LOC entities, indicating ongoing challenges in boundary delimitation and fine-grained classification. For ORG and TIME, all models show lower precision, reflecting the linguistic variability and contextual ambiguity with which such entities are expressed in literary narratives.

Conceptualization of Named Entities. Finally, part of the observed error patterns can be attributed to differences in how named entities are conceptualized in *PPORTAL_ner* and in traditional NER corpora such as HAREM. By including both proper names and descriptive nominal expressions, *PPORTAL_ner* captures a broader range of referential phenomena typical of literary texts. While this choice increases linguistic coverage, it also raises the difficulty of strict boundary and type matching, particularly under the **Strict** evaluation scenario. As a result, some increase in Incorrect (I) errors is expected and reflects the intrinsic complexity of literary entity annotation rather than deficiencies of specific models.

4.6 Overall Considerations

In this chapter, we investigated pre-training strategies for enhancing Named Entity Recognition in Portuguese-language literary texts. We introduced two domain-adaptive models, LitBERT-CRF and LitBERTimbau, and compared them against a cross-domain transfer learning setup and a general-domain baseline (BERT-CRF). This comparative analysis provides insights into the benefits and limitations of different adaptation strategies when applied to literary NER.

Overall, all models exposed to literary data during fine-tuning outperform the general-domain baseline, confirming the importance of domain adaptation for character identification. Across evaluation scenarios, LitBERT-CRF and FT BERT-CRF achieve comparable performance levels, with small differences that reflect trade-offs between precision and recall rather than clear dominance of one strategy. These results indicate that both domain-adaptive pre-training and cross-domain transfer learning are viable approaches for literary NER, depending on modeling constraints and available resources.

A closer inspection of the PERSON entity category, which corresponds directly to literary characters, shows that fine-tuned models achieve consistently high performance, with F1-scores above 85%. This confirms that exposure to literary annotation schemes such as *PPORTAL_ner* substantially improves the recognition of character mentions, including both proper names and descriptive nominal expressions.

Nevertheless, recurring challenges remain with complex or multi-token character mentions. Typical error cases include incomplete boundary detection in expressions such as *"the youngest son of the captain"*, partial recognition of honorific constructions (e.g., *"Dona Maria das Dores"*), or fragmentation of long descriptive references that mix proper names and relational nouns. Representative examples of these errors are documented and analyzed in Appendix D, where boundary mismatches and partial entity detections are shown to be a dominant source of residual errors.

Error analysis further shows that domain-adapted models substantially reduce the number of missing entities compared to the baseline, indicating improved sensitivity to literary entity mentions. At the same time, spurious predictions remain more frequent in fine-tuned models, suggesting a tendency to over-generate entities in ambiguous narrative contexts. This behavior reflects the inherent difficulty of distinguishing between character mentions, metaphorical references, and generic descriptions in literary prose, rather than simple overfitting.

This chapter concludes the first stage of the proposed computational framework for gender-bias analysis, addressing **RG2** and the foundational task of identifying literary characters. Within *PORTALIA*, this stage is critical, as errors in character identification propagate to subsequent modules and directly affect the reliability of downstream gender analyses. Despite the demonstrated improvements of domain-adaptive models, limitations remain, particularly regarding boundary detection and variability in historical writing styles, which are inherent to the predominantly public-domain literary corpus available for model adaptation.

Chapter 5

Inferring Gender of Named Entities

After identifying literary characters, the next stage in *PORTALIA* is to infer the character's gender. Gender inference, also referred to as *gender tagging* or *gender classification*, refers to inferring the gender of named entities within a text [160]. This task is essential for studying gender bias, as it allows researchers to analyze gender representation. In the literary domain, where nuanced patterns in character representation may reflect or reinforce societal norms [12, 13, 45, 109], accurate gender inference is particularly valuable. By identifying the gender of entities, researchers can uncover representation trends, analyze character roles, and evaluate potential biases embedded in the text.

As discussed in Section 2.3.2.2, many studies of literary gender bias rely on manual annotation to classify the gender of characters [98, 112, 114, 115, 122, 134, 136]. While such approaches allow for nuanced interpretive judgments, they are inherently limited in scale and reproducibility. These limitations have motivated the adoption of automated methods, typically based on machine learning or natural language processing, to support large-scale literary analysis [157–159].

However, most automated gender inference approaches depend on large, annotated datasets to achieve robust performance [161]. For Portuguese, and especially for Portuguese-language literary texts, such resources are scarce or nonexistent. This scarcity is compounded by the historical nature of most publicly available literary corpora, which reflect linguistic conventions, naming practices, and gender norms that differ substantially from contemporary usage. These constraints limit the direct applicability of supervised or data-hungry approaches, including large language models trained predominantly on modern, general-domain data.

In such a context, this chapter adopts a linguistically grounded heuristic approach to gender inference, designed to be interpretable, reproducible, and compatible with low-resource literary settings. Rather than positioning this approach as a replacement for recent LLMs, we treat it as a complementary and methodologically transparent solution that aligns with the goals of historical literary analysis. Importantly, the proposed method does not rely on coreference resolution or global character linking; instead, it infers gender locally, based on evidence available in character mentions and their immediate textual context. This design choice reflects a deliberate trade-off between contextual depth and

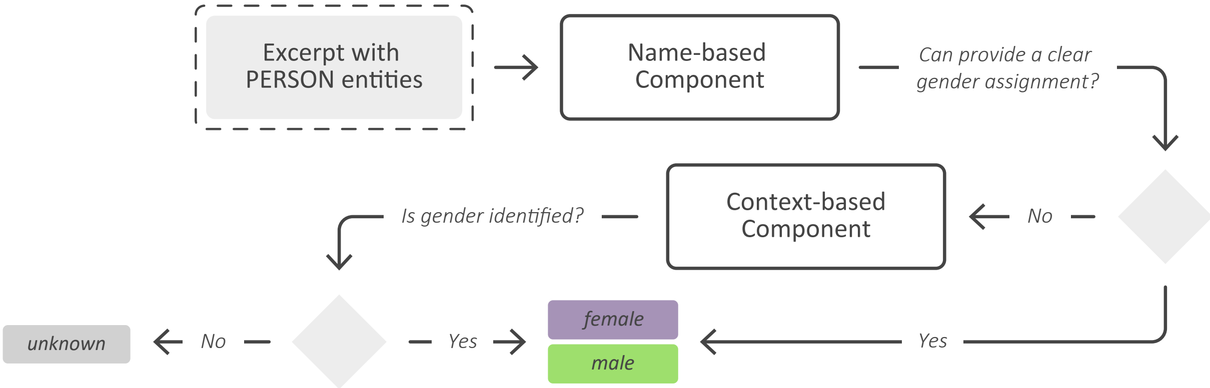


Figure 5.1: Overview of the proposed gender inference heuristic, combining name-based and context-based components to assign gender labels to entities.

analytical robustness, prioritizing reproducibility and linguistic interpretability.

The proposed heuristic integrates two complementary components: (i) a name-based gender prediction module grounded in anthroponymic data, and (ii) a context-based inference module that exploits morphosyntactic gender markers in the surrounding text, such as agreement patterns in determiners, adjectives, and participles. By combining onomastic cues with grammatical evidence, the method captures gender marking even in cases of ambiguous, rare, or absent proper names. Next, Section 5.1 details the proposed heuristic approach, including the rules and linguistic cues employed by each component. Section 5.2 presents a quantitative and qualitative evaluation of the method. Finally, Section 5.3 summarizes the main findings and discusses their implications for subsequent stages of the *PORTALIA* framework.

This chapter addresses the third research goal of this dissertation, **RG3** (Section 1.2), which aims to develop and evaluate an automated approach for gender inference in Portuguese-language literary texts. The heuristic proposed here operationalizes this goal by combining name-based rules with contextual linguistic cues, providing a scalable and interpretable method that is well-suited to the scarcity of gender-annotated corpora in Portuguese. The work conducted under **RG3** has resulted in two publications that apply the proposed gender inference heuristic to quantitatively analyze descriptions of male and female body parts in literary works in Portuguese [37, 38].

5.1 Gender Inference Heuristic

Figure 5.1 illustrates the proposed heuristic for gender inference in Portuguese-language literary texts. For each entity identified as `PERSON`, the method assigns one of three labels: *female*, *male*, or *unknown*.[1] The heuristic is composed of two sequential and

[1]Given the historical nature of the analyzed corpus, gender marking in the texts overwhelmingly follows a binary grammatical system. Accordingly, this work focuses on binary gender inference, while recognizing that contemporary literary contexts may require broader representations.

complementary components. First, a *name-based component* (Section 5.1.1) attempts to infer gender using onomastic information derived from census data. When this step yields no confident assignment, due to name ambiguity, rarity, or the absence of a proper name, the heuristic falls back to a *context-based component* (Section 5.1.2), which infers gender from morphosyntactic cues in the local textual context.

This design reflects a deliberate trade-off between coverage and interpretability. Name-based inference efficiently handles frequent and unambiguous cases, while context-based inference captures gender marking expressed through grammatical agreement, which is particularly salient in Portuguese. As with all downstream components of *PORTALIA*, the accuracy of gender inference depends on the quality of the preceding *Character Identification* step; errors in character identification can propagate to the gender inference step, potentially affecting downstream analyses of literary texts.

5.1.1 Name-based Component

The name-based component relies on the *genderBR* package,[2] which associates first names with gender probabilities based on data from the Brazilian Institute of Geography and Statistics (IBGE) 2010 Census.[3,4] This census covers approximately 190.8 million individuals and includes over 130,000 distinct first names, providing broad empirical coverage of Portuguese-language naming practices. Although the *genderBR* database is derived from Brazilian census data, many Portuguese personal names are shared across Brazilian and European Portuguese traditions. For historical literary texts, the name-based component thus provides a useful approximation, while acknowledging that naming conventions may vary across time and regions.

For a given entity, the heuristic attempts to match its head name against the *genderBR* database. Each name is associated with the proportion of individuals registered as female or male. Following the default configuration of the package, names are classified as *female* if the proportion of female usage exceeds 0.9, as *male* if it is below or equal to 0.1, and as *unknown* otherwise. This conservative threshold prioritizes precision over coverage, minimizing incorrect gender assignments for ambiguous or weakly attested names.

Table 5.1 shows examples of gender assignments by the name-based component. Names such as "Capitu" and "Iracema" are clearly associated with the female gender due to their high probability of female usage, while "Bentinho" is similarly classified as male. Other entities, such as "Diadorim" and "Cris", either lack clear gender data or fall below the confidence threshold, leading to an *unknown* classification. Moreover, expressions

[2] *genderBR*: https://github.com/meirelesff/genderBR

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

[4] When newer census statistics become available, the genderBR tables can be updated simply by recomputing the female–male proportions using the IBGE's publicly released name-frequency data.

Table 5.1: Examples of gender assignments using the name-based component.

Entity	Prob.	Gender	Entity	Prob.	Gender	Entity	Prob.	Gender
Capitu	1.00	female	Diadorim	–	unknown	minha mãe	–	unknown
Bentinho	0.00	male	Tiê	0.55	unknown	o garoto	–	unknown
Iracema	0.99	female	Cris	0.74	unknown	a cachorra Baleia	–	unknown

Prob.: *Probability that the name is associated with female individuals in the IBGE census*
Note: *Capitu and Bentinho are protagonists of Machado de Assis' Dom Casmurro; Iracema is the titular character of José de Alencar's novel; Diadorim is a central character in Guimarães Rosa's Grande Sertão: Veredas; Tiê and Cris appear as given names with contemporary usage, included here to illustrate ambiguous or insufficiently represented names.*

Table 5.2: Examples of gender assignments using contextual linguistic cues.

Entity	Context	Gendered Cues	Gender
Diadorim	Diadorim é um bravo guerreiro.	"bravo", "guerreiro"	male
minha mãe	Ela é minha mãe.	"ela", "minha", "mãe"	female
Tiê	Tiê caiu do morro.	–	unknown

female — male

such as *"minha mãe"* (my mother), *"o garoto"* (the boy), and *"a cachorra Baleia"* (the dog Baleia) illustrate cases where the name-based component is not applicable.

We position the name-based component as the first step in our gender inference heuristic to quickly and efficiently handle straightforward cases where census data can provide confident gender predictions. By assigning a gender label directly from a database of names, we reduce the need for more complex and computationally intensive analysis for well-known and clearly gendered names. However, this approach has limitations when dealing with gender-neutral or rare names, which are less represented in census data and may require context-sensitive interpretation.

5.1.2 Context-based Component

When the name-based component yields an *unknown* label, the context-based component infers gender from grammatical agreement patterns in the surrounding text. It identifies gender markers in nearby words, including pronouns, articles, adjectives, and verbs, which in Portuguese usually agree with the gender of the referenced entity. For instance, *"o amigo"* refers to a male friend, while *"a amiga"* refers to a female friend.

To extract such cues, we employ dependency parsing using the *spaCy* library, identifying tokens syntactically linked to the entity, including determiners, adjectival modifiers, predicate complements, and core pronouns. Gender-bearing words associated with the entity are then aggregated: if the majority of cues exhibit masculine morphology, the entity is classified as *male*; if they show feminine morphology, it is classified as *female*. When no reliable cues are present, the label remains *unknown*.

When nouns lack explicit gender markers, we rely on the surrounding context and linguistic dependencies to infer the gender. If most gender-bearing words associated with a named entity are marked with masculine gender morphology, the entity is classified as male. Conversely, if most gender-bearing words have feminine gender morphology, the entity is classified as female. Table 5.2 provides examples of how the context-based component assigns gender using linguistic cues.

For example, in the phrase *"Diadorim é um bravo guerreiro"* (Diadorim is a brave warrior), the masculine adjective *"bravo"* and noun *"guerreiro"* indicate a male gender assignment. Similarly, in *"minha mãe"* (my mother), the pronoun *"ela"* (she), the possessive adjective *"minha"*, and the noun *"mãe"* confirm a female gender assignment. On the other hand, in the case of *"Tiê caiu do morro"* (Tiê fell from the hill), where there are no clear gendered words, the entity is classified as unknown.

By combining these two complementary components, the heuristic provides a lightweight yet effective method for gender inference in Portuguese-language literary texts. Its main limitations arise from cases with insufficient contextual cues, entities described by gender-invariable nouns (e.g., *"estudante"*, *"presidente"*), or stylistic variations that obscure grammatical agreement, which are common challenges in literary language.

5.2 Experimental Evaluation

To evaluate the proposed gender inference heuristic, we performed a quantitative assessment using a gold-standard dataset of 1,000 `PERSON` entities randomly sampled from the *PPORTAL_ner* corpus. Each entity is assigned one of three labels: *female*, *male*, or *unknown*, the latter corresponding to cases in which gender could not be reliably inferred from the available textual evidence.

Methodological positioning. This evaluation focuses on a rule-based heuristic rather than on supervised classifiers or large language models due to two reasons. First, at the time this research started, high-quality instruction-tuned language models for Portuguese (particularly for historical and literary domains) were either unavailable or insufficiently reliable for fine-grained entity-level gender inference. Second, supervised approaches would require large-scale, consistently annotated datasets of entity-level gender labels, which do not currently exist for Portuguese-language literary texts. In this context, the proposed heuristic offers a transparent, linguistically grounded, and reproducible alternative that is compatible with limited labeled data and historical corpora.

Annotation reliability. To build the gold annotations, two LLMs (Gemini 2.5 Flash and Gemini 2.5 Pro) and one human annotator independently labeled all entities. Inter-annotator agreement is measured using Cohen's κ for each pair and Fleiss' κ for overall consistency. The pairwise κ values ranged from 0.63 to 0.67, with raw agreement rates around 80%, and an overall Fleiss' κ of 0.65, indicating substantial agreement among

Table 5.3: Quantitative results for each heuristic component. Best results are in bold.

Gender	Frequency (%)
Female	138 (13.8%)
Male	613 (61.3%)
Unknown	249 (24.9%)

Component	Accuracy	Coverage	Macro-F1	Weighted-F1
Name-based	0.583 ±0.029	34.3%	0.611	0.598
Context-based	0.652 ±0.029	84.0%	0.542	0.631
Heuristic	**0.762** ±0.025	**93.0%**	**0.641**	**0.713**

annotators. The final gold labels are derived using a majority-vote strategy, with human annotation prevailing in cases of complete disagreement.

The resulting label distribution comprises 61.3% *male*, 24.9% *unknown*, and 13.8% *female* entities. This imbalance is not an artifact of sampling but reflects a well-documented structural characteristic of canonical Portuguese-language literature, in which male characters tend to be both more numerous and more explicitly marked [23, 24, 32, 95, 97, 102–104, 106, 107, 116]. As a result, random samples from literary corpora in Portuguese naturally contain a higher proportion of male-referenced entities.

5.2.1 Evaluation Setup

Since the heuristic is deterministic, evaluation consists of applying it once to the annotated dataset and comparing predicted labels against the gold standard. Three configurations are evaluated separately to assess their individual contribution: (i) *Name-based*, which relies exclusively on the *genderBR* database; (ii) *Context-based*, which infers gender from grammatical agreement patterns; and (iii) *Heuristic (combined)*, which integrates both components through a priority-based rule system.

For each configuration, we report accuracy, precision, recall, and F1-score, along with coverage (i.e., the proportion of entities receiving a non-*unknown* label). Confidence intervals for accuracy are estimated via 1,000 bootstrap resamples. Given the class imbalance, both macro-averaged and weighted F1-scores are reported. Results are presented both for the full three-way classification task and for the restricted binary setting that excludes the *unknown* label, the latter serving as a diagnostic analysis rather than a standalone performance claim.

5.2.2 Results and Analysis

Table 5.3 summarizes the quantitative performance of each component of the proposed heuristic, alongside the overall gender distribution in the annotated dataset. The *Heuristic (combined)* component outperforms the isolated name- and context-based modules across all metrics, confirming the effectiveness of integrating lexical and grammatical cues. Figure 5.2 further details the per-class performance, illustrating the variation in precision, recall, and F1-score across gender labels and components.

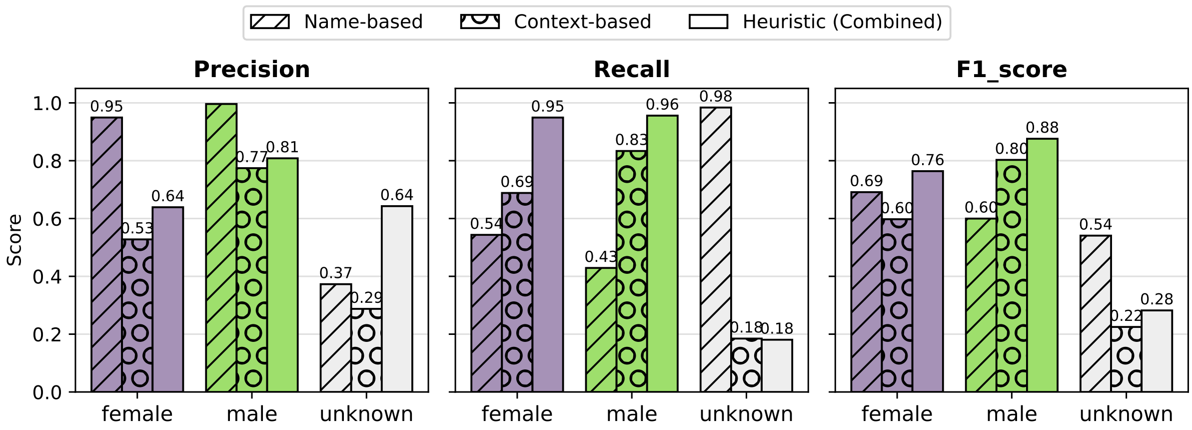


Figure 5.2: Per-class precision, recall, and F1-score across gender labels (*female*, *male*, *unknown*) for each heuristic component. The pattern '/' corresponds to the Name-based component, the pattern 'o' corresponds to the Context-based component, and bars without a pattern represent the Heuristic (combined) component.

Name-based component. The name-based component achieved an accuracy of 58.3% with a coverage of 34.3%, meaning that approximately one third of the entities could be confidently labeled based on name information alone. Precision for both *female* (0.949) and *male* (0.996) labels was extremely high, confirming that when a name is confidently matched in the census database, gender assignment is highly reliable. However, recall was substantially lower (0.543 for *female*, 0.429 for *male*), reflecting limited coverage rather than misclassification. This conservative behavior is a direct consequence of the 90% confidence threshold adopted from *genderBR*. Lowering this threshold (e.g., to 80%) would likely increase coverage at the cost of introducing more false gender assignments, particularly for ambiguous or historically uncommon names. Given the downstream use of these labels in bias analysis, we prioritize precision over recall to avoid propagating systematic misclassifications.

Context-based component. The context-based component, which infers gender from syntactic and morphological cues, yielded higher accuracy (65.2%) and substantially greater coverage (84.0%). This improvement indicates that gender agreement in Portuguese grammar provides robust information even when the entity is not a proper name. The *male* class exhibited strong performance (F1=0.803), while *female* predictions were more balanced but slightly weaker (F1=0.597). The *unknown* class showed low recall (0.185), suggesting that this component tends to make a gendered decision when minimal cues are available, occasionally leading to overgeneralization.

Heuristic (combined). Integrating both components yields an overall accuracy of 76.2% with 93.0% coverage, substantially improving upon either component alone. This indicates the complementarity between lexical and contextual signals. Precision and recall were high for both gendered classes (*female*: Precision=0.639, Recall=0.949, F1=0.764;

male: Precision=0.808, Recall=0.956, F1=0.876), indicating consistent and balanced performance. The *unknown* category remained the most challenging (F1=0.282), as expected for entities lacking explicit gender markers or with conflicting contextual evidence.

Excluding the *unknown* label. To better understand the model's behavior in unambiguous cases, we re-evaluated the heuristic considering only *male* and *female* entities. Accuracy increased from 76.2% to 98.8%, and the macro F1-score reached 0.98. This confirms that most residual errors involve entities with inherently uncertain gender references rather than systematic misclassification between *male* and *female* categories. Thus, the heuristic performs robustly when gender is textually or semantically grounded.

Qualitative error analysis. A total of 417 qualitative errors were identified in the name-based component, primarily involving entities with gender-neutral or rare names, or referring to noun phrases (e.g., *"a professora"*, *"meu pai"*). In the context-based component, 348 errors were observed, most arising when gendered cues referred to secondary entities or when stylistic inversion obscured syntactic agreement. The integrated heuristic reduced the number of qualitative errors to 238, indicating a substantial improvement in interpretability and stability. This reduction suggests that the rule-based integration strategy effectively prioritized reliable lexical evidence while selectively leveraging contextual cues when name-based information was unavailable or ambiguous.

A detailed inspection of the name-based component revealed that most errors did not stem from incorrect gender attribution but from lexical and semantic mismatches. The dominant sources of misclassification include: (i) entities that do not correspond to individuals, such as collective or abstract nouns (e.g., *"os libertos"*, *"os cavaleiros"*, *"multidão"*); (ii) occupational or kinship terms lacking proper names (*"o médico"*, *"a professora"*, *"meu pai"*); and (iii) proper names absent from the census database, particularly foreign, archaic, or literary ones (*Ari*, *Cruelcia*, *Medina*, *Rostopchin*). Overall, these patterns indicate that the main limitation of the name-based component lies in coverage and entity-type recognition rather than in the accuracy of gender assignment itself: when a valid name match is found, the gender prediction is almost invariably correct.

Regarding the context-based component, different inconsistencies were observed, particularly when contextual cues override the surface semantics of the entity. Instances such as *"a pobre criança"* or *"as vítimas"* illustrate misclassifications in which the surrounding narrative implies a feminine context, while the entities themselves remain gender-neutral in isolation. In other cases, the context shifted the gender attribution toward the opposite class, likely due to co-occurring feminine markers or references to female characters in proximity. These discrepancies suggest that the component over-relies on local lexical cues rather than on syntactic or discourse-level dependencies, leading to overgeneralizations in sentences with mixed gender indicators.

Finally, across both components, additional errors were traced to irregularities in the NER model. Some entities were truncated or tokenized incorrectly, often splitting multiword expressions (e.g., *"D. Maria"* → *"D."*, *"Maria"*), while others were wrongly labeled as `PERSON` despite referring to places, objects, or abstract concepts (e.g., *"a torre"*, *"sua casa"*). Such limitations reduce the accuracy of the gender inference stage because the heuristic relies on NER outputs as its input; consequently, misclassification propagates downstream and can bias subsequent analyses of representational patterns. In other words, mistakes in character identification will affect which excerpts are analyzed, which syntactic cues are aggregated per entity, and the bias indicators computed in Chapter 6.

Discussion. Overall, the results indicate that the proposed heuristic achieves a robust balance between accuracy, coverage, and interpretability. While it cannot resolve all cases of referential ambiguity inherent to literary texts, its conservative design minimizes systematic misclassification and provides transparent signals for downstream gender bias analysis. These properties make it well-suited as an intermediate step within the broader *PORTALIA* framework, where explainability and linguistic grounding are essential.

5.3 Overall Considerations

In this chapter, we proposed and evaluated a heuristic approach for inferring the gender of named entities in Portuguese-language literary texts. The method combines two complementary components: a name-based classifier grounded in lexical statistics derived from census data and a context-based module that exploits syntactic and morphological gender agreement in Portuguese. Overall, the experimental results indicate that the proposed heuristic offers a balanced and interpretable solution to gender inference.

The name-based component achieves extremely high precision whenever a reliable match is found in the *genderBR* database, confirming the robustness of census-based lexical statistics for unambiguous names. However, its limited coverage highlights a well-known limitation of name-based approaches in literary corpora, where characters are frequently referred to through archaic, foreign, metaphorical, or descriptive expressions rather than conventional first names. In contrast, the context-based component substantially expands coverage by leveraging grammatical agreement patterns, enabling gender inference even for non-proper or implicitly referenced entities. At the same time, its reliance on local contextual cues makes it vulnerable to stylistic inversions and mixed-gender constructions, which are common in narrative prose.

By integrating both components through a rule-based decision mechanism, the combined heuristic improves both accuracy and coverage, achieving 76.2% accuracy with 93.0% coverage on the annotated dataset. These results confirm the complementarity between lexical and grammatical signals, while also illustrating the trade-offs inherent to rule-based inference in literary texts. When excluding inherently ambiguous cases labeled as *unknown*, accuracy reached 98.8%, indicating that most residual errors stem from referential uncertainty rather than systematic misclassification. These results suggest that the heuristic is well-suited for large-scale gender inference in Portuguese-language literary corpora, where explicit annotations are scarce.

The qualitative analysis provided further insights into the strengths and limitations of the system. Most residual errors were attributed to irregularities in the NER output, entities with collective or abstract references, or syntactic structures in which gender cues referred to secondary elements in the sentence. Nonetheless, the integration of both modules significantly reduced interpretative errors and enhanced the consistency of the inferred labels. The heuristic not only provides reliable gender annotations for literary characters but also forms the foundation for subsequent analyses of gender representation and bias measurement explored in the next chapter.

Overall, this chapter concludes the second stage of the *PORTALIA* framework and our **RG3**, devoted to assigning reliable gender labels to literary characters. The heuristic developed produces gender annotations that directly inform the analyses of representation, characterization, and bias conducted in Chapter 6. However, because gender is a fundamental variable in bias measurement, inaccuracies at this stage necessarily propagate to downstream modules. For example, misgendered entities can distort frequency distributions, skew co-occurrence patterns, and affect the attribution of character traits or narrative roles. At the same time, the transparency and interpretability of the heuristic allow the propagation of errors to be clearly diagnosed and quantified.

Chapter 6

Measuring Gender Bias in Language and Methods

The final step of the *PORTALIA* framework focuses on the measurement and analysis of gender bias in Portuguese-language literary texts. Building upon the previous stages of *Character Identification* and *Gender Inference*, this chapter operationalizes the final component of the framework, *Gender Bias Measuring*, and directly addresses the fourth research goal of this dissertation, **RG4** (Section 1.2). At this stage, the framework moves from entity- and character-level annotation to the systematic quantification of representational asymmetries associated with gender.

More concretely, this chapter investigates how gendered patterns emerge at two interrelated levels. First, it explores *bias in language*, that is, how male and female characters are linguistically described and associated with actions, attributes, and roles within literary narratives. Second, it analyzes *bias in methods*, assessing how computational models trained on or applied to literary texts encode and reproduce these gendered patterns. Together, these analyses aim to make explicit both the textual manifestations of gender bias and how computational methods may reflect, preserve, or amplify them.

Understanding gender bias at both the linguistic and method levels is essential for two reasons. From a literary perspective, it enables large-scale analyses of representation that complement traditional close reading, revealing systematic asymmetries that may not be readily observable in individual works. From a computational perspective, it provides insight into the behavior of NLP models applied to literary data, highlighting how biases embedded in training corpora can surface in downstream representations and generative outputs. Within *PORTALIA*, the gender labels inferred in the previous chapter serve as the anchor for these analyses, enabling comparisons between male- and female-associated excerpts across multiple dimensions.

As previously discussed in Section 3.4, we approach the problem from two complementary perspectives. The first one, *bias in language*, investigates how male and female characters are portrayed within the corpus, exploring descriptive patterns, contextual associations, and representation disparities. The second perspective, *bias in methods*, evaluates how computational models—both embedding-based and generative—encode and

Table 6.1: Literary corpora used to construct the combined literary corpus.

Corpus	#Works	Coverage	Geographical Origin	Refs.
PPORTAL	1,638	13th–21st	Brazil and Portugal	[99, 166, 167]
OBras	400	19th–20th	Brazil	[180]
ELTeC-por	100	1840–1920	Portugal	[181]
Colonia	100	1500–1936	Portugal and Brazil	[182]

reproduce gendered associations present in the corpus. The embedding layer captures lexical associations learned from the corpus, while the LLM layer reveals how such associations manifest in generative behavior.

Next, Section 6.1 outlines the literary corpus analyzed in both linguistic and computational experiments. The analyses rely on a unified corpus assembled from four different corpora to increase linguistic diversity, reduce sampling biases, and ensure sufficient robustness for both embedding-based and LLM-based experiments. Section 6.2 presents our analysis of gender bias in language, detailing the metrics and methods employed. Section 6.3 explores gender bias in computational methods. Finally, Section 6.4 discusses overall considerations and summarizes insights.

6.1 Literary Corpus

This section describes the preparation and characterization of the corpus used to measure and analyze gender bias in Portuguese-language literary texts. First, we outline the data sources and explain the motivation behind selecting these collections (Section 6.1.1). We then detail the steps taken to prepare the corpus, including merging, standardization, and temporal categorization (Section 6.1.2). Next, we present the characteristics of the final corpus (Section 6.1.3). Finally, we discuss limitations related to representativeness, historical bias, and metadata availability (Section 6.1.4).

6.1.1 Data Sources

The primary objective of this chapter is to evaluate the large-scale application of the proposed framework, focusing on measuring and analyzing gender bias in a corpus that is as diverse and representative as possible. To achieve this, we selected four distinct corpora of public-domain Portuguese-language literary works, containing both Brazilian and European texts, which are widely recognized in literary studies and linguistic research. These corpora are described next and in Table 6.1.

PPORTAL. A cross-collection dataset containing metadata for public domain Portuguese-language works [99, 166]. In addition to metadata, *PPORTAL* provides download links for over 9,500 public domain literary works. From these available links, we extract the raw texts of 1,638 works from the *Domínio Público* collection, specifically focusing on works categorized as "Literatura" (Literature) and "Literatura Infantil" (Children's Literature).

OBras (Obras Brasileiras). A collection of Brazilian public domain literary works [180]. OBras is a dynamic corpus to which new works are continuously added; the version used here, last updated in July/2024, contains 400 works. The corpus is part of the Literateca collection[1] and is integrated into the AC/DC project [183], a corpus search service developed and maintained by Linguateca. All materials in OBras are publicly accessible and can be explored through the AC/DC interface for advanced searches or downloaded directly from the project page.[2]

ELTeC-por. A collection of Portuguese novels curated for the European Literary Text Collection (ELTeC), part of the COST Action *Distant Reading for European Literary History* [181].[3] This corpus consists of 100 novels written in European Portuguese, primarily focusing on works from the 19th and 20th centuries.

Colonia. A collection of Portuguese manuscripts published between 1500 and 1936, representing both Portugal and Brazil [182]. The collection includes 48 European Portuguese texts and 52 Brazilian Portuguese texts.

6.1.2 Corpus Preparation

After collecting the raw texts from each corpus, we applied several preprocessing steps to standardize the data and prepare it for analysis. First, duplicate entries across corpora are removed by identifying redundancies based on work titles and author names. Next, text formats are standardized to ensure consistency, which is particularly important because the OBras, ELTeC-por, and Colonia corpora include syntactic and/or semantic annotations in XML formats, whereas the *PPORTAL* corpus provides plain text files.

Additional metadata for all selected works is collected using the Goodreads API.[4] Specifically, we focused on original publication years, which are essential for temporal categorization, and literary genres, as indicated by Goodreads popular shelves. Metadata fields such as author names, publication dates, and genres are then harmonized to ensure uniform formatting and alignment across the entire corpus.

Finally, to investigate long-term tendencies in gender representation, the merged corpus is divided into broad temporal subsets based on publication century: up to the 18th century, 19th century, 20th century, and 21st century. This categorization enables a coarse-grained analysis of historical variation while avoiding overly fine temporal slicing that would result in sparse data, particularly for earlier periods.

[1] Literateca: https://www.linguateca.pt/Literateca/
[2] Projeto OBras: https://www.linguateca.pt/OBRAS/OBRAS.html
[3] Distant Reading for European Literary History: https://www.distant-reading.net/
[4] Goodreads API: https://www.goodreads.com/api

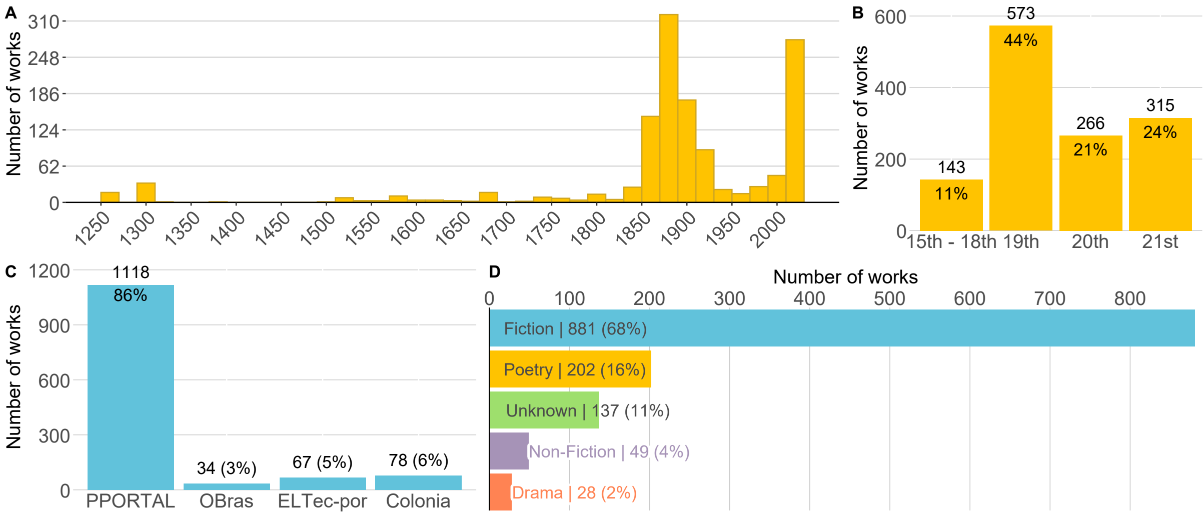


Figure 6.1: Combined corpus distribution by (**A**) publication year, (**B**) century, (**C**) source, and (**D**) literary genre.

6.1.3 Corpus Characterization

The final combined corpus initially contained 1,836 literary works drawn from the four selected corpora. After deduplication, format standardization, and metadata harmonization, the corpus was reduced to 1,297 unique works (71%). Figure 6.1 summarizes the distribution of the corpus by publication year, century, source, and genre. As shown in Figure 6.1(**B**), the corpus is heavily concentrated between the mid-19th and early 20th centuries. This concentration reflects the structure of contemporary digital literary repositories, which disproportionately preserve, curate, and disseminate texts from this period, particularly canonical works that are more frequently republished and digitized.

In contrast, works published before 1800 are much less represented,[5] totaling 143 works (11%), as shown in Figure 6.1(**B**). Rather than reflecting a substantially lower volume of literary production in earlier centuries, this underrepresentation is more plausibly explained by modern processes of digitization, canon formation, and editorial interest. Older texts are less frequently digitized, less accessible in standardized formats, and often attract less attention from contemporary publishers and digital humanities initiatives, especially when compared to 19th-century canonical literature.

Figure 6.1(**C**) shows the corpus distribution by source. The majority of works (86%) originate from the *PPORTAL* collection, followed by Colonia (6%), ELTeC-por (5%), and OBras (3%). The predominance of *PPORTAL* reflects both the deduplication process, which prioritized this collection, and the fact that it provides raw texts critical for direct analysis, along with extensive metadata that facilitates categorization

[5] To mitigate sparsity effects and enable meaningful comparison, all works published up to the 18th century are grouped into a single temporal category. This aggregation does not aim to capture fine-grained historical change within early periods, but rather to allow early literary texts to be included in longitudinal analyses without overinterpreting patterns derived from small samples.

and contextualization. The smaller corpora (Colonia, ELTeC-por, and OBras) contribute valuable diversity, offering specialized foci such as colonial literature or annotated texts, complementing the larger *PPORTAL* dataset.

Regarding literary genres, Figure 6.1(**D**), fiction works dominate, accounting for approximately 68% of the corpus. Poetry and drama follow, contributing 16% and 2%, respectively. This distribution reflects the historical prominence of narrative and poetic forms in Portuguese-language literature, particularly during the 19th and early 20th centuries, when movements such as Romanticism and Realism shaped literary production.

Non-fiction works, including historical accounts and sermons, make up roughly 4% of the corpus, primarily sourced from Colonia. These texts provide insights into historical and rhetorical perspectives on gender, particularly within contexts of colonization and religious discourse. Notable examples include *A Carta* by Pero Vaz de Caminha and *Sermões* by Padre Antonio Vieira, which illustrate early constructions of gender roles in Portuguese exploration and colonial settings.

6.1.4 Corpus Limitations

Despite its size and diversity, the combined corpus presents limitations that must be acknowledged. One key limitation is that the corpus has exclusively public-domain works. Since copyright restrictions primarily allow free access to older texts, the resulting dataset is necessarily skewed toward works published up to the early 20th century. This constraint directly affects several dimensions of diversity in the corpus.

First, there is an imbalance in temporal coverage: the corpus is dominated by works from the 19th and early 20th centuries, while texts from the 18th century and contemporary literature are significantly underrepresented. Consequently, the patterns identified here are more reflective of historical literary conventions than of current linguistic or cultural dynamics, potentially ignoring any major shift in recent decades.

Second, although the corpus includes works from both Brazilian and European Portuguese traditions, it presents a stronger concentration of Brazilian Portuguese, particularly due to the predominance of the *PPORTAL* and OBras collections. This distribution may introduce regional biases in vocabulary, syntactic structures, naming conventions, and literary style, thus limiting the generalizability of findings when considering Portuguese-language literature as a whole.

Third, while reliable metadata on author gender is incomplete, a manual inspection of the corpus indicates a marked predominance of male authors: approximately 93% of the works are authored by men, with only around 7% authored by women. This distribution aligns with historiographical studies showing that women's participation in Portuguese and Brazilian literary production was historically constrained by social, educational, and editorial barriers [98]. Furthermore, archival research demonstrates that

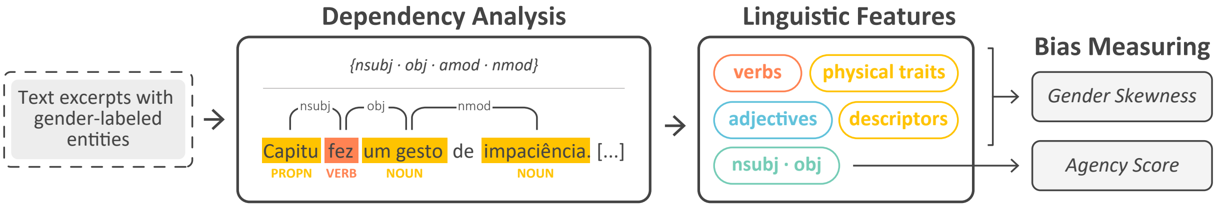


Figure 6.2: Overview of the gender bias in language workflow.

many women's works were erased, misattributed to male authors, or published under masculine pseudonyms [97], further reducing their visibility in public domain collections.

Finally, the exclusive reliance on public domain texts makes it difficult to incorporate contemporary works or texts authored by women, Black authors, Indigenous authors, LGBTQIA+ authors, and other historically marginalized groups whose voices are central to understanding present-day transformations in literary representation and discourse. This limitation restricts the scope of conclusions regarding the evolution of gender bias and the diversity of literary production in Portuguese.

Despite these constraints, the corpus provides a robust and historically grounded basis for large-scale analysis, enabling the systematic examination of gender representation across genres, regions, and long temporal spans. Still, awareness of these limitations is crucial for contextualizing the findings presented in the following sections and for guiding future efforts toward more inclusive corpus construction.

6.2 Gender Bias in Language

This section focuses on the linguistic dimension of gender bias in Portuguese-language literary texts. Following the character identification and gender inference stages, we now analyze how language encodes asymmetries in the portrayal of male and female characters. Specifically, we explore the lexical and syntactic patterns that reveal how characters of different genders are described, represented, and situated within narratives.

The underlying assumption is that linguistic choices, such as adjectives used to describe characters, nouns denoting social or emotional roles, and verbs indicating actions or states, reflect and perpetuate culturally embedded gender norms. By systematically quantifying these patterns, we can identify how gender bias manifests not only in the frequency of male versus female representation but also in the ways each gender is linguistically constructed.

To guide the analysis, the linguistic component of the framework unfolds in three sequential steps, illustrated in Figure 6.2: (i) *Dependency Analysis*, which identifies syntactic relations linking `PERSON` entities to surrounding lexical items (verbs, nouns, adjectives) (Section 6.2.1); (ii) *Lexical Features*, which collects and categorizes the words associated with characters through dependency relations considered in the analysis (Sec-

Table 6.2: Universal dependency relations used in the analysis.

Depend.	Meaning	Definition
nsubj	Nominal Subject	Nominal which is the syntactic subject and the proto-agent of a clause
obj	Direct Object	The noun phrase that denotes the entity acted upon or which undergoes a change of state or motion (the proto-patient)
amod	Adjectival Modifier	Any adjectival phrase that serves to modify the meaning of the noun
nmod	Noun Modifier	Nominal modifiers that depend on another noun. It corresponds to an attribute, or genitive complement

Source: https://universaldependencies.org/pt/dep/

tion 6.2.2); (iii) *Bias Measuring*, which computes metrics (*Agency Score* and *Gender Skewness*) quantifying asymmetries in linguistic usage. (Section 6.2.3). Finally, Section 6.2.4 presents and discusses the results, highlighting the main trends and asymmetries observed across literary periods and genres.

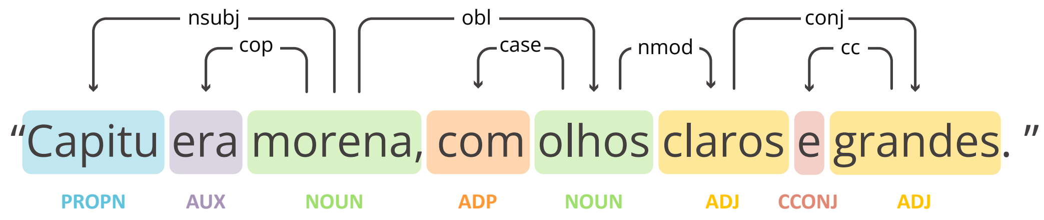
6.2.1 Dependency Analysis

In the dependency analysis step, we explore the syntactic relationships between identified `PERSON` entities (i.e., literary characters) and surrounding linguistic contexts, focusing on verbs, nouns, and adjectives. To do so, we use the `pt_core_news_lg` model from spaCy, whose dependency parser is trained on the *Universal Dependencies (UD) Portuguese Bosque Treebank*.[6] This treebank, developed collaboratively by the Floresta Sintá(c)tica project, provides consistent syntactic annotations and is a reliable resource for identifying grammatical relations in Portuguese-language texts.

The dependency parser identifies the syntactic roles and relations between words, such as subjects, objects, and modifiers, that structure a sentence. These relations are important for understanding how gender is linguistically represented in association with literary characters. Specifically, we focus on the following dependency relations and linguistic components, defined in Table 6.2.

Nominal Subjects (`nsubj`). The nominal subject is the noun or noun phrase that occupies the syntactic subject position of a clause and controls verbal agreement. While this position often coincides with semantic agency in active clauses, it may also encode non-agentive or patient roles in passive or stative constructions. For this reason, we interpret `nsubj` as a syntactic indicator of prominence rather than a direct marker of semantic responsibility. This relation typically connects a noun or noun phrase (e.g., "the dog", "a woman", "the tall boy") to a verb that agrees with it in number (singular or plural) and person (first, second, or third). **Example:** "O cachorro late alto" (The dog barks loudly), where "O cachorro" (the dog) is the nominal subject, and the verb "late" (barks) agrees with it in the third person singular.

[6] https://universaldependencies.org/treebanks/pt_bosque/index.html



Universal dependencies: `nsubj`: nominal subject, `cop`: copula, `obl`: oblique nominal, `case`: case marking, `nmod`: nominal modifier, `conj`: conjunct, `cc`: coordinating conjunction

POS tags: `PROPN`: proper noun, `AUX`: auxiliary verb, `NOUN`: noun, `ADP`: adposition, `ADJ`: adjective, `CCONJ`: coordinating conjunction

Figure 6.3: Dependency analysis applied to a synthetic sentence—translated to English as "Capitu was brunette, with large and clear eyes."

Direct Object (`obj`). The direct object is the noun or noun phrase that receives or is directly affected by the action of the verb. Analyzing direct objects helps uncover how actions and roles are distributed among characters, particularly in terms of gender representation. Gendered nouns or pronouns in this position can reveal implicit associations or stereotypes related to agency and passivity in narratives. **Example:** "A mulher viu o homem" (The woman saw the man), where "o homem" (the man) is the direct object of the verb "viu" (saw). The gender of the object "homem" (man) may reflect how male and female entities are positioned within the narrative structure.

Adjectival Modifiers (`amod`). Adjectives modify nouns by adding descriptive information about their qualities, states, or attributes. In Portuguese, adjectives typically agree with the nouns they modify in both gender and number, making them key indicators of gendered representation. The distribution and semantics of adjectives can thus reveal how characters are portrayed within traditional gender frameworks. **Example:** "O homem bonito" (The handsome man), where "bonito" (handsome) agrees in gender with the noun "homem" (man), exemplifying an association between beauty and masculinity.

Noun Modifiers (`nmod`). Noun modifiers add semantic detail to a noun, often expressing relationships such as possession, association, or characterization. This syntactic relation frequently involves prepositional phrases or nominal complements that link one noun to another, enriching the contextual meaning of entities. In literary texts, noun modifiers can be particularly revealing of relational or social roles, shedding light on how characters are positioned or described in gendered terms. **Example:** "Os olhos de Capitu" (The eyes of Capitu), where "de Capitu" functions as a modifier of "olhos" (eyes), establishing a relational context that connects the described feature to the character.

Figure 6.3 illustrates the dependency structure of a synthetic example. Each token is represented as a node in a directed graph, with edges indicating grammatical relations. In this sentence, the entity *"Capitu"* is associated with descriptors such as *"morena"* (brunette) and *"olhos claros"* (clear eyes), which together construct her physical portrayal. This dependency graph highlights how syntactic structures can be leveraged to trace lexical features that encode gendered representation.

It is worth noting that while the term *"morena"* (brunette) functions as a descriptive adjective, it was incorrectly labeled as a noun. This misclassification highlights a limitation in the dependency parser's handling of specific context-dependent terms in Portuguese, where words that can shift between noun and adjective roles based on context are sometimes misclassified. Such limitations are especially significant when analyzing gendered language, as misclassifications may obscure the intended descriptive relationships between entities and their attributes.

6.2.2 Lexical Features

Building on the dependency analysis, this section focuses on the extraction and categorization of lexical features that encode gendered patterns in literary texts. Once syntactic dependencies establish the relations between characters and surrounding lexical items, we analyze the specific words and constructions that reveal how gender is linguistically represented. These features (adjectives, verbs, and nouns) serve as key indicators of how agency, attributes, and social roles are distributed across male and female entities.

Adjectives. Adjectives are important linguistic markers of gender representation, as they often convey evaluative and descriptive meanings that can reinforce or challenge traditional stereotypes [10, 32, 45]. For instance, Fast et al. [46] observed that adjectives associated with male characters frequently reflect traits such as "strong", "arrogant", and "dominant", while adjectives linked to female characters include terms like "submissive", "dependent", and "hysterical". In our analysis, we extract adjectives connected to characters either through the `amod` dependency relation or via copulative constructions (e.g., "ser", "estar"). Analyzing the distribution of adjectives in this way provides insights into how gendered traits are lexicalized in literary texts.

Verbs. Verbs encode agency and action, reflecting how characters engage with events and exert influence over their narrative environments [10, 15, 111, 118]. In our analysis, we extract verbs associated with characters through the `nsubj` and `obj` dependency relations, which indicate when a character functions as the subject or object of an action. This approach allows us to assess whether male and female entities are more frequently depicted as agents, actively performing actions, or as patients, being acted upon. Prior research indicates that male characters are more often described using action-oriented verbs, whereas female characters tend to be associated with verbs expressing passivity, emotions, or social interactions, thereby reinforcing traditional gender stereotypes [45].

Nouns. Nouns are central to capturing how gendered identities and social roles are linguistically constructed. We focus on nouns that either function as direct descriptors of entities (e.g., "herói", "mãe", "senhora") or appear in modifier relationships (`nmod`). To systematically assess gender bias in these nouns, we employ two complementary lexical resources. First, regarding *physical traits*, we manually compiled a dictionary of body parts, containing 55 core body parts and 104 synonyms (see Table E.1). This lexicon allows us to match nouns referencing specific body parts (e.g., "olhos", "mãos", "cabelo"), enabling the analysis of potential biases in how physicality is emphasized and whether such emphasis differs across genders. Second, for *descriptors*, we use the lexicon proposed by Freitas and Santos [32], categorizing each word directly associated with a `PERSON` entity via dependency relations into four categories: "social", "emotional", "physical" (appearance), and "character" [32]. "Social" descriptors encompass professions, occupations, or social status; "emotional" descriptors capture feelings or tendencies; "physical" descriptors describe appearance; and "character" includes personality traits, including cognitive attributes such as intelligence. These categories are not mutually exclusive, so a single word may belong to multiple categories depending on context. Words that do not align with any predefined category are labeled as "other".

Overall, these lexical features, interconnected through dependency relations, form the basis for identifying and quantifying patterns of gender bias in language. The following section describes the computational strategies used to measure these patterns and assess their variation within the literary corpus.

6.2.3 Bias Measuring

Previous studies have proposed a variety of metrics to quantify gender bias in language, including stereotype scores [66], composite scores [19], and genderedness measures [138]. While informative, these approaches are often limited to specific lexical features (e.g., adjectives or occupational terms) or rely on co-occurrence counts, which can be sparse and unstable in literary corpora.

Hoyle et al. [20] introduced *Pointwise Mutual Information (PMI)* to measure the strength of co-occurrence between gendered terms and descriptive attributes. Similarly, Barré and Dupont [45] proposed an *association score* that quantifies the link between words and gender based on the proportion of masculine and feminine characters associated with each word. Despite their usefulness, these metrics present limitations: PMI is highly sensitive to low-frequency words, often inflating scores for rare attributes, and its unbounded scale complicates interpretation. Association scores, while more interpretable, do not explicitly account for overall attribute frequency and may be biased in corpora with uneven gender distributions.

To overcome these limitations, we introduce *Gender Skewness*, a metric designed to capture the imbalance in the occurrence of any linguistic attribute between male and female characters. Unlike PMI or association scores, Gender Skewness normalizes the difference between conditional probabilities by their sum, yielding a bounded score between -1 and 1. Positive values indicate a skew toward female characters, negative values toward male characters, and values near zero indicate neutral usage. This normalization allows for fair comparison across attributes with different base frequencies and mitigates the influence of rare occurrences, making the metric particularly suitable for large and heterogeneous literary corpora.

In this study, we combine *Gender Skewness* with the *Agency Score* [45] to capture complementary dimensions of bias: *syntactic agency* and *qualitative framing*. Both metrics are computed for the lexical features extracted in Section 6.2.2, including verbs, adjectives, descriptors, and physical traits.

Agency Score. Proposed by Barré and Dupont [45], the *Agency Score* evaluates the extent to which characters of a given gender tend to occupy syntactic positions associated with agency or patienthood in a narrative. Specifically, it quantifies the propensity of a character to appear as a *syntactic* subject (`nsubj`) or as a direct object (`obj`) of a sentence. Formally, the *Agency Score* is defined as:

$$A(G) = \frac{P(\texttt{nsubj}|G) - P(\texttt{obj}|G)}{P(\texttt{nsubj}|G) + P(\texttt{obj}|G)}, \tag{6.1}$$

where $P(\texttt{nsubj}|G)$ represents the probability that entities of gender G occur in the syntactic subject position, and $P(\texttt{obj}|G)$ represents the probability that entities of gender G occur as direct objects. A higher agency score indicates a greater agency presence for the gender, while a lower score indicates more passivity [45].

Importantly, this metric captures *syntactic* rather than semantic agency. In particular, the dependency relation `nsubj` does not always correspond to the semantic agent of an action, especially in passive constructions (e.g., `nsubj:pass`), copular clauses, or stative predicates. As a result, the *Agency Score* should be interpreted as a measure of syntactic prominence and grammatical positioning within clauses, rather than a direct proxy for intentional or causal agency. Nonetheless, prior work has shown that, when aggregated over large corpora, such syntactic distributions provide a meaningful and robust signal of narrative agency and gendered asymmetries in character representation [15, 45].

Gender Skewness. While the *Agency Score* captures asymmetries in syntactic positioning, it does not account for the distribution of qualitative or descriptive attributes across genders. To address this limitation, we introduce *Gender Skewness*, a metric that quantifies the imbalance in the occurrence of linguistic attributes associated with male and female characters. For a given attribute x (e.g., a verb, adjective, descriptor, or physical trait), *Gender Skewness* is calculated as:

$$S(x) = \frac{P(x|F) - P(x|M)}{P(x|F) + P(x|M)}, \tag{6.2}$$

where $P(x|F)$ denotes the probability of attribute x occurring in excerpts associated with female `PERSON` entities, and $P(x|M)$ denotes the probability of attribute x occurring in excerpts associated with male ones. The values of $S(x)$ are interpreted as follows:

$$S(x) = \begin{cases} \text{skewed toward female entities} & \text{if } S(x) > 0, \\ \text{skewed toward male entities} & \text{if } S(x) < 0, \\ \text{equally associated with both genders} & \text{if } S(x) = 0. \end{cases}$$

By normalizing the difference by the sum of probabilities, $S(x)$ is bounded between -1 and 1, allowing for direct comparison across attributes with different overall frequencies. Analyzing the distribution of *Gender Skewness* across lexical features enables the identification of systematic patterns of gender bias in literary language.

6.2.4 Results and Discussion

This section presents and discusses the results obtained from applying the proposed framework to the dimension of *Gender Bias in Language*. This dimension focuses on the linguistic expressions through which gender bias manifests in literary discourse, including patterns of agency, descriptive asymmetries, and differential use of linguistic attributes across genders. The goal is to explore how language contributes to the construction and perpetuation of gendered representations in Portuguese-language literature.

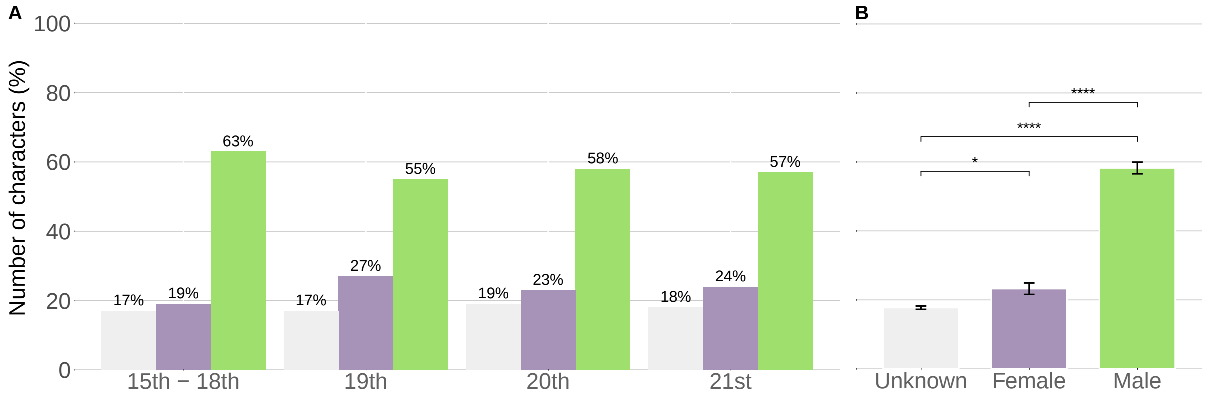
To guide this investigation, we define six research questions (RQs) that address complementary aspects of gender bias in language. The first three questions (**RQ1–RQ3**) examine how gender representation and linguistic patterns evolve across centuries, revealing *temporal trends*. The remaining three questions (**RQ4–RQ6**) focus on *language-based analyses*, exploring how specific lexical, syntactic, and descriptive features differentiate the portrayal of male and female entities.

RQ1 (Temporal Trends): *How does overall gender representation vary across centuries in the literary corpus?*

RQ2 (Temporal Trends): *How are linguistic markers distributed between genders across centuries in the literary corpus?*

RQ3 (Temporal Trends): *How do agency and passivity biases shift across centuries in the literary corpus?*

RQ4 (Language-based Analysis): *Are specific words or categories disproportionately skewed to male or female entities in the literary corpus?*



(T-test: "*" $p <= 0.05$, "**" $p <= 0.01$, "***" $p <= 0.001$, and "****" $p <= 0.0001$)

Figure 6.4: Overall gender representation. (**A**) Distribution of gender representation (`PERSON` entities) across centuries in the corpus, highlighting the proportion of male, female, and unknown individuals. (**B**) Overall distribution of gender representation in the entire corpus, with statistical significance.

RQ5 (Language-based Analysis): *What are the most frequently mentioned physical traits for male and female entities in the literary corpus, and how do these traits reinforce or challenge gender stereotypes?*

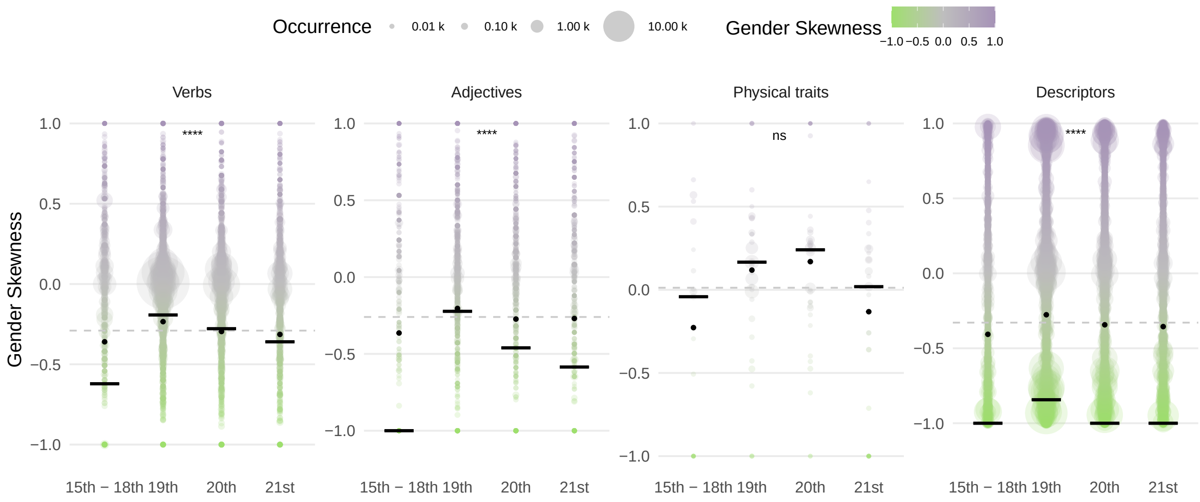
RQ6 (Language-based Analysis): *How do interactions between male and female entities (e.g., subject-object relationships) reflect power dynamics in the literary corpus?*

The results are organized into two complementary parts: (i) *Temporal Trends* (Section 6.2.4.1), which explore the diachronic evolution of gender representation and linguistic attributes, addressing **RQ1–RQ3**; and (ii) *Language-based Analysis* (Section 6.2.4.2), which investigates gendered linguistic markers at a finer level of granularity, addressing **RQ4–RQ6**. Finally, Section 6.2.4.3 provides an integrative discussion that synthesizes these findings, situating them within broader historical and cultural contexts of Portuguese-language literature.

6.2.4.1 Temporal Trends

[RQ1] Overall gender representation. Figure 6.4(**A**) shows the gender distribution of `PERSON` entities across centuries, while Figure 6.4(**B**) presents the overall proportions in the corpus. Male entities consistently dominate, representing more than half of all identified individuals in every century. Female representation remains stable but under-represented, ranging from 19% to 24% across the centuries. There is also a significant share of entities labeled as "unknown" gender, likely attributed to a lack of explicit gender markers or incomplete linguistic information.

The 18th century and earlier display the highest male proportion (around 63%), suggesting that early Portuguese-language literature was even more male-centered. This



(Kruskal-Wallis test: "*" $p <= 0.05$, "**" $p <= 0.01$, "***" $p <= 0.001$, and "****" $p <= 0.0001$)

Figure 6.5: Gender skewness distribution for male and female entities across centuries. Dashed lines represent mean values within the attributes, and the black points and cross-bars represent mean and median values for each century, respectively.

aligns with historical patterns that excluded women from public and literary life, restricting them to domestic or peripheral roles [184]. Although women's social visibility expanded in the 19th and 20th centuries, with increased access to education and authorship, the literary representation of female characters remained largely unchanged. Similar persistence has been observed in English literature, where female visibility has improved but not yet reached parity [15, 24, 135].

In sum, to answer **RQ1**, our findings indicate that the corpus analyzed exhibits a predominantly male-centered narrative structure, reflecting and possibly reinforcing historical gender hierarchies. These findings raise further questions about how female characters are portrayed, whether as active agents or confined to stereotypical roles, within predominantly male-driven narratives.

[RQ2] Temporal gender skewness. Figure 6.5 shows the distribution of gender skewness across centuries, capturing imbalances in the association of linguistic attributes (verbs, adjectives, descriptors, and physical traits) with male and female characters. Positive values indicate a bias toward female entities, negative values toward male entities, and values near zero suggest balance.

Results show a persistent male bias across centuries, most pronounced before the 19th century. Verbs in early periods display stronger associations with male entities, reflecting narrative patterns in which male characters are more frequently positioned in action-oriented or event-driving roles. Adjectives, in contrast, do not encode agency directly but contribute to evaluative and symbolic framing. Their stronger association with male entities in earlier periods suggests that men are more frequently characterized, whereas women are comparatively less described through adjectival qualification in these

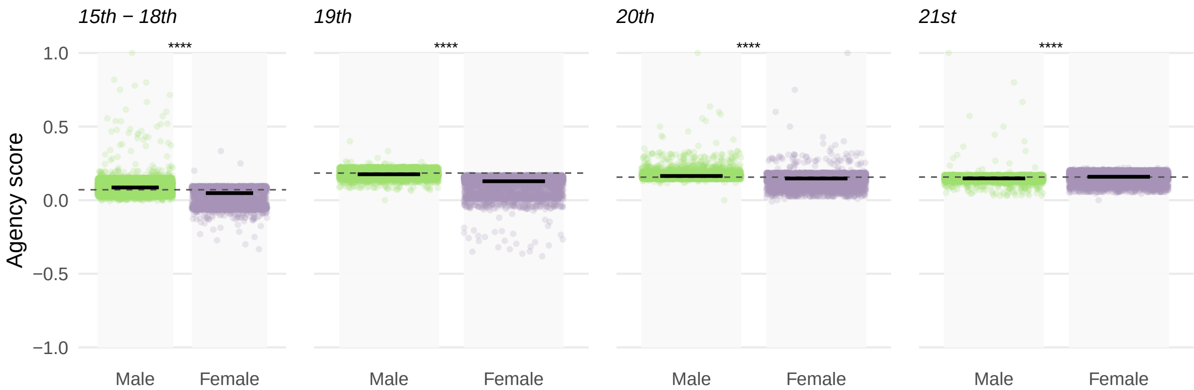


Figure 6.6: Agency score distribution by gender and century. Statistical significance of differences is marked by "****" (T-test, $p < 0.0001$). Dashed lines represent mean values, and black crossbars indicate medians for male and female entities.

periods. From the 19th century onward, gender asymmetry declines, possibly mirroring women's gradual social and literary inclusion. Yet, in the 20th and 21st centuries, asymmetry resurges slightly, suggesting that while overt bias decreased, gendered linguistic patterns remain culturally ingrained.

Regarding physical traits, the average gender skewness is close to zero, indicating a more balanced portrayal across genders. The distributions across centuries do not show any statistically significant differences, suggesting that physical traits have been more equally attributed to both genders over time. Still, in the 19th and 20th centuries, certain descriptors show mild positive skewness, suggesting subtle female predominance in physical characterization, perhaps reflecting increased attention to women's appearance in modern narratives.

Overall, to answer **RQ2**, our findings highlight that although gendered language has evolved, imbalances persist, especially in action-related attributes, which continue to privilege male entities.

[RQ3] Agency and passivity bias. Figure 6.6 shows the distribution of agency scores across centuries. The score measures how often male and female entities appear as agents (subjects) versus patients (objects). Results reveal significant gender disparities across all centuries (T-test, $p < 0.0001$), with mean values consistently close to zero, reflecting a general tendency for entities to be depicted as passive or neutral. However, male entities consistently show positive scores, reflecting their predominance in active, action-performing roles, while female entities display predominantly negative scores.

Overall, the disparity in agency scores is most pronounced in pre-19th-century texts but narrows in later centuries, suggesting gradual, though limited, progress. By the 20th century, female agency increases slightly, reflecting broader social changes, yet the disparity persists. These results mirror findings in French and English corpora, where female passivity remains a recurrent narrative feature [15, 45].

Addressing **RQ3**, we find that the corpus analyzed preserves entrenched gendered portrayals of action and influence. Although female characters gradually gain narrative agency, male dominance continues to shape the dynamics of power within the analyzed works, highlighting the slow and uneven reconfiguration of gendered roles in fiction.

6.2.4.2 Language-based Analysis

This section investigates specific linguistic attributes (verbs, adjectives, physical traits, and descriptor categories) associated with male and female entities in the corpus. By analyzing these attributes, we address **RQ4**, **RQ5**, and **RQ6**, exploring how specific words or descriptor categories contribute to gender bias in our corpus. To do so, we employ two complementary approaches:

1. **Quantitative Analysis:** We use the *Gender Skewness* metric to quantify the imbalance in the frequency of linguistic attributes between male and female entities.
2. **Qualitative Analysis:** We explore specific examples of verbs, adjectives, descriptors, physical traits, and subject-object relationships to interpret how these lexical features reflect or challenge traditional gender roles and power dynamics.

To capture overall tendencies, we average gender skewness values across the entire corpus and apply chi-squared tests to assess whether associations between linguistic attributes and gender are statistically significant.

[RQ4] Skewed descriptors. To address **RQ4**, we analyze the gender skewness distribution of the extracted linguistic attributes, including verbs, adjectives, physical traits, and descriptor categories. As described in Section 6.2.2, a descriptor corresponds to any word directly associated with a `PERSON` entity through dependency relations. Using the lexicon proposed by Freitas and Santos [32], we categorize each descriptor into four non-exclusive types: *social*, *emotional*, *physical* (appearance), and *character*. This categorization serves as a proxy for identifying common stereotypes found in literary works.

It is important to note that extreme gender skewness values for individual words may be driven by low absolute frequencies, particularly in the case of rare or context-specific terms. For this reason, gender skewness should not be interpreted in isolation at the level of individual lemmas. Instead, it serves as an aggregate indicator whose interpretability increases when jointly considered with frequency, category-level distributions, and recurring patterns across multiple attributes. Our analysis, therefore, emphasizes trends at the level of lexical groups and categories rather than isolated words.

Figure 6.7 shows that *appearance*, *emotion*, and *social* descriptors are most skewed toward female entities, portraying women as emotionally expressive, socially defined, and physically characterized. For male entities, the most skewed categories are *character*, *social*, and *emotion*, highlighting portrayals of agency, authority, and individuality. The

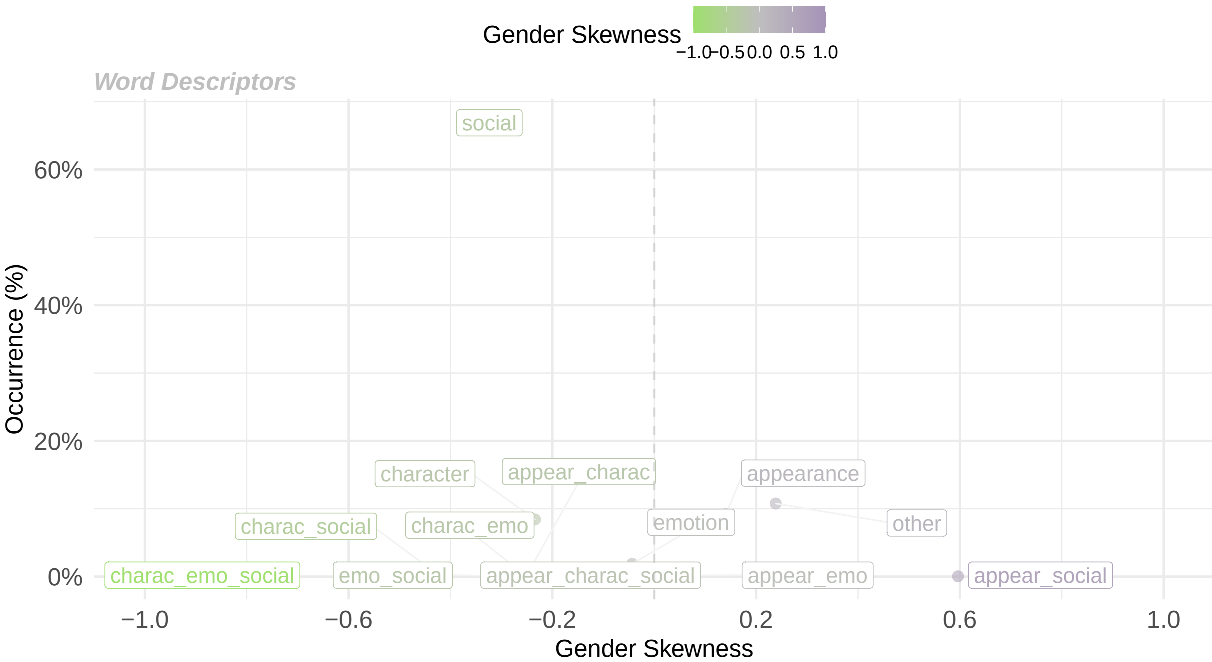


Figure 6.7: Gender skewness of descriptors associated with `PERSON` entities, grouped into four categories: *social*, *emotional*, *appearance*, and *character*. The x-axis represents the gender skewness score, where positive values indicate a stronger association with female entities, and negative values indicate a stronger association with male entities. The y-axis represents the word occurrence as a percentage of the total mentions in the corpus.

most frequent overall category is *social*, suggesting that both genders are often framed through social relations or occupations, though with distinct nuances.

Figure 6.8 presents the most gender-skewed and neutral descriptors. The most frequent descriptors for female entities are closely tied to the *social* and *appearance* categories. Examples include "benfeitora" (benefactress), "empregada" (maid), and "cortesã" (courtesan) in the *social* category, and "velhota" (old woman) and "resplandecente" (radiant) in the *appearance* category. These descriptors reinforce traditional gender roles that emphasize women's social and physical characteristics, often linking their identity to caregiving, servitude, or physical beauty.

In contrast, descriptors frequently associated with male entities predominantly fall within the *social* category, reflecting roles of authority, exploration, or labor. Examples include "agricultor" (farmer), "comandante" (commander), "embaixador" (ambassador), and "explorador" (explorer). These terms align with stereotypical depictions of men as active professionals, leaders, and adventurous agents, reinforcing societal norms prioritizing male contributions to public and occupational spheres [10, 106].

The gender-neutral descriptors, positioned closer to the zero-skewness axis, often describe general qualities or traits that are not overtly gendered. These include words such as "triste" (sad) or "small" (pequeno), which are applied relatively equally to male and female entities. As indicated in Figure 6.5, the average gender skewness of descriptors is more balanced, close to zero, suggesting that there are specific attributes in the corpus

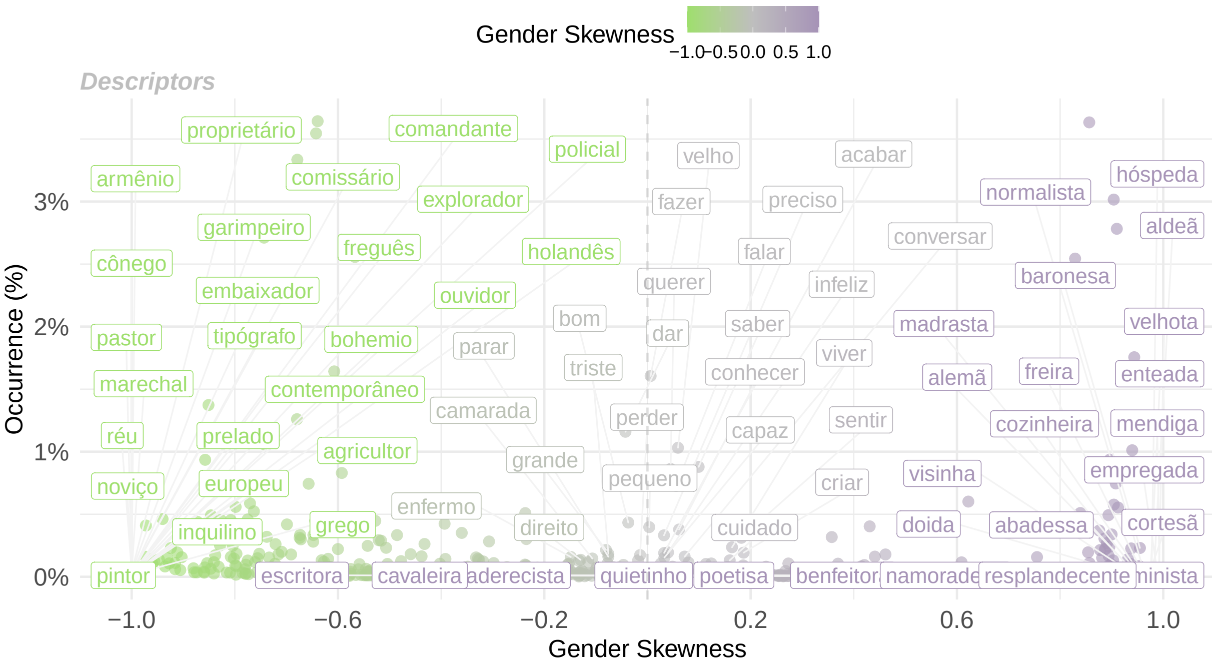


Figure 6.8: Most skewed descriptors in the corpus. Values near zero indicate gender-neutral words. The x-axis represents the gender skewness score, where positive values indicate a stronger association with female entities, and negative values indicate a stronger association with male entities. The y-axis represents the word occurrence as a percentage of the total mentions in the corpus.

that do not strongly favor either gender. This can be interpreted as evidence of some degree of linguistic neutrality in how both male and female characters are portrayed.

Besides descriptors, adjectives and verbs show similar tendencies. Figures 6.9 and 6.10 show the gender distribution of these lexical features. The most skewed adjectives towards female entities are “morena” (brunette), “graciosa” (gracious), “deslumbrante” (dazzling), “amiga” (friend), and “galante” (charming). These adjectives are often linked to traditional feminine qualities associated with physical appearance, charm, and social roles, reinforcing stereotypes of women as delicate, pleasant, and relational beings. There are also adjectives such as “infiel” (unfaithful), “fria” (cold), “má” (bad), “quieta” (quiet), which, while more negatively charged, also tend to reinforce passive or non-agentive roles, suggesting an association with negative aspects of traditional femininity.

In contrast, the most skewed adjectives towards male entities are “assíduo” (diligent), “empenhado” (dedicated), “duvidoso” (doubtful), “bárbaro” (barbaric), and “invejoso” (envious). These adjectives often describe attributes associated with social roles in which men are typically expected to display strength, activity, or certain moral failings. There are also adjectives like “escravo” (slave), “cristão” (Christian), “brutal” (brutal), and “soberbo” (arrogant), which can reflect rigid, often hierarchical gender roles, with men frequently portrayed as dominant, aggressive, or morally complex positions. These adjectives contribute to reinforcing traditional views of masculinity as being tied to authority, power, and an active or controlling presence in society.

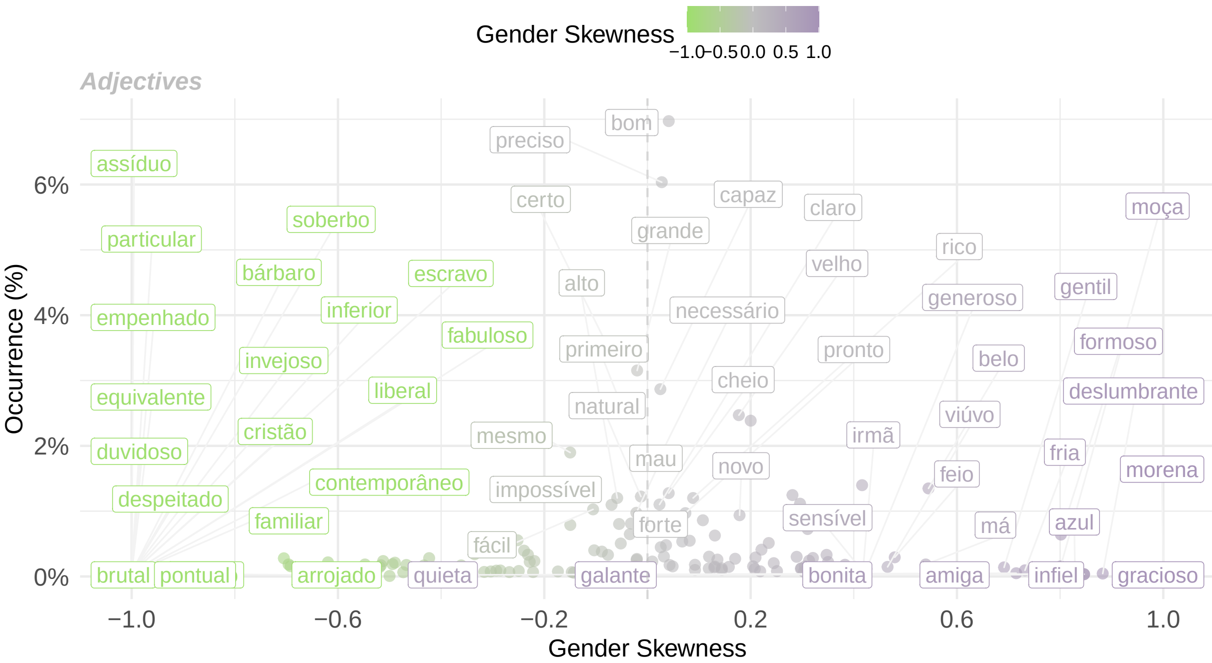


Figure 6.9: Most skewed adjectives in the corpus. Adjectives closer to zero are gender-neutral, reflecting balanced usage for both genders.

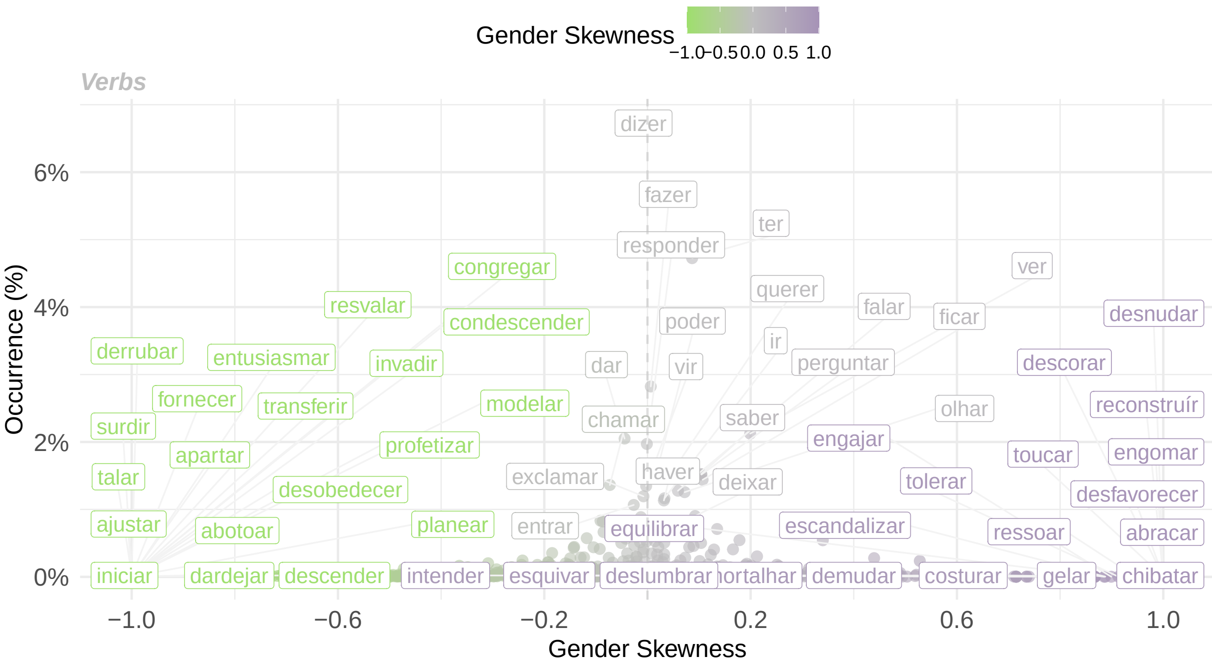


Figure 6.10: Most skewed verbs in the corpus. Verbs closer to zero are gender-neutral, reflecting balanced usage for both genders.

Regarding verbs, the verbs most skewed towards female entities are “desnudar” (to undress), “engomar” (to starch), “abraçar” (to embrace), “reconstruir” (to reconstruct), and “descorar” (to bleach), which are often associated with nurturing, caretaking, or domestic roles. These verbs highlight actions that are traditionally linked to femininity, such as caring for others, attending to domestic tasks, or performing physically intimate actions. In addition, we also have verbs such as “chibatar” (to whip) and “costurar” (to sew), which can reflect a subservient or manual labor role, again reinforcing traditional expectations of women’s work being related to service or crafting.

As for the verbs commonly associated with males, we find “talar” (to carve), “derrubar” (to fell), “surdir” (to surge), “fornecer” (to supply), and “apartar” (to separate), which tend to reflect active, assertive actions that are often linked with power, authority, and physical labor. These verbs describe activities typically attributed to male roles, emphasizing physicality, control, or leadership. In addition, we also have verbs such as “invadir” (to invade), “desobedecer” (to disobey), and “dardejar” (to strike with a dart), which connote aggression, defiance, and combativeness—traits often ascribed to traditional notions of masculinity.

Overall, addressing **RQ4**, our results indicate that specific lexical categories are disproportionately associated with male or female entities in the corpus. Rather than assigning fixed semantic meanings to individual terms, the analysis highlights systematic asymmetries at the level of descriptor types and action patterns. Then, female entities tend to co-occur more frequently with descriptors related to physical appearance, emotions, and social positioning, while male entities are more strongly associated with verbs and descriptors linked to authority, action, and public roles. These tendencies emerge from aggregate distributions across the corpus and should be interpreted as probabilistic patterns rather than deterministic semantic assignments.

Importantly, the framework is designed to minimize subjective interpretation by relying on normalized association metrics and category-level analyses. While individual words may admit multiple plausible readings depending on context, the convergence of results across categories, time periods, and syntactic roles provides a more robust indication of gendered representational bias than isolated lexical examples.

[RQ5] Physical traits associations. Figure 6.11 shows the gender skewness distribution of physical traits, revealing a clear tendency for certain traits to be more frequently associated with one gender over the other. For female entities, the only physical trait that is highly skewed is “cílio” (eyelash), with a gender skewness of 1. This association reflects a stereotypical focus on beauty-related attributes when describing women, as eyelashes are often linked to ideals of physical attractiveness and delicacy.

In contrast, male entities are associated with a broader range of physical traits, including “boca” (mouth), “bochecha” (cheek), “cabeça” (head), “coluna” (spine), and “punho” (wrist). While related to the body, these traits are less explicitly tied to beauty or aesthetic qualities and may instead imply physicality, strength, or functionality. For example, “punho” (wrist) could suggest action or manual labor, aligning with traditional masculine roles, while “coluna” (spine) might metaphorically evoke resilience or strength. The disparity in the types of traits associated with each gender highlights the gendered lens through which physical descriptions are framed in literature. Female traits are often

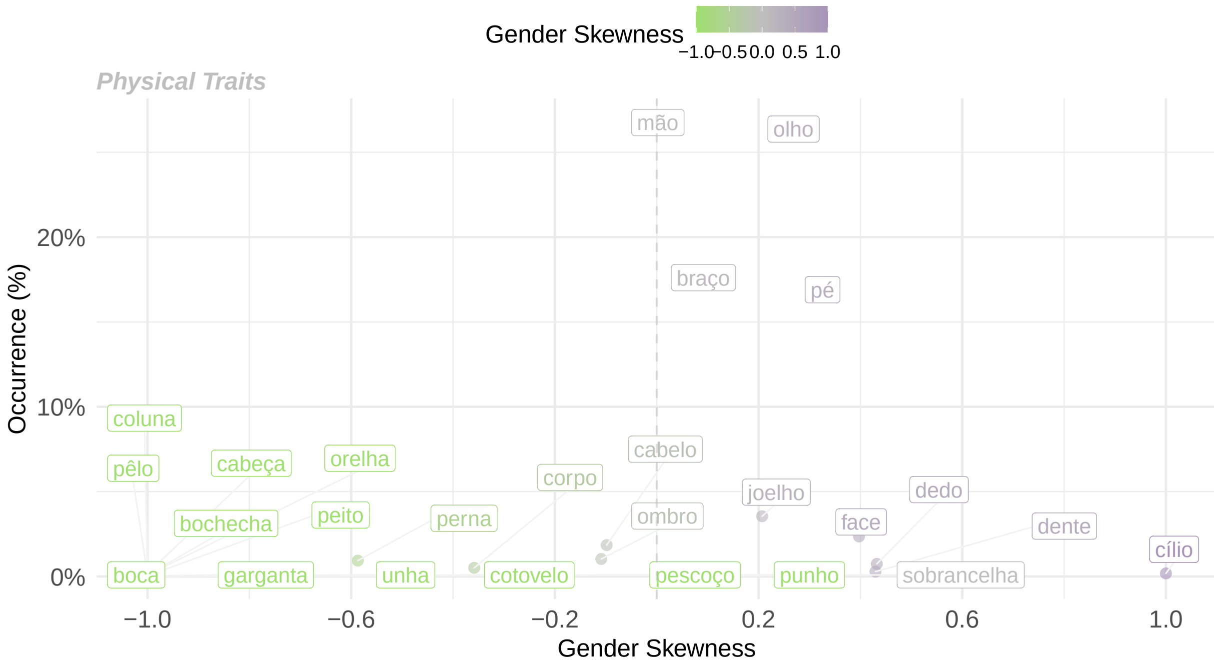


Figure 6.11: Gender skewness of physical traits associated with PERSON entities.

limited to features that reinforce their aesthetic value, whereas male traits are tied to action, utility, or robustness, reinforcing traditional gender roles.

Overall, answering **RQ5**, our findings reveal distinct linguistic patterns in how physical traits are described regarding gender. While some associations align with traditional stereotypes, such as delicacy and beauty for women or strength and action for men, the data also uncovers diverse and context-specific descriptions that vary across traits and genders. These findings highlight how language can reflect and construct gendered perceptions in literature.

[RQ6] Subject-object relationships. To assess how interactions between male and female entities (e.g., subject-object relationships) reflect power dynamics, we investigate the most frequently used verbs in sentences where male and female entities are either the subject or the object. Figures 6.12 and 6.13 show the verbs most frequently associated with male and female entities as subjects or objects, respectively.

For male entities as subjects, verbs such as "comandar" (to command), "realizar" (to accomplish), and "derrotar" (to defeat) appear prominently, indicating a depiction of male characters in positions of authority and control. There are also verbs such as "trepar" (to climb or to fuck) and "abusar" (to abuse), which may highlight problematic aspects of male representation in literature. The verb "trepar", depending on context, can convey both physical prowess and sexual dominance, reinforcing the stereotype of men as aggressive. Similarly, "abusar" reflects themes of power imbalance and exploitation, suggesting darker dimensions of male characters' roles in narratives. These verbs suggest an association of male characters with actions emphasizing physical and social power.

For female entities as subjects, verbs such as "encarregar" (to take charge), "cos-

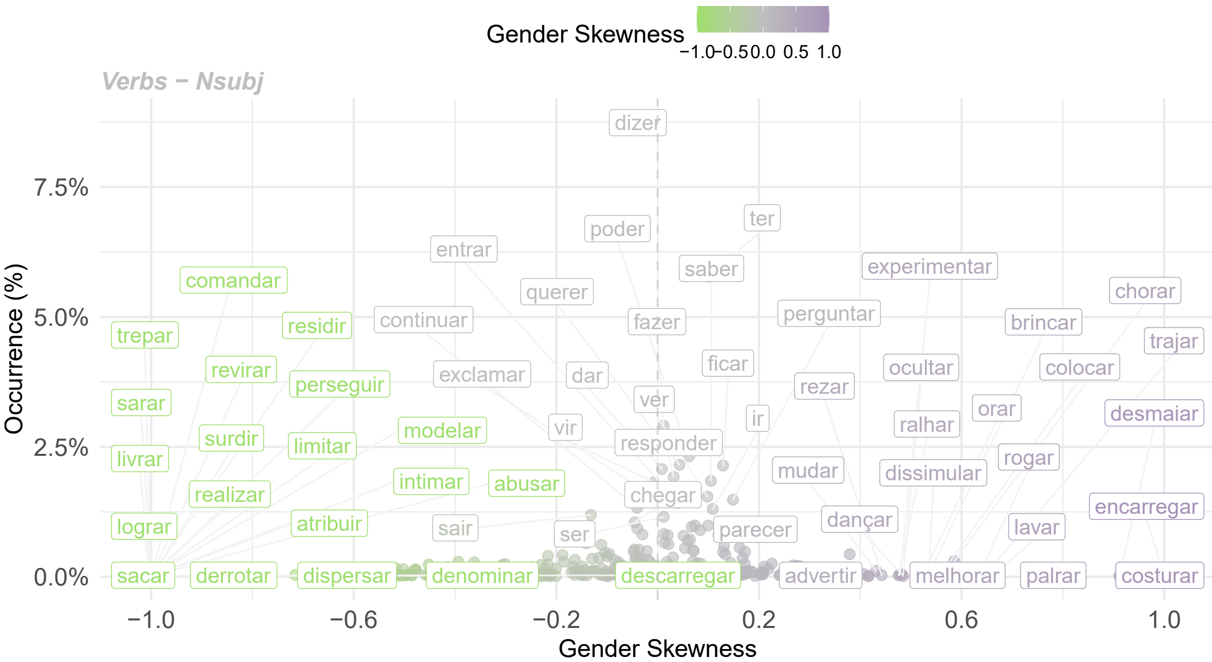


Figure 6.12: Verbs most frequently associated with male and female entities as subjects.

turar" (to sew), "desmaiar" (to faint), "lavar" (to wash), and "chorar" (to cry) are most frequent, highlighting traditional gender roles that emphasize caregiving, domestic labor, and emotional expression. These verbs reflect long-standing stereotypes of women as nurturers, laborers in the private sphere, or emotionally fragile beings. In particular, the presence of verbs like "costurar" and "lavar" reinforces the association of women with domestic tasks, while "chorar" underscores the stereotype of heightened emotionality often attributed to female characters.

Gender-neutral verbs, such as "chegar" (to arrive), "continuar" (to continue), "dar" (to give), "dizer" (to say), "entrar" (to enter), "exclamar" (to exclaim), and "fazer" (to do), are relatively balanced in their distribution across genders. These verbs describe actions not inherently tied to specific gendered roles or stereotypes. Their neutrality highlights areas of narrative where gender does not heavily influence the portrayal of character actions, offering a counterpoint to the more heavily gendered verbs.

When male entities appear as objects, verbs such as "comandar" (to command), "combater" (to combat), "derrubar" (to overthrow), "enfrentar" (to face), "expulsar" (to expel), and "governar" (to govern) are more frequently associated with them. These patterns suggest that while male characters often occupy subject roles of power, they are also depicted in adversarial or competitive situations, where conflict, challenge, and dominance are emphasized. These verbs suggest portraying men not just as leaders but as figures engaged in confrontational or power-assertive relationships.

On the other hand, when female entities are the objects of action, verbs such as "desdenhar" (to scorn), "merecer" (to deserve), "velar" (to watch over), "namorar" (to court), and "casar" (to marry) are more prevalent. These verbs often depict women as

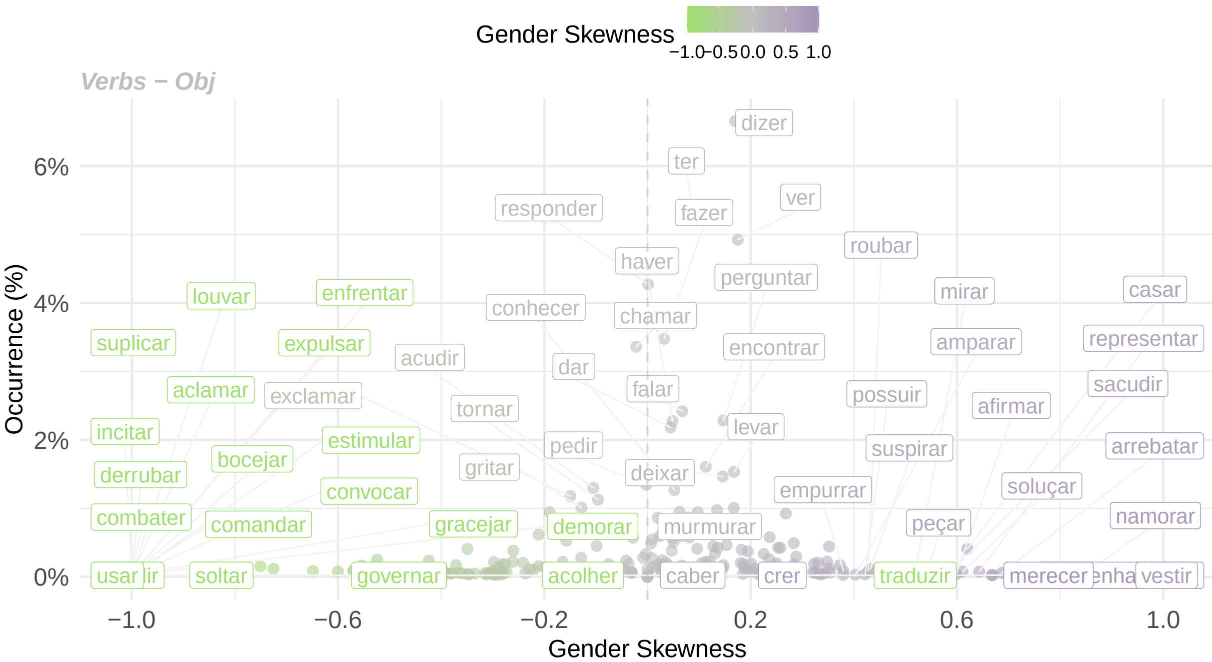


Figure 6.13: Verbs most frequently associated with male and female entities as objects.

recipients of male actions, placing them in roles that are more passive or relational in nature. These associations are commonly framed in terms of romantic, protective, or possessive dynamics, where female characters are frequently seen as objects of desire, affection, or control. Moreover, the language used to depict female objects of action typically emphasizes dependence, subjugation, or the fulfillment of relational roles rather than active participation in decision-making or agency.

Overall, answering our **RQ6**, our findings highlight the distinct ways male and female entities are portrayed in subject-object relationships. Male characters are often shown as active agents in positions of power, leadership, or conflict, while female characters are more commonly depicted in passive, supportive, or relational roles. These patterns reflect broader gender dynamics in literary representations, where men are often associated with dominance and authority, while women are portrayed within narrower roles defined by emotionality, dependence, and relational interaction.

6.2.4.3 Discussion

Overall, the results obtained through the *Gender Bias in Language* dimension of our framework reveal consistent patterns of asymmetry in how male and female entities are represented across the analyzed corpus. Temporal patterns indicate that, although portrayals of women have evolved slightly over time, persistent gendered conventions continue to shape narrative structures. Female entities are still frequently associated with relational or supportive roles, while male entities dominate in positions of agency and authority [15, 24, 135]. This observation raises important questions about how literature

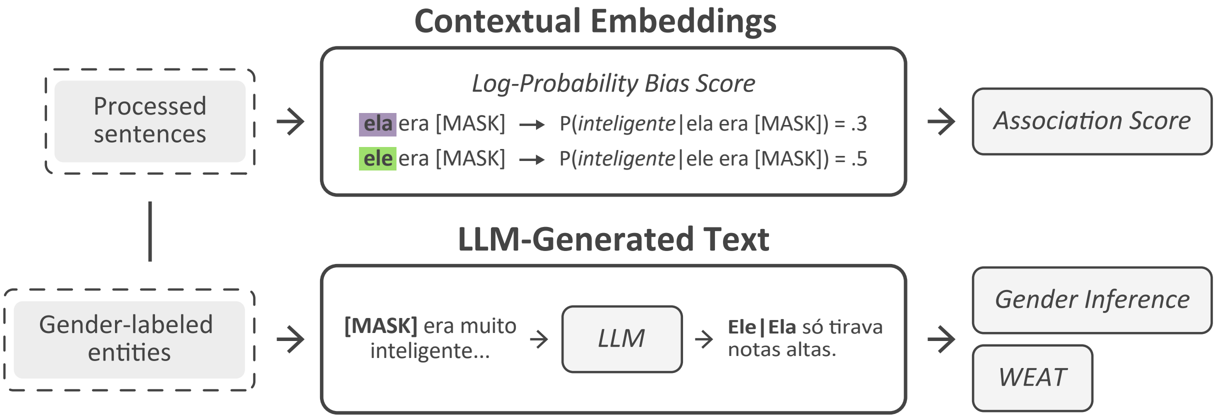


Figure 6.14: Overview of the gender bias in computational methods workflow.

can evolve alongside societal changes in gender equality.

The language-based analyses further expose how these asymmetries are embedded at the linguistic level. By quantifying descriptors, verbs, and adjectives associated with gender, we identified subtle yet systematic differences that traditional literary analyses might overlook. Computational methods thus extend the interpretive scope of literary inquiry, revealing how gender is encoded not only in character roles but in the very language used to construct them. This linguistic embedding of bias reinforces implicit gendered perspectives that shape narrative meaning.

Finally, our findings align with prior research emphasizing women's underrepresentation and stereotyped portrayals in literary discourse [10, 15, 32, 105, 114, 130]. Female figures often remain confined to passive, emotional, or aestheticized depictions, while male figures embody action, intellect, and power. These tendencies suggest that literature, while capable of contesting social norms, often reproduces the patriarchal structures and symbolic hierarchies of its historical and cultural context [98]. At the same time, the observed temporal variations signal that literary language can also serve as a site of gradual transformation, reflecting evolving understandings of gender and agency.

6.3 Gender Bias in Computational Methods

While the previous section focused on how gender is encoded in language, this section explores how computational methods internalize, amplify, or reshape these patterns. Specifically, we analyze how models trained or fine-tuned on our corpus capture gendered associations, both in their latent representations and in their generative behavior. As illustrated in Figure 6.14, our investigation is organized along two complementary levels of analysis:

1. **Bias in Contextual Embeddings:** We explore how associations between gendered terms (e.g., *homem* vs. *mulher*) and descriptive attributes are structured in the embedding spaces of contextualized models (Section 6.3.1).

2. **Bias in Generated Text:** We explore how gender bias emerges in the outputs of LLMs, assessing how narrative continuations reflect stereotypical or unbalanced portrayals of male and female characters (Section 6.3.2).

6.3.1 Bias in Contextual Embeddings

Language models have become central to Natural Language Processing (NLP), powering applications from text generation [185] to literary analysis [120]. However, research has shown that these models are not neutral: they encode and can even amplify biases present in their training data [26, 73]. Such biases often manifest as gender, racial, or social stereotypes, potentially reinforcing historical inequalities and discriminatory language patterns in downstream applications [44].

Early work on bias focused on static word embeddings, such as Word2Vec and GloVe, which were shown to encode gender stereotypes [26, 186]. The advent of transformer-based models, like BERT [187], introduced contextualized embeddings, making biases more dynamic and context-dependent [73, 188]. Methods to quantify bias in these models include log-probability differences in MLM [73]. Despite these advances, most studies focus on English [30], leaving gendered languages like Portuguese underexplored.

Portuguese presents specific challenges due to its gendered grammatical system, which requires explicit morphological agreement (e.g., *ela é bonita* vs. *ele é bonito*). This obligatory gender marking may reinforce stereotypes [148]. Existing research on Portuguese NLP has mostly analyzed general-domain corpora [189, 190], overlooking literary texts, which play a significant role in shaping cultural perceptions over time [15, 125]. In literature, female characters are often described in terms of appearance and emotions, while male characters are linked to agency and intellect [32].

In this section, we investigate gender bias in Portuguese-language literary texts using a masked language modeling approach. We fine-tune BERTimbau [168] on a curated corpus of 592 prose works spanning 1804–1998. To probe gendered associations, we construct a template-based sentence corpus using common linguistic patterns from the considered corpus, with gendered terms as targets and adjectives or verbs as attributes. We then apply the log probability difference metric [73] to compare associations before and after fine-tuning, evaluating how exposure to literary texts amplifies or reshapes gender biases in the model.

6.3.1.1 Corpus

To investigate gender bias in Portuguese-language literary texts, we adapted the literary corpus introduced in Section 6.1. From the initial collection of 840 works, we excluded non-narrative genres such as poetry and plays to ensure analytical consistency. This selection criterion is motivated by two main reasons: poetry often employs non-standard

Table 6.3: Composition of the filtered Portuguese prose corpus (1804–1998) used to investigate gender bias in contextual embeddings.

Corpus	**Coverage**	**#Works**	**#Sentences**	**#Tokens**
OBras [180]	1855–1984	23	54,317	1,005,266
Colonia [182]	1844–1948	35	171,741	2,797,503
ELTeC-por [181]	1844–1973	37	209,614	3,139,395
PPORTAL [166]	1804–1998	497	788,542	10,641,252
Total	**1804–1998**	**592**	**1,224,214**	**17,583,416**

syntax, symbolic language, and unconventional structures that hinder meaningful comparison with prose; and dramatic texts primarily consist of dialogue and typically lack the extended narrative descriptions necessary for analyzing gendered language patterns in character portrayals [32].

The final corpus thus consists of 592 prose works (approximately 70% of the original collection), spanning the period 1804–1998, with 1.2 million sentences and 17.6 million tokens. All the texts are pre-processed using our proposed framework, including text cleaning and sentence segmentation. Table 6.3 summarizes the corpus composition, including temporal coverage, number of works, sentences, and tokens.

6.3.1.2 Fine-tuning

To investigate how literary texts influence gender representations in established Portuguese language models, we fine-tuned BERTimbau [168] with the MLM task on our literary corpus (Section 6.3.1.1). This domain adaptation enables the model to capture literary-specific linguistic patterns (e.g., archaic vocabulary, stylistic variations) while preserving its general Portuguese language understanding capabilities.

Setup. Fine-tuning the BERTimbau Base model[7] 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. Model adaptation is assessed using perplexity (PPL) on a held-out validation set (20% of the corpus). After fine-tuning, the model achieved a PPL of 2.71, compared to 3.16 pre-fine-tuning, representing a 14.2% reduction. This improvement, al-

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

Table 6.4: 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**

though moderate, indicates that the model successfully adapted to the linguistic patterns of Portuguese-language literary texts.

6.3.1.3 Gender Bias Assessment

Building on previous methodologies [73, 188], we develop a framework to quantify gender bias in Portuguese-language literary texts using MLM probing. Our approach extends prior studies by: (*i*) incorporating Portuguese-specific grammatical gender constraints, and (*ii*) distinguishing between descriptive stereotypes (captured via adjectives) and agency stereotypes (captured via verbs). The framework consists of two main components: template design (Section 6.3.1.3.1) and bias measurement (Section 6.3.1.3.2).

6.3.1.3.1 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 [191, 192], 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 [193], 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. 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) using the *adj*, *posemo*, and *negemo* categories of BP-LIWC2015. 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). To ensure linguistic diversity, we include multiple conjugations for specific verbs, resulting in a total of nine verb variations. For the *verb-based* templates, we select 120 verbs from the verb, *cogproc* (cognitive processes), *percept* (perception), *social*, and *work* categories of BP-LIWC2015. These verbs are used to explore gendered subject-verb associations related to cognition, perception, social interactions, and professional roles.

In total, ten sentence templates are designed to cover both *adjective-based* and *verb-based* structures, as samples in Table 6.4. Using such templates and systematically combining gendered noun phrases, adjectives, and verbs, we generate a total of 12,000 unique sentences. The complete set of generated sentences is available [194].

6.3.1.3.2 Bias Measurement To quantify gender bias, we use the Log-Probability Bias Score (LPBS) proposed by Kurita et al. [73], which extends the WEAT [67] to masked language models. The LPBS 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 6.3.1.3.1.

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 [188]:

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: $P_T = P(ela = [MASK]|sent)$

Table 6.5: Average gender association scores for pre-trained and fine-tuned models.

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	$+1.45$		

Values are mean $\pm$ standard deviation; Δ indicates the change from pre-trained to fine-tuned; p_{pre} and p_{pos} denote statistical significance of gender differences (female vs. male) before and after fine-tuning, respectively.

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 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.

6.3.1.4 Experimental Results

This section presents the results of our gender bias experiments, comparing the pre-trained BERTimbau model with its version fine-tuned on our Portuguese-language literary corpus (Section 6.3.1.1). We compute association scores for *adjective-based* and *verb-based* templates to evaluate the impact of fine-tuning on gender bias patterns.

Overall Trends. Table 6.5 shows the average association scores for female and male targets across template types. For *adjective-based* templates, the pre-trained model exhibits negative associations for both female (-0.79 ± 1.70) and male (-0.71 ± 1.52) targets, with no significant gender difference ($p_{pre} = 0.56$). Fine-tuning increases the association score for female targets ($-0.79 \rightarrow 0.15$, $\Delta = +0.94$), while male targets remain largely stable ($-0.71 \rightarrow -0.64$, $\Delta = +0.08$). This indicates that fine-tuning increased the model's tendency to link female-denoting words with adjectives.

For *verb-based* templates, the pre-trained model shows a stronger negative bias for female targets (-1.91 ± 3.42) than male targets (-0.88 ± 2.79), suggesting verbs related to cognition, perception, social, and work domains were less likely to be predicted for female subjects. After fine-tuning, association scores increase for both genders: for male-denoting words, from -0.88 to 0.57 ($\Delta = +1.45$), and for female-denoting words, more

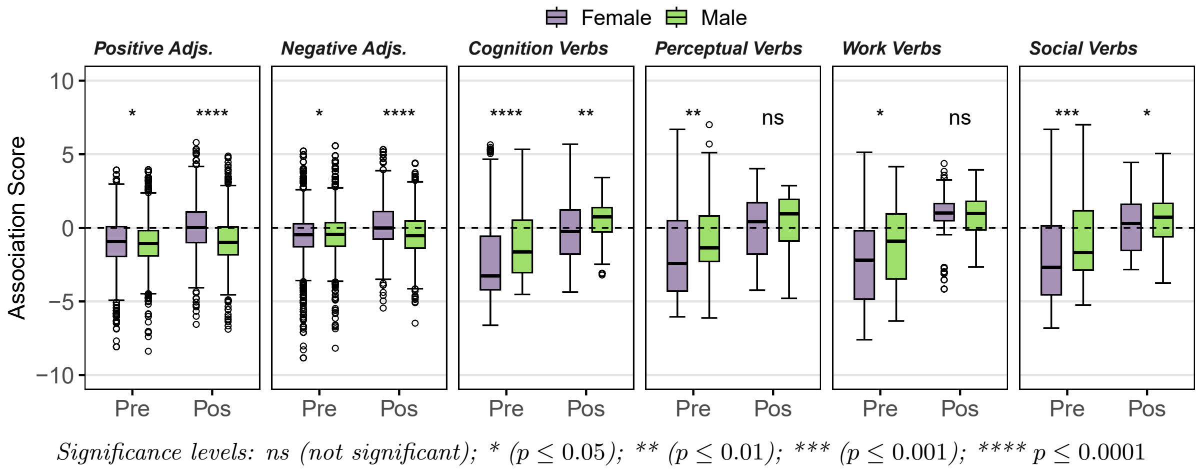


Figure 6.15: Gender association scores for adjective and verb categories, comparing pre-trained (Pre) and fine-tuned (Pos) models. Significance levels are indicated, with *ns* denoting non-significant results.

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 6.15 shows average association scores for female and male targets across adjective polarities and verb categories. In the pre-trained model, both positive and negative adjectives yield negative scores for both genders, with no significant difference (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, pre-trained scores are negative across all categories, more pronounced for female targets in cognition, social, and work domains. 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. 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.

Discussion. Our findings show that fine-tuning BERTimbau on a curated corpus of Portuguese-language 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. This shift, particularly towards positive adjectives, 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 [32].

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. 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 6.15). 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 [32].

Overall, our findings show that fine-tuning on literary texts does not uniformly reduce or amplify gender bias. Instead, it attenuates some pre-existing biases, such as the strong disassociation of female subjects from verbs, while simultaneously reinforcing others that are characteristic of literary discourse, including the stronger association of female characters with descriptive attributes and male characters with agentic roles.

6.3.2 Bias in Generated Text

The rapid development of Large Language Models (LLMs) has transformed Natural Language Processing, enabling applications ranging from conversational agents to literary text generation [185, 195]. These models can produce coherent and stylistically rich narratives, expanding possibilities in digital humanities and computational creativity [196]. However, despite their impressive generative capacity, LLMs often reproduce or amplify social biases present in their training data, including gender stereotypes that influence how characters are portrayed and described [28, 120].

When generating or continuing narratives, especially when gender is under-specified, LLMs may assign genders or traits based on biased linguistic patterns learned from data. For instance, as shown in Figure 6.16, two different models (Mistral and Llama) produce coherent continuations for a masked excerpt from *Iracema* by José de Alencar but assign different gender pronouns to the same masked character. While such variation may appear neutral in isolation, systematic analysis across many generations can reveal persistent patterns of gendered associations that reflect historical inequalities [12].

These biases largely stem from the uncurated nature of large-scale training corpora, often composed of web texts and public-domain books written in historical contexts domi-

Martim banhou-se nágua do rio, e passeou na praia para secar o corpo ao vento e ao sol. `[gender = male]`.

`[Mistral]`
`[MASK]` **banhou-se nágua do rio, e passeou na praia para secar o corpo ao vento e ao sol.** Depois de passear pela praia, a pé, Personagem se sentou sobre uma pedra à beira-mar e olhou a paisagem que estendia-se diante dela. O mar brilhante refletia a luz do Sol, que se puxava para o horizonte.

`[Llama]`
`[MASK]` **banhou-se nágua do rio, e passeou na praia para secar o corpo ao vento e ao sol.** Depois de passar uma hora relaxada na praia, `[MASK]` decidiu que era hora de encontrar algo para comer. Ele sabia que havia um restaurante cerca de 100 metros da praia, chamado "O Pescador", conhecido por seus frutos do mar frescos e deliciosos.

Figure 6.16: Example of two LLMs (Mistral and Llama) assigning different gender pronouns to a masked character (`[MASK]`) in generated narrative continuations from an excerpt of "Iracema" by José de Alencar.

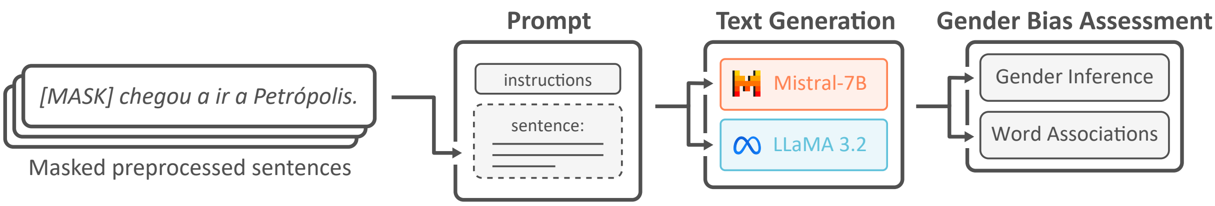


Figure 6.17: Overview of the steps followed to assess gender bias in Portuguese-language narrative generation.

nated by rigid gender norms [14, 27]. As a result, models internalize linguistic regularities that reinforce stereotypical gender representations in generated texts. Although these issues have been widely studied in English and other high-resource languages [29, 53], research focusing on Portuguese remains scarce [197, 198], despite the language's global significance and its rich literary tradition.

In this section, we investigate gender bias in Portuguese-language literary text generation by analyzing how LLMs extend narratives from public-domain works. We extract masked sentences, generate continuations using different LLMs, and apply gender inference and association analyses to evaluate bias patterns. Figure 6.17 provides an overview of the methodological pipeline adopted in this analysis.

6.3.2.1 Corpus

The corpus employed in this section derives directly from the Portuguese-language literary dataset described in Section 6.3.1.1. As summarized in Table 6.3, the base corpus contains 592 prose works spanning from 1804 to 1998. Here, beyond the initial preprocessing step, the *PORTALIA* framework is further applied to the texts to perform both *Character Identification* and *Gender Inference* stages. Each literary work is thus transformed into a structured collection of sentences in which character entities are explicitly identified and annotated with gender information.

For the purposes of LLM-based narrative generation, we apply additional filtering and balancing steps to the same dataset. Sentences shorter than five tokens are removed, as are those lacking clear narrative structure or containing explicit gender markers (e.g., *ele*, *ela*, *dele*, *dela*). Next, only sentences where the `PERSON` entity is classified as a proper noun (`PROPN`) and serves as the nominal subject or direct object are retained, ensuring focus on character-centered actions relevant to gender analysis. Finally, we filter out sentences containing explicit gender markers, such as pronouns (*ele*, *ela*) or possessives (*dele*, *dela*), then allowing the model to infer gender without linguistic cues.

To balance representation across texts while maintaining computational efficiency, we select the first 100 sentences of each literary work. After filtering, the final corpus consists of 28,317 sentences and 774,516 tokens. The dataset is publicly available at [199].

6.3.2.2 Text Generation

To generate narrative continuations, we employ two instruction-tuned generative models: *Mistral-7B-Instruct* and *LLaMA 3.2-3B* (see Appendix F). These models are selected for their ability to produce coherent, high-quality continuations while remaining open source and suitable for controlled experimentation [200, 201]. As both have undergone instruction tuning, they are particularly well-suited to structured prompting in creative and narrative contexts.

To encourage more diverse and creative responses, we set the temperature[8] parameter to 0.9, following previous studies that recommend higher temperatures for text generation [120, 202]. This configuration increases the probability of selecting less frequent tokens, leading to greater variation in the generated continuations while maintaining coherence. A higher temperature is particularly suitable for creative tasks such as narrative generation, where diversity in output is desirable.

To ensure consistency across different inputs, we use a structured prompting strategy that explicitly instructs the model to continue the narrative in Portuguese, focusing on describing the actions of the character mentioned in the input sentence. To mitigate direct gender bias in the model response, we mask the character's name using the placeholder `[MASK]`, requiring the model to infer or generate a continuation without explicit gender cues. This approach ensures that gender-specific linguistic markers do not influence the model's output, allowing us to evaluate implicit biases in narrative generation. The exact prompt format is as follows.

[8] In the context of language models, *temperature* is a parameter that controls the randomness of token selection during text generation. Lower values (e.g., close to 0) make the model more deterministic, favoring the most probable tokens, and resulting in more repetitive or conservative outputs. In contrast, higher values increase the likelihood of selecting less probable tokens, promoting diversity and creativity.

> **Instructions:** continue writing in pt-br the story from the following sentence, describing what the character did next. The character is `[MASK]`.
>
> ---
>
> **Sentence:** {`sentence`}

While the main instructions guiding the models are made in English, they explicitly direct the model to generate text in Portuguese (pt-br), as specified in the prompt. This design choice is aligned with common practices in multilingual instruction-tuned models, where English instructions are often used to control behavior across multiple languages without degrading performance in the target language [203–205]. It also provides a controlled setup to ensure that variation in outputs is driven by the model's internal representations rather than inconsistencies in prompting.

6.3.2.3 Gender Bias Assessment

To assess how generative models infer and represent gender in narrative continuations, we conduct two distinct analyses: (*i*) *gender inference*, which identifies linguistic markers to determine the gender of characters; and (*ii*) *word association* analysis, which explores semantic associations of gendered terms to uncover broader stereotypes or biases.

6.3.2.3.1 Gender Inference Analysis Because Portuguese encodes grammatical gender through pronouns, articles, determiners, and adjectives, gender can often be inferred using rule-based heuristics [32]. Therefore, we infer the gender of entities in the generated texts by relying on the linguistic markers present. We use the same *Gender Inference* module (Section 5.1) applied to the original literary corpus to ensure a fair and consistent comparison. However, this approach may struggle with ambiguous cases in which conflicting gender markers appear or when the context does not provide explicit linguistic cues.

Specifically, we use the *spaCy* library with the pre-trained Portuguese model[9] for syntactic and dependency parsing. This allows us to analyze the syntactic structure of each generated text and identify key contextual elements that may signal gender. The parser identifies modifiers syntactically linked to the masked entity (e.g., articles; adjectives such as "bonito" for male vs. "bonita" for female) and associated pronouns (e.g., "ele" for he, "ela" for she), which serve as contextual gender indicators. Dependency relations are inspected to ensure that these gendered markers refer to the masked entity rather than another character in the sentence. If the masked entity is not explicitly mentioned in the generated continuation, we assume that the first nominal subject or object (`PROPN`) serves as its referent.

[9] `pt_core_news_lg`, trained on the *UD Portuguese Bosque* treebank.

Gender classification is then assigned based on the dominant morphology of the gendered markers associated with the entity. If most detected markers display masculine morphology, the entity is labeled as *male*; if predominantly feminine, as *female*. In cases with conflicting or absent gender cues the gender is marked as *unknown*.

6.3.2.3.2 Word Association Analysis To investigate potential gendered patterns in the generated narratives, we analyze whether words associated with male and female entities exhibit systematic differences in their semantic contexts. We perform this analysis by: (i) training word embedding models on both original and generated texts; and (ii) measuring biases in word associations.

Training Word Embeddings. We train two widely used word embedding models, *Word2Vec* [206] and *FastText* [207], on the filtered literary corpus and on the LLM-generated continuations. *Word2Vec* is trained using the skip-gram architecture with a window size of five, negative sampling of five, and 300-dimensional embeddings. *FastText*, which incorporates subword information and is particularly suitable for morphologically rich languages such as Portuguese, is trained with identical hyperparameters. Both models are trained for 50 epochs to ensure stable and well-converged embeddings.

Measuring Word Associations. Biases in semantic associations are quantified using the WEAT [67]. This method measures the relative strength of associations between gendered word sets (e.g., male vs. female) and predefined thematic categories (e.g., career vs. family, agency vs. passivity) using cosine similarity in the embedding space. The WEAT effect size d is calculated as:

$$d = \frac{\text{mean}_{x \in X} S(x, A, B) - \text{mean}_{y \in Y} S(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} S(w, A, B)}, \tag{6.3}$$

where X and Y are target word sets, A and B are attribute word sets, and $S(w, A, B)$ is the difference in average cosine similarity between word w and the attribute sets:

$$S(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b). \tag{6.4}$$

A positive effect size ($d > 0$) indicate stronger association of X with male-related terms, while negative values indicate stronger association with female-related terms. Values close to zero suggest little or no measurable bias. For instance, if X represents career-related words, Y family-related words, A male-associated terms, and B female-associated terms, a positive d implies a closer association of career words with male terms.

Word Sets. Male- and female-associated word sets, as well as target word sets, are defined using BP-LIWC2015 [191, 192], the Brazilian Portuguese adaptation of the Linguistic Inquiry and Word Count lexicon. LIWC categorizes words into 73 thematic groups capturing psychological, social, and linguistic characteristics. Examples of relevant categories are shown in Table 6.6.

Table 6.6: Example words from BP-LIWC2015 categories.

Category	Examples
`male`	senhor, pai, marido, homem, amigo, menino
`female`	senhora, mãe, esposa, mulher, amiga, menina
`cogproc`	refletir, concluir, identificar, curioso, racional
`feel`	sentir, sentimento, acariciar, suavemente, sentia
`health`	saudável, saúde, medicina, medicamento
`home`	casa, lar, cama, cozinha, quintal, sala, banheiro
`insight`	saber, entender, buscar, aprender, estudar, supor
`leisure`	brincadeira, entretenimento, divertir, jogar
`negemo`	ódio, raiva, triste, desespero, problema
`percept`	ver, ouvir, sinto, olhou, olhar, escutar, tocar
`posemo`	bom, alegria, feliz, alegre, amor, beleza
`risk`	perigo, risco, ameaça, perdido, problems, crise
`work`	carreira, emprego, trabalho, projeto, profissão

To establish gendered reference sets, we use the *female* and *male* categories from LIWC, ensuring a linguistically grounded approach to gender representation. To construct the target word sets, we analyze the following category pairs:

- **Cognitive vs. Feel:** using the `cogproc` and `feel` categories, this set contrasts analytical and logical thinking with emotions and feelings, reflecting potential gendered stereotypes in cognitive versus affective attributes.
- **Insight vs. Percept:** using the `percept` and `insight` categories, this set contrasts perceptual with insight processes, investigating whether gendered differences emerge in association with sensory experience versus introspection.
- **Positive vs. Negative:** using the `posemo` and `negemo` categories, this set contrasts words associated with positive and negative emotions, reflecting gendered portrayals in terms of emotional qualities.
- **Risk vs. Health:** using the `health` and `risk` categories, this set contrasts terms related to well-being with those associated with danger and recklessness, revealing potential gendered perceptions of cautiousness and risk-taking.
- **Work vs. Home:** using the `work` and `home` categories, this contrasts career and domestic roles, often associated with gendered expectations in societal contexts.
- **Work vs. Leisure:** using the `leisure` and `work` categories, this contrasts work-related terms with those associated with relaxation and hobbies.

To ensure comparability, we randomly select 30 words from each LIWC category to create balanced word sets. Statistical significance of WEAT effect sizes is assessed via permutation testing, in which target labels are shuffled and effect sizes recomputed

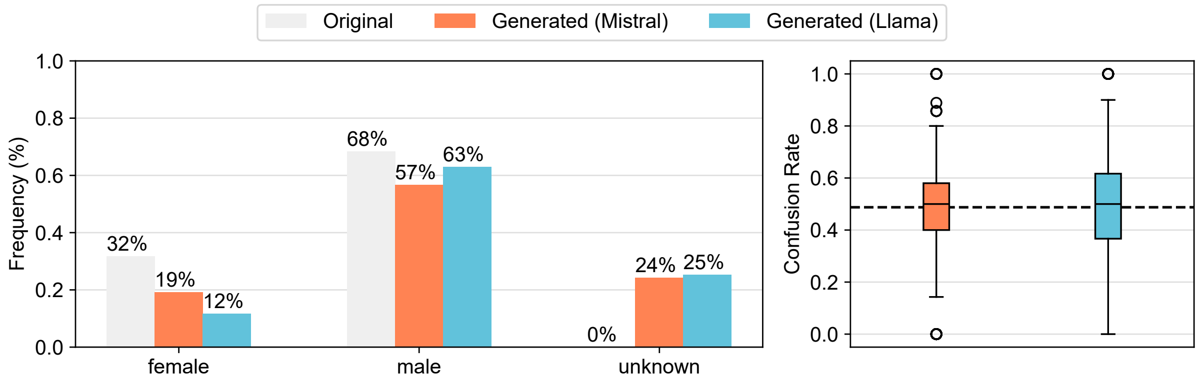


Figure 6.18: **(left)** Percentage of male, female, and unknown entities in the original texts and the outputs of Mistral and Llama models. **(right)** Gender confusion rate, indicating how often entities were assigned a different gender compared to their original reference (dashed line indicates the overall average value).

across multiple iterations to compute p-values. This procedure ensures that observed associations reflect systematic biases rather than random distributional variation.

6.3.2.4 Experimental Results

This section presents the results of our analysis on gender representation and bias in narrative generation. We first explore the gender distribution of entities inferred by the models, followed by semantic associations of gendered terms, and conclude with an interpretation of observed patterns.

6.3.2.4.1 Gender Distribution To assess how generative models assign gender in narrative continuations, we compare the gender distribution of entities in the generated texts to that of the original texts (Section 6.3.2.1). Specifically, we assess whether the models replicate the gender proportions found in the original texts or exhibit any shifts in representation. We also analyze the gender confusion rate (or error rate), which quantifies the frequency with which an entity's gender in the generated text differs from its original reference. This is calculated as the proportion of gender mismatches between the original and generated texts.

Figure 6.18(left) illustrates the overall gender distribution in the original and generated texts. The original texts (which, due to our input selection focusing on specific characters, have these characters' genders as either male or female, i.e., 0% unknown for the source entities) show a clear gender imbalance, with 68% male and 32% female entities, consistent with previous findings on gender representation in literary works [32].

Both models reflect this male over-representation in their outputs, with male entities accounting for approximately 60% of identified entities, and female entities for approximately 15%. Statistical analysis using chi-squared tests indicates that these overall gender distributions from Mistral ($\chi^2 = 0.35$, $p = 0.84$) and Llama ($\chi^2 = 0.60$, $p = 0.74$)

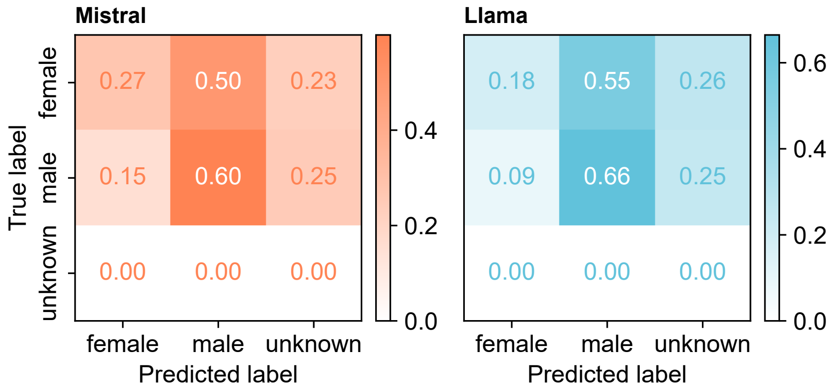


Figure 6.19: Confusion matrices for gender classification in generated texts. The matrices show the true vs. predicted gender assignments.

are not significantly different from the original texts. Moreover, the output distributions of Mistral and Llama are not significantly different from each other ($\chi^2 = 0.05$, $p = 0.97$). These results indicate that the models largely replicate existing gender imbalances, with a notable tendency to produce gender-neutral attributions when context is ambiguous.

In particular, both models also generate a substantial portion of entities labeled *unknown* (around 25%). This could indicate instances where the models generate ambiguous or gender-neutral descriptions, making it difficult for the inference method to assign a gender. Alternatively, it may reflect the models' uncertainty in continuing masked sentences, resulting in non-explicit gender references. In either case, generative models may avoid committing to a specific gender when the context is ambiguous, aligning with previous findings [120].

Regarding the confusion rate, Figure 6.18(right) shows that both models show an average error rate of around 50%, with no statistically significant difference between them. This indicates that both models are equally likely to assign genders to entities in the generated texts incorrectly. However, Figure 6.19 reveals that these errors are unequal between genders. Both models achieve relatively high accuracy for male entities (around 60%) but significantly lower accuracy for female entities, with Mistral correctly classifying only 27% and Llama only 18% of female entities.

In addition, both models exhibit a systematic bias towards misclassifying female entities as male. That is, when models misclassify an entity, it is far more likely to be a female entity wrongly assigned as male rather than vice versa. This bias is consistent with previous research showing that male-associated words dominate in language models trained on literary corpora, leading to stronger male-centered defaults [12].

6.3.2.4.2 Word Associations To further investigate gender bias in generated texts, we analyze word associations using word embeddings trained on three different corpora: the original texts, and texts generated by both the Mistral and Llama models. We train

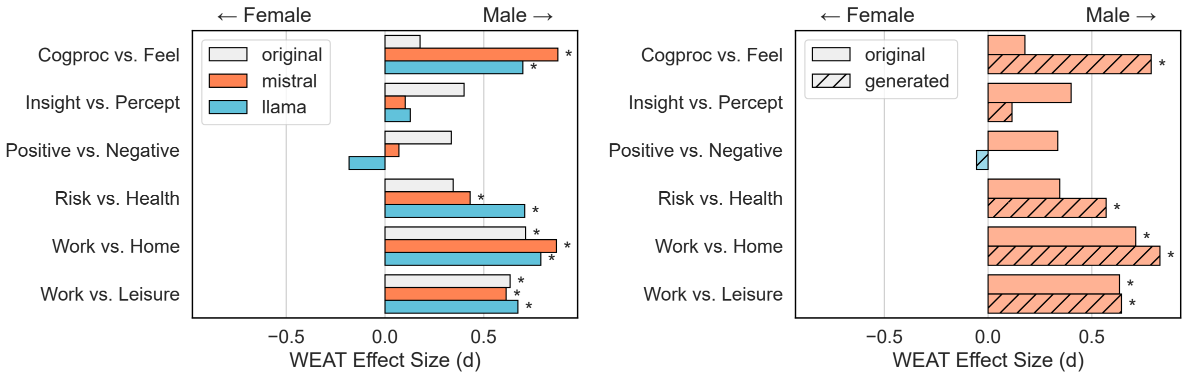


Figure 6.20: **(left)** Average WEAT effect sizes (d) grouped by target word sets and dataset (original, Mistral, Llama). **(right)** Average WEAT effect sizes grouped by target word sets and type of text (original vs. generated). Asterisks (*) indicate statistically significant results ($p < 0.05$).

Word2Vec and FastText models and compute the WEAT effect sizes to quantify the relative strength of gender associations in ten pairs of predefined categories (see Section 6.3.2.3.2).

To focus on broader trends in gender bias, we aggregate the results from both embedding models. Although these models differ in architecture—Word2Vec relies on local context windows, whereas FastText incorporates subword information—they capture similar semantic relationships and gender associations.[10] By aggregating the results, we can highlight consistent patterns of bias across embedding methods rather than emphasizing minor differences between the models. This approach aligns with our goal of investigating gender bias in generative models at a high level rather than focusing on the specific nuances of individual embedding techniques.

Overall Trends in Gender Associations. Figure 6.20(left) presents the overall average WEAT effect sizes (d) grouped by the target word sets and the dataset, while Figure 6.20(right) groups them by target word sets and text type (original vs. generated). The results include an asterisk to indicate statistical significance at the 5% level ($p < 0.05$). Here, a positive effect size ($d > 0$) indicates a stronger association of the first target category with male-associated words, while a negative effect size ($d < 0$) indicates a stronger association with female-associated words.

Our findings reveal that both original and generated texts exhibit similar gendered word associations, suggesting that generative models tend to preserve existing biases. However, effect sizes vary across categories, indicating that some associations become more or less pronounced in the generated texts compared to the original corpus. For example, in *Cogproc vs. Feel*, the effect size is considerably smaller in the original texts ($d = 0.18$) than in the texts generated by Mistral ($d = 0.87$) and Llama ($d = 0.70$).

[10]Complete results are shown in Appendix F.

Table 6.7: Effect sizes (d) and p-values for comparisons of target words across different datasets. Asterisks (*) indicate statistically significant results ($p < 0.05$). Cell colors indicate the magnitude of effect sizes: light for small ($|d| < 0.5$), medium for moderate ($0.5 \leq |d| < 0.8$), and dark for large ($|d| \geq 0.8$).

Target Words	Data	d	p	s
Cogproc vs. Feel	Original	0.18	5.19e-1	-
	Mistral	0.87	3.28e-6	*
	Llama	0.70	4.68e-4	*
Insight vs. Percept	Original	0.40	6.04e-2	-
	Mistral	0.10	8.15e-1	-
	Llama	0.13	7.01e-1	-
Positive vs. Negative	Original	0.34	1.72e-1	-
	Mistral	0.07	9.06e-1	-
	Llama	-0.18	1.00e+0	-

Target Words	Data	d	p	s
Risk vs. Health	Original	0.35	1.43e-1	-
	Mistral	0.43	3.64e-2	*
	Llama	0.71	4.07e-4	*
Work vs. Home	Original	0.71	4.21e-4	*
	Mistral	0.87	7.17e-6	*
	Llama	0.79	7.57e-5	*
Work vs. Leisure	Original	0.63	2.11e-3	*
	Mistral	0.61	2.59e-3	*
	Llama	0.67	7.73e-4	*

This suggests that generative models can amplify associations between male terms and cognitive processes, while female terms are more strongly linked to emotional attributes.

Conversely, in *Insight vs. Percept*, the effect size is higher in the original texts ($d = 0.40$) than in the generated texts (Mistral: $d = 0.10$, Llama: $d = 0.13$), indicating that the association between male terms and concepts related to insight is slightly reduced in the generated texts. This suggests that generative models introduce subtle shifts in gendered attributes, likely due to differences in how they generalize from training data.

Stereotypes in Gendered Associations. Table 6.7 presents the effect sizes and p-values for comparisons of target words in different datasets. While most effect sizes remain below 0.8 (a commonly used threshold for large effects[11]), they still reflect meaningful gender biases. Two sets of words exhibit particularly strong associations: *Cogproc vs. Feel* ($d = 0.87$, Mistral) and *Work vs. Home* ($d = 0.87$, Mistral). These results suggest persistent stereotypes linking male-associated words to professional and cognitive terms, while female-associated words remain closely tied to domestic and emotional ones.

Regardless of the effect size magnitudes, the bias direction remains consistent across datasets. Categories such as *Cognitive Processes*, *Risk*, and *Work* show a stronger association with male terms, reinforcing traditional stereotypes that link men with analytical thinking, risk-taking, and professional domains. In contrast, *Home*, *Leisure*, *Health*, and *Feel* tend to be more closely associated with female terms, reflecting persistent gendered expectations that emphasize domesticity, relaxation, and caregiving roles.

These findings align with previous research on gender bias in language models and social stereotypes. For example, Huang et al. [12] found that in stories generated by GPT-2, female protagonists tend to have motivations related to body, sexuality, and family; whereas male protagonists' actions are driven by power, risk, and violence. Similarly, Lucy and Bamman [120] observed that GPT-3-generated stories often associate female

[11]Conventional thresholds for Cohen's d classify 0.2 as small, 0.5 as medium, and 0.8 as large [67].

characters with topics related to family, emotions, and body parts; while male characters are more frequently related to politics, war, sports, and crime.

Generative Models. Our results show that gender bias amplification is consistent across both Mistral and Llama models, indicating that this is a general characteristic of such generative models rather than a model-specific artifact. This amplification effect is particularly evident in domains traditionally linked to gender stereotypes, such as the association of men with analytical thinking and professional settings and women with emotional expression and domesticity.

6.3.2.4.3 Discussion Overall, our results demonstrate that generative language models reproduce and, in some cases, amplify gender biases present in historical literary texts. Both Mistral and Llama reflect the over-representation of male characters and under-representation of female entities, suggesting that these models perpetuate historical disparities learned from their training data. Indeed, while the specific literary texts used for our narrative continuations may not have been part of the models' exact training sets, it is reasonable to assume these LLMs were trained on similar datasets that invariably include texts reflecting historical and societal gender imbalances; such training data can subsequently influence their outputs when generating new narratives.

Moreover, high rates of gender misclassification, particularly the tendency to label female entities as male, highlight asymmetric gender inference in generated narratives. Word association analyses further reveal persistent stereotypes: male terms are more strongly linked to cognitive processing, professional success, and risk-taking, while female terms remain associated with emotional expression, domesticity, and caregiving. These patterns echo societal norms and prior findings on bias in language models [12, 67, 120].

Another notable aspect of our findings is the presence of a substantial proportion of entities classified as *unknown* in the generated texts. This could indicate instances where the models generate more ambiguous descriptions, potentially reflecting a shift toward gender-neutral language in uncertain contexts. While this may suggest some level of neutrality, further investigation is needed to explore whether these cases stem from an unbiased generative process or simply reflect model uncertainty and avoidance.

The consistency of these patterns across both models implies that such biases are not architecture-specific but rather inherent to generative models trained on corpora reflecting societal norms [44]. Mitigating these biases requires multifaceted approaches, including transparent data curation, bias-aware training strategies, and systematic evaluation frameworks that consider linguistic and cultural dimensions.

6.4 Overall Considerations

This chapter operationalized the final component of the *PORTALIA* framework, *Gender Bias Measuring*, by exploring gender asymmetries in both Portuguese-language literary texts and the computational models trained or applied to them. Building upon the gender-labeled excerpts produced in the *Character Identification* and *Gender Inference* stages, the analyses presented here demonstrated how the framework enabled a unified and systematic examination of gender bias across linguistic and computational dimensions.

Section 6.2 focused on gender bias in language, applying the excerpt-based analytical units defined earlier in the framework to quantify representation disparities, explore lexical and syntactic descriptors, and evaluate stereotypical patterns in character portrayal. These findings reveal persistent imbalances throughout the corpus, with male characters occupying the majority of narrative space and female characters being more frequently associated with domesticity, emotion, and passivity. By relying on dependency-based feature extraction, the analysis provides a level of granularity and comparability that would be difficult to achieve through manual approaches alone.

Section 6.3 complemented this perspective by investigating *bias in methods*. Using the same excerpt structure and gender labels derived earlier in the framework, we probed how embeddings and generative language models encode gendered associations found in the corpus. The results showed that contextual embeddings learned from literary data, as well as LLMs exposed to similar patterns, tend to reflect historical imbalances: male terms appear more strongly associated with professional, agentive, and cognitive domains, while female terms are linked to affective or domestic spheres. The recurrence of these patterns across models suggests that bias is not model-specific, but emerges from the interaction between training data and linguistic representations.

Taken together, the two analyses illustrate how the final stage of the framework consolidates insights produced by the earlier components. The literary analysis shows *where* and *how* gendered patterns originate in the texts, while the computational analysis demonstrates *how these same patterns are transmitted or reshaped* by NLP methods. This alignment between linguistic and computational evaluations exemplifies the modularity and extensibility of the framework: the same excerpt-based units can support traditional literary studies, embedding-based analyses, and LLM probing tasks.

Overall, the results presented in this chapter highlight the effectiveness and utility of our proposed framework. By integrating character identification, gender inference, and bias measurement into a reproducible workflow, *PORTALIA* enables large-scale analyses that bridge literary studies and computational modeling. From a literary perspective, the findings reinforce how Portuguese-language prose reflects and reinforces societal gender norms; from a computational perspective, they show how these norms can propagate to downstream models if not explicitly addressed.

Chapter 7

Conclusion and Future Work

In this chapter, we provide a synthesis of the contributions and findings of this Ph.D. dissertation. Section 7.1 summarizes the main results, highlighting the insights gained from the analyses of gender bias in Portuguese-language literary texts and computational methods. Section 7.2 discusses the limitations of the research and potential threats to validity. Section 7.3 outlines directions for future work. Finally, Section 7.4 presents the research products generated during the Ph.D., organized by dissertation chapter and including relevant byproducts, proving the broader impact of this dissertation.

7.1 Summary of Results

This dissertation aims to design and evaluate a computational framework for detecting and analyzing gender bias in Portuguese-language literary texts. Figure 7.1 provides a

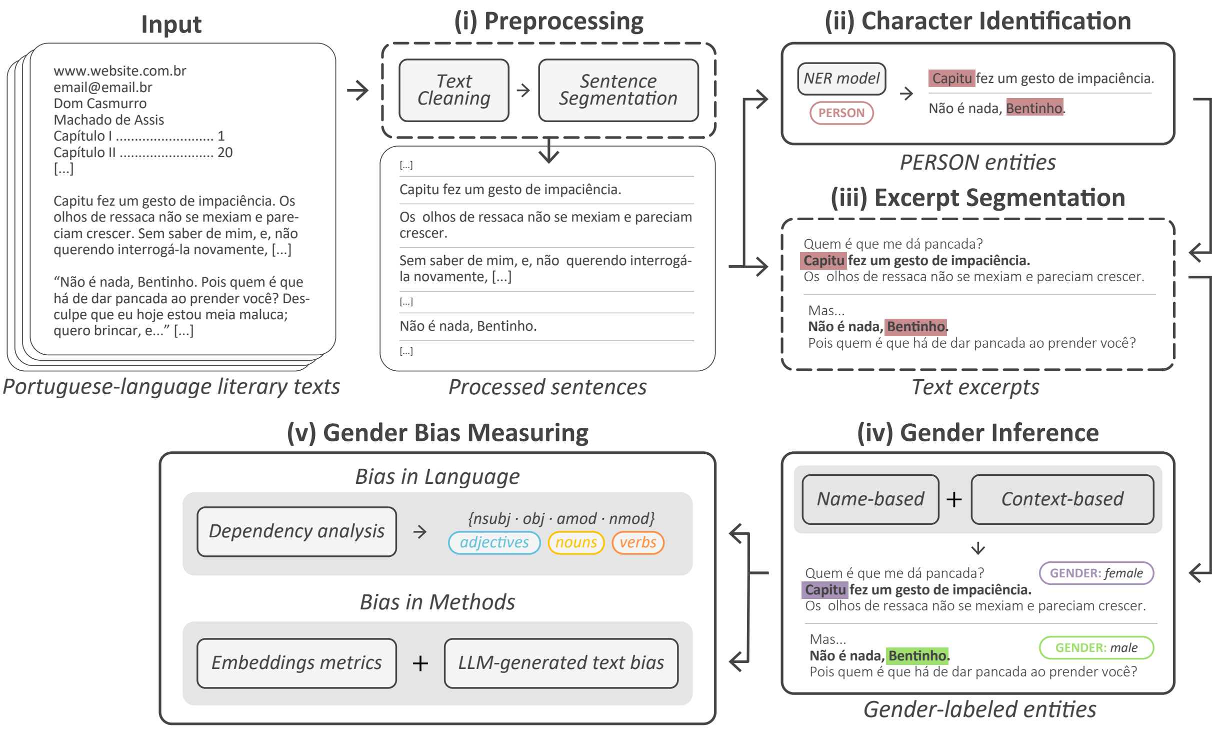


Figure 7.1: Schematic overview of the *PORTALIA* framework, illustrating the modular workflow from character identification to gender bias measurement in literary texts.

consolidated overview of the proposed *PORTALIA* framework, summarizing the complete methodological workflow developed throughout this work. The figure integrates all analytical stages discussed in the thesis, from character identification to gender bias measurement, and illustrates how the individual research goals are operationalized within a coherent, modular framework for large-scale literary analysis.

To ground the development of this framework, the dissertation begins with a systematic mapping of related work (Chapter 2), which establishes the conceptual and methodological foundations of gender bias research across Digital Humanities and Computer Science. This review identified persistent limitations in prior studies, such as restricted corpus diversity, a strong focus on English-language data, and a heavy reliance on manual annotation. These findings directly informed the formulation of four research goals (**RGs**), each addressed in a dedicated chapter and collectively guiding the design, implementation, and evaluation of the *PORTALIA* framework.

[**Related Work Review**] In Chapter 2, we conducted a comprehensive related work review to establish the conceptual and methodological foundations of this research, bridging perspectives from Digital Humanities and Computer Science. From the Humanities side, the review mapped how gender bias has been analyzed in literary studies across multiple dimensions (e.g., authorship, character traits, language, relationships, representation, and themes), revealing both the evolution and the limitations of prior work. From the computational perspective, it explored advances in gender bias detection within NLP, identifying common methodological issues such as binary gender modeling, restricted linguistic coverage, and the absence of standardized evaluation metrics.

The review exposed a series of intersecting gaps: limited corpus diversity (often centered on Western canonical authors), lack of longitudinal and intersectional analyses, methodological dependence on manual annotation, and the dominance of English-centric computational studies. It also emphasized the need for integrated frameworks capable of combining interpretative depth with computational scalability. By synthesizing insights from both domains, this chapter provided a unified conceptual foundation that guided the design of the computational framework proposed in the following chapters.

[**RG1**] **Develop a computational framework for measuring and analyzing gender bias in Portuguese-language literary texts.** In Chapter 3, we present *PORTALIA* (**POR**tuguese **T**ext **A**na**L**ysis for Gender-b**IA**s), a unified and modular framework that operationalizes gender bias analysis and measurement across multiple linguistic and interpretative levels. Designed to address persistent challenges identified in previous studies, the framework integrates NLP techniques into a unified and reproducible workflow. It is structured around three main analytical components: *Character Identification*, *Gender Inference*, and *Gender Bias Measuring*. Together, these modules enable a stepwise process that begins with the recognition of literary entities, continues with gender inference through linguistic and contextual evidence, and culminates in the quantification of gender asymmetries in both textual representations and computational models.

The modular architecture of the framework ensures flexibility and scalability, allowing its components to be independently replaced, extended, or adapted to different corpora, historical periods, and literary traditions. Moreover, by integrating computational precision with interpretative depth, *PORTALIA* bridges disciplinary boundaries between literary studies and computer science, enabling a more systematic and context-aware investigation of gender bias in literature. Through its implementation for Portuguese-language corpora, the framework also contributes to expanding the scope of gender bias research beyond English, supporting more inclusive, cross-linguistic, and reproducible analyses of literary texts. The direct outcome of this research goal is the conference paper [33], which introduces and evaluates the *PORTALIA* framework.

[RG2] Develop and evaluate domain-adaptive strategies for identifying literary characters in Portuguese-language texts. In Chapter 4, we investigated the task of literary character identification, a foundational step for gender bias analysis in narrative texts. Unlike conventional NER tasks, recognizing literary characters requires handling diverse linguistic and narrative phenomena (e.g., symbolic naming, implicit references, and pronominal anaphora) that are especially prevalent in Portuguese-language prose. To address these challenges, we developed two domain-adaptive models, *LitBERT-CRF* and *LitBERTimbau*, which extend pre-trained transformer architectures through additional pre-training and fine-tuning on literary corpora. Complementing these models, we introduced *PPORTAL_ner*, an annotated dataset of Portuguese-language literary entities, specifically designed to support domain adaptation and evaluation in literary contexts.

Experimental results demonstrate that domain-adaptive pre-training significantly enhances performance compared to general-domain baselines, particularly in the identification of `PERSON` entities corresponding to literary characters. Both LitBERT-CRF and LitBERTimbau achieved F1-scores above 85%, confirming the effectiveness of integrating literary data into model training. Detailed error analysis revealed that adaptation reduces omission errors and improves boundary precision, though challenges remain in capturing complex or multi-token names and in balancing the detection of spurious entities.

Overall, this research goal advances the state of the art in literary NER for Portuguese, bridging gaps between computational modeling and narrative analysis. The resulting models and resources provide a significant step for large-scale, automated analyses of character representation, thereby enabling subsequent stages of gender bias detection. Moreover, by addressing language- and genre-specific challenges, the contributions of this chapter extend beyond the immediate context of gender studies, offering methodological insights applicable to domain adaptation and entity recognition in other forms of cultural and historical text analysis. The resulting resources and models underpin large-scale automated character analyses and include three publications: [34], [35], and [36].

[RG3] Develop and evaluate an automated approach for gender inference in Portuguese-language literary texts. In Chapter 5, we addressed the challenge of automatically inferring the gender of named entities in Portuguese-language literary texts. Unlike standard gender classification tasks, literary data introduce additional complexity due to stylistic variation, indirect references, and the frequent presence of unusual or ambiguous names. Furthermore, the scarcity of annotated resources in Portuguese-language literary language limits the applicability of fully supervised approaches. To address these challenges, we proposed a gender inference heuristic that integrates two complementary components: a *name-based* module relying on probabilistic gender associations from the *genderBR* database, and a *context-based* module that exploits morphological and syntactic gender agreement patterns derived from dependency parsing.

The experimental evaluation, based on a gold standard of annotated entities, demonstrated that the integrated heuristic substantially outperforms its individual components, achieving 76.2% accuracy and 93.0% coverage in the full three-way classification (*female*, *male*, *unknown*). When ambiguous entities were excluded, accuracy reached 98.8%, confirming that most residual errors result from referential uncertainty rather than systematic misclassification. Precision and recall were balanced across gendered classes, indicating that the combined system effectively integrates lexical reliability with grammatical flexibility.

Qualitative error analysis revealed that most inaccuracies stemmed from the underlying NER output (e.g., truncated or misclassified entities) or from complex narrative constructions involving collective, metaphorical, or contextually gender-neutral expressions. Nonetheless, the rule-based integration strategy significantly reduced the number of interpretative errors by prioritizing reliable name-based evidence and selectively incorporating contextual cues when lexical information was missing or ambiguous. Overall, this research goal resulted in an automated heuristic for gender inference in Portuguese-language literary corpora. The proposed approach is transparent, interpretable, and computationally efficient, enabling considerably reliable gender attribution. Two publications report this research goal's results: [37] and [38].

[RG4] Measure and analyze gender bias in both literary language and computational methods. In Chapter 6, we addressed the final research goal of this dissertation by investigating how gender bias manifests in Portuguese-language literary texts and how such patterns are reproduced, reshaped, or amplified by computational models trained on or exposed to these texts. Building upon the previous steps of character identification and gender inference, this stage operationalized the measurement and analysis of gender bias through two complementary analytical dimensions: bias in language and bias in computational representations.

The first dimension focused on *gender bias in literary language*, investigating lexical patterns, descriptive asymmetries, and narrative portrayals of male and female characters.

Through quantitative metrics and linguistic analyses, we identified enduring gender disparities across two centuries of Portuguese-language works, with male characters remaining dominant in visibility and agency, while female characters were frequently associated with domestic, emotional, or passive traits. These findings highlight how literature serves not only as a mirror of historical gender norms but also as a medium that reinforces them through stylistic and thematic conventions.

The second dimension investigated *gender bias in computational methods*, focusing on how modern NLP models encode and propagate these literary patterns. Experimental results showed that fine-tuning a pre-trained language model on the literary corpus does not simply increase or decrease gender bias, but reconfigures it. Fine-tuning reduced a strong pre-existing tendency of the model to dissociate female subjects from verbs related to action and cognition, but also reinforced a stronger association of female-denoting words with descriptive adjectives and the persistent linkage of male-denoting words with cognitive and professional domains. Similarly, generative models reproduced historical gender imbalances and frequently misclassified female entities as male, while word association analyses revealed stable stereotypes aligning masculinity with work, risk, and analytical reasoning and femininity with domesticity, relationality, and emotional expression.

Taken together, these findings demonstrate that gender bias in computational models is not an artifact of modeling alone, but a continuation of cultural and linguistic patterns embedded in the texts on which they are trained. By bridging literary analysis and computational evaluation, this research goal establishes a methodological foundation for tracing how gender bias moves from cultural production to algorithmic inference. More broadly, it reinforces the need for critical, interpretable, and corpus-aware approaches to fairness in natural language processing, particularly when working with culturally and historically situated texts. Three publications report these analyses: [33], [39], and [40].

7.2 Limitations and Threats to Validity

Despite the contributions and insights provided by this dissertation, some limitations must be acknowledged. These limitations pertain both to individual modules of *PORTALIA* framework and to general aspects of the study, and they highlight opportunities for future research and methodological improvements.

***PORTALIA* Framework.** The overall framework relies heavily on existing NLP tools, such as SpaCy models for dependency parsing, sentence segmentation, and part-of-speech tagging, as well as pre-trained NER models developed for literary texts. The quality and robustness of these tools directly influence the accuracy of downstream analyses. Errors in dependency parsing or entity recognition can propagate through the framework, resulting in incomplete or incorrect character identification, gender inference, or bias measurement. Additionally, while *PORTALIA* is designed to be flexible and modular, its implementation has been validated primarily on Portuguese-language literary texts, which may limit generalizability to other languages or textual genres.

Character Identification. The character identification module depends on the annotated *PPORTAL_ner* corpus, comprising 25 literary works. Although this corpus spans different periods and literary genres, it may not fully represent the diversity of Portuguese-language literature. Furthermore, manual correction and refinement were performed by a single annotator, introducing potential subjectivity and limiting assessment of inter-annotator agreement. Finally, the focus on works from the 19th and early 20th centuries introduces historical bias, limiting applicability to contemporary literature.

The pre-training and fine-tuning of domain-adaptive models also face constraints. Class imbalances, particularly for less frequent entity types such as `ORG` and `TIME`, may have influenced model performance. Complex or multi-token character names, stylistic variations, and historical linguistic differences further challenge the models, potentially leading to boundary errors or missing entities.

Gender Inference. The gender inference module integrates two complementary components: a name-based classifier grounded in the *genderBR* database and a context-based heuristic that exploits morphological and syntactic cues extracted through dependency parsing. While this approach achieves robust overall performance, several limitations must be acknowledged.

The name-based component is constrained by the coverage and provenance of the underlying resource. It relies on the *genderBR* database, which is derived from the Brazilian Census and provides high-precision gender associations for Brazilian Portuguese names. However, this design choice introduces potential biases when the method is applied to European Portuguese literary texts or to works from earlier historical periods. Naming conventions, name frequencies, and gender associations vary substantially across Portuguese-speaking regions and over time. As a result, names that are rare or ambiguous in Brazilian Portuguese may exhibit clearer gender associations in European Portuguese, or vice versa, and many archaic or foreign names commonly found in literary texts are not covered by contemporary census-based resources.

Consequently, the name-based component may underperform in European Portuguese texts or in pre-20th-century literature, where onomastic practices diverge markedly from modern Brazilian usage. Although the context-based component partially mitigates this limitation by leveraging grammatical agreement patterns, the overall system remains affected by a Brazilian-centric onomastic bias. This limitation underscores the importance of developing region- and period-aware name resources when extending gender inference to broader Portuguese-language literary traditions.

The context-based component, while broader in scope, is itself sensitive to the accuracy of dependency parsing and to the explicitness of gender markers in the text.

Literary language frequently employs stylistic inversions, ellipses, indirect reference, and figurative constructions, which can obscure grammatical agreement and lead to uncertain or conflicting gender cues. Finally, both components operate under a binary gender assumption, treating gender as a categorical variable and thus excluding non-binary or fluid identities. While this simplification reflects the dominant representations in the historical corpus analyzed, it remains a conceptual and ethical limitation that should be addressed in future extensions of the framework.

Gender Bias Measuring. The gender bias measurement module is constrained by both linguistic and computational assumptions. The linguistic analysis relies on syntactic dependencies extracted using the UD Portuguese Bosque treebank. Although this resource is widely adopted and robust, its annotation conventions do not fully capture the stylistic and diachronic variability of Portuguese-language literary works, especially in texts before the 20th century. Misclassifications may introduce noise into the dependency relations used to link characters to descriptive attributes. These errors do not invalidate the core analytical patterns but may obscure finer distinctions in how gendered traits are expressed.

Furthermore, the linguistic categories employed (e.g., physical, emotional, social, character) depend on lexicons that were constructed in contemporary contexts and may not perfectly align with historical semantic usage. As a result, some lexical items shift meaning across periods but are treated uniformly in the analysis. This may lead to anachronistic interpretations or an underestimation of nuance in character representation. The analysis also assumes that morphological agreement in adjectives and verbs reliably signals the gender of the referenced entity; however, literary texts frequently employ stylistic inversion, ellipsis, and implicit subjects, limiting the precision of this inference. Thus, the linguistic patterns identified should be interpreted as probabilistic tendencies, rather than deterministic or exhaustive portraits of gender representation.

On the computational side, the methods used to assess gender bias in masked language models and generative models also introduce constraints. In the contextual embedding analysis, the log-probability association metric assumes that language models encode gendered attributes in local token-level predictions. While effective for capturing lexical co-occurrence biases, this approach cannot reveal higher-level narrative functions, discourse strategies, or irony—central features of literary style. Template-based probing provides controlled comparisons but artificially constrains syntactic and semantic variation, potentially overemphasizing stable stereotype associations.

Similarly, in the narrative generation analysis, the use of masked literary sentences foregrounds how models infer gender when explicit markers are absent. However, this setup does not account for broader narrative structures, authorial style, or genre. Moreover, the generative behavior of LLMs is sensitive to decoding parameters (e.g., temperature, top-k), which means alternative configurations may yield different gender assignments or descriptive focuses. The models themselves are trained on large-scale corpora whose composition and biases are only partially documented. Therefore, the observed patterns reflect not only biases from the literary corpus but also the much broader training environments of Mistral and LLaMA-based models.

Threats to Validity. Beyond module-specific limitations, several broader factors affect the validity and generalizability of this dissertation. The literary corpus, although historically and stylistically diverse, may not represent all Portuguese-language literature, particularly contemporary or non-canonical works. Evaluation metrics, while systematic, capture only a subset of the complex phenomenon of gender bias and may not fully account for intersectional effects with race, class, or sexuality. Lastly, the findings related to computational models reflect the specific architectures, prompts, and multilingual settings explored; results may vary for other LLMs or alternative experimental designs.

Large Language Models. This dissertation was developed in a context where LLMs were either not yet available or not sufficiently reliable for systematic literary analysis in Portuguese. Since then, the rapid evolution of instruction-tuned and multilingual LLMs, such as GPT-based models, Qwen, and other foundation models, has significantly changed the landscape of NLP. These models are now capable of performing tasks such as entity recognition, gender inference, co-reference resolution, and even interpretative summarization in a zero-shot or few-shot setting.

While this progress opens new opportunities, it also introduces important methodological trade-offs. LLM-based pipelines often operate as black boxes, offering limited transparency, reduced reproducibility, and sensitivity to prompt design. In contrast, the *PORTALIA* framework prioritizes interpretability, modularity, and explicit linguistic grounding, which are critical when analyzing culturally and historically situated texts. As a result, although modern LLMs could potentially replace individual modules of the framework, their use would shift the analysis from an interpretable, rule- and evidence-based process toward a more opaque, model-driven approach, with implications for accountability and scholarly interpretation.

7.3 Future Directions

Building on the findings and limitations discussed in the previous sections, several avenues for future research emerge, both at the level of individual modules and across the *PORTALIA* framework as a whole.

Character Identification. Future work will expand the annotated corpus to include texts from a broader range of authors, genres, and cultural contexts, improving its diversity and representativeness. Annotation quality and consistency can be enhanced through refined guidelines and a double-annotation process with adjudication to resolve discrepancies. Further research may explore hyperparameter optimization and advanced training protocols to improve the developed NER models' effectiveness.

Gender Inference. The heuristic for gender inference can be further refined in multiple ways. The name-based component may benefit from the integration of additional name-gender datasets and from dynamic thresholding to better handle ambiguous or rare names. The context-based component could incorporate deep learning models for contextual representation, capturing subtler linguistic cues and improving performance in stylistically complex or ambiguous sentences. Incorporating co-reference resolution would also enhance the inference of gender across multiple mentions of the same character throughout a literary work, enabling a more holistic analysis.

Gender Bias Measuring. Future work on gender bias measurement can deepen the analysis by moving beyond lexical patterns toward structural and narrative dimensions of character representation. One promising direction is the integration of narrative role and agency modeling, analyzing how characters participate in plot development, decision-making, and social interaction networks. Such approaches would allow the measurement of gender bias not only in how characters are described, but also in what they do, what is done to them, and how central they are to the narrative. Additionally, incorporating co-reference and character tracking across entire works would support more robust computation of character trajectories, enabling diachronic analyses of agency within a story. On the computational side, future research could explore bias mitigation techniques, such as adversarial training or representation-space regularization, assessing not only whether bias can be reduced but also how such interventions affect literary meaning, narrative coherence, and interpretability.

Overall Framework. At the framework level, future work should focus on expanding the scope, diversity, and interoperability of the system. First, extending the corpus beyond public domain works is essential for capturing contemporary literary production and more diverse authorial voices. This may involve negotiating access to non-public corpora, working with digital libraries, or partnering with research groups in literary studies. Second, improving interoperability between modules, particularly by integrating co-reference resolution and character coreference clustering, would enable more consistent tracking of characters across narratives, supporting richer longitudinal analyses. Third, future versions of the *PORTALIA* framework may incorporate interactive or semi-automated annotation workflows, allowing literary scholars to inspect and refine model outputs, thus strengthening interpretability and facilitating human-in-the-loop research. Finally, evaluating the framework across different text styles and languages would test its generality and open paths for comparative studies of gender representation in textual data.

7.4 Research Products

This section presents the publications resulting from this Ph.D. dissertation, including both direct products (Section 7.4.1) and byproducts (Section 7.4.2).

7.4.1 Direct Products

This section presents the direct products that resulted from this PhD research, organized by publication date. Direct products are publications and other materials that derive directly from the results and original contributions of this PhD, reflecting the core objectives, methods, and findings developed throughout the dissertation.

Dissertation-related Publications

1. **Mariana O. Silva**, Luiza de Melo-Gomes, and Mirella M. Moro. Gender Representation in Literature: Analysis of Characters' Physical Descriptions. In *Anais do XI Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2023)*, p. 17-24, 2023. [37] **Second Best Paper Award.**

2. **Mariana O. Silva**, Luiza de Melo-Gomes, and Mirella M. Moro. From Words to Gender: Quantitative Analysis of Body Part Descriptions within Literature in Portuguese. *Information Processing & Management*, p. 103647, 2024. [38]

3. **Mariana O. Silva** and Mirella M. Moro. Evaluating Pre-training Strategies for Literary Named Entity Recognition in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, p. 384–393, 2024. [34]

4. **Mariana O. Silva** and Mirella M. Moro. NLP Pipeline for Gender Bias Detection in Portuguese Literature. In *Anais do LI Seminário Integrado de Software e Hardware (SEMISH 2024)*, p. 169-180, 2024. [33]

5. **Mariana O. Silva** and Mirella M. Moro. PPORTAL_ner: An Annotated Corpus of Portuguese Literary Entities. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 12927–12937, 2024. [36]

6. **Mariana O. Silva** and Mirella M. Moro. Aprimorando o Reconhecimento de Entidades Nomeadas em Textos Literários em Português com Modelos Adaptativos. *Linguamática*, p. 77–94, 2025. [35]

7. **Mariana O. Silva**, Michele A. Brandão, and Mirella M. Moro. Rewriting Stories with LLMs: Gender Bias in Generated Portuguese-language Narratives. *Journal of the Brazilian Computer Society*, p. 1120–1136, 2025. [39]

8. **Mariana O. Silva**, Michele A. Brandão, and Mirella M. Moro. Gender Bias in Portuguese Literary Texts: A Masked Language Model Approach. In *Proceedings of the 16th Symposium in Information and Human Language Technology (STIL 2025)*, p. 407–419, 2025. [40]

Under Review

Mariana O. Silva, Michele A. Brandão, and Mirella M. Moro. Gender bias detection in literary texts: A systematic literature review. *ACM Computing Surveys*, 2025.

Datasets

1. **Mariana O. Silva**, and Mirella M. Moro. PPORTAL_ner: An Annotated Corpus of Portuguese Literary Entities. 2024. [176]
2. **Mariana O. Silva**, Michele A. Brandão, and Mirella M. Moro. Gender Bias in Portuguese Literary Texts: A Masked Language Model Approach. 2025. [194]
3. **Mariana O. Silva**, Michele A. Brandão, and Mirella M. Moro. Rewriting Stories with LLMs: Gender Bias in Generated Portuguese-language Narratives. 2025. [199]

Code and Models

1. Computational framework for gender bias in Portuguese-language literary texts proposed in Chapter 3: `https://github.com/marianaossilva/gender_pipeline`
2. NER models developed in Chapter 4: `https://huggingface.co/marianaossilva`

Further Publications

1. **Mariana O. Silva**, Gabriel P. Oliveira, and Mirella Moro. Data Insights on Gender Representation: Analyzing the Book and Music Industries. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados - Data Science for Social Good Brazilian Workshop (DS4SG)*, p. 338-347, 2024. [92]
2. **Mariana O. Silva**, Gabriel P. Oliveira, and Mirella Moro. Premiação de Mulheres na Literatura e na Música: Análises de Dados da Billboard e do Goodreads. In book: *A Internet como Campo de Disputas de Gênero*, p. 185-197, 2024. [DOI]
3. **Mariana O. Silva**, Gabriel P. Oliveira, and Mirella M. Moro. Analyzing Character Networks in Portuguese-language Literary Works. In *Anais do XII Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2023)*, p. 115-126, 2023. [208]

4. **Mariana O. Silva**, Clarisse Scofield, Luiza de Melo-Gomes, Juliana E. Botelho, Gabriel P. Oliveira, Danilo B. Seufitelli, and Mirella M. Moro. Brazilian Reading Preferences in Goodreads: Cross-state and Cross-region Analyses. *Brazilian Journal of Information Systems (ISys)*, p. 25:1–25:20, 2022. [209]

5. Clarisse Scofield, **Mariana O. Silva**, and Mirella M. Moro. What Makes a Book Successful? A Study on Portuguese-language Literature. In Anais Estendidos do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web - Concurso de Trabalhos de Iniciação Científica (WebMedia/CTIC 2022), p. 69-72, 2022. [DOI]

6. **Mariana O. Silva**, Clarisse Scofield, Luiza de Melo-Gomes, and Mirella M. Moro. Cross-collection Dataset of Public Domain Portuguese-language Works. *Journal of Information and Data Management (JIDM)*, v.13, n.1, 2022. [99]

7. Clarisse Scofield, **Mariana O. Silva**, Luiza de Melo-Gomes, and Mirella M. Moro. Book Genre Classification Based on Reviews of Portuguese-Language Literature. In *Computational Processing of the Portuguese Language (PROPOR 2022)*, p. 188-197, 2022. [210]

8. **Mariana O. Silva**, Clarisse Scofield, and Mirella M. Moro. PPORTAL: Public domain Portuguese-language literature Dataset. In *Anais do III Dataset Showcase Workshop (SBBD/DSW 2021)*, p. 77-88, 2021. [166] **Best Paper Award.**

9. **Mariana O. Silva**, Clarisse Scofield, Gabriel P. Oliveira, Danilo B. Seufitelli, and Mirella M. Moro. Exploring Brazilian Cultural Identity Through Reading Preferences. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2021)*, p. 115-126, 2021. [211] **Second Best Paper Award.**

Scientific Initiation Co-advisorships

1. Bárbara Martins Ribeiro Duarte (UFMG 2025). *Desvendando a Trama: Gênero Identitário e Conexões Sociais em Textos Narrativos Brasileiros.*

2. Marcele Louise Silva Araponga (UFMG 2025). *Desvendando a Trama: Gênero Identitário e Conexões Sociais em Textos Narrativos Brasileiros.*

3. Gabriella de Lima (UFMG 2025). *Desvendando a Trama: Gênero Identitário e Conexões Sociais em Textos Narrativos Brasileiros.*

7.4.2 Byproducts

Besides the products directly related to the Ph.D., the knowledge acquired during the process has also contributed to other publications and datasets on the contexts of social and music computing, as well as digital government. All of them are cited next.

Social and Music Computing

1. Gabriel R. G. Barbosa, Bruna C. Melo, Gabriel P. Oliveira, **Mariana O. Silva**, Danilo B. Seufitelli, and Mirella M. Moro. Hot Streaks in the Brazilian Music Market: A Comparison Between Physical and Digital Eras. In *Anais do XVIII Simpósio Brasileiro de Computação Musical*, p. 152-159, 2021. [DOI]

2. **Mariana O. Silva**, Gabriel P. Oliveira, Danilo B. Seufitelli, Anisio Lacerda, and Mirella M. Moro. Collaboration as a Driving Factor for Hit Song Classification. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia 22)*, p. 66-74, 2022. [DOI] **Best Paper Runner-up.**

3. Luiza de Melo-Gomes, Danilo B. Seufitelli, Gabriel P. Oliveira, **Mariana O. Silva**, Mirella M. Moro. Análise do Sucesso Musical no Brasil Utilizando Dados do Twitter (*Analysis of Musical Success in Brazil Using Twitter Data*). In: *Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados (WTAG 2022)*, p. 40-46, 2022. [DOI]

4. Danilo B. Seufitelli, Gabriel P. Oliveira, **Mariana O. Silva**, Gabriel R. G. Barbosa, Bruna C. Melo, Juliana E. Botelho, Luiza de Melo-Gomes, and Mirella M. Moro. From Compact Discs to Streaming: A Comparison of Eras within the Brazilian Market. *Vórtex Music Journal*, p. 1-28, 2022. [DOI]

5. Gabriel P. Oliveira, Gabriel R. G. Barbosa, Bruna C. Melo, Juliana E. Botelho, **Mariana O. Silva**, Danilo B. Seufitelli, and Mirella M. Moro. Musical Success in the United States and Brazil: Novel Datasets and Temporal Analyses. *Journal of Information and Data Management*, 2023. [DOI]

6. **Mariana O. Silva**, Gabriel P. Oliveira, Danilo B. Seufitelli, and Mirella M. Moro. Collaboration-Aware Hit Song Prediction. *Journal on Interactive Systems*, p. 201-214, 2023. [DOI]

7. Danilo B. Seufitelli, Gabriel P. Oliveira, **Mariana O. Silva**, Clarisse Scofield, and Mirella M. Moro. Hit song science: a comprehensive survey and research directions. *Journal of New Music*, p. 41-72, 2023. [DOI]

8. **Mariana O. Silva**, Gabriel P. Oliveira, Danilo B. Seufitelli, and Mirella M. Moro. Temporal Success Analyses in Music Collaboration Networks: Brazilian and Global Scenarios. *Vórtex Music Journal*, p. 1-27, 2023. [DOI]

9. Danilo B. Seufitelli, Gabriel P. Oliveira, **Mariana O. Silva**, Clarisse Scofield, and Mirella M. Moro. MGD+: An Enhanced Music Genre Dataset with Success-based Networks. In *Anais do V Dataset Showcase Workshop (DSW 2023)*, p. 36-47, 2023. [DOI]

Digital Government

1. **Mariana O. Silva**, *et al.* LiPSet: Um conjunto de Dados com Documentos Rotulados de Licitações Públicas. In *Anais do IV Dataset Showcase Workshop (DSW 2022)*, p. 13-24, 2022. [DOI]

2. **Mariana O. Silva**, *et al.* Análise de Sobrepreço em Itens de Licitações Públicas. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico (WCGE 2023)*, p. 118-129, 2023. [DOI]

3. **Mariana O. Silva**, Gabriel P. Oliveira, Lucas G. L. Costa, and Gisele L. Pappa. Evaluating Domain-adapted Language Models for Governmental Text Classification Tasks in Portuguese. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados (SBBD 2024)*, p. 247-259, 2024. [DOI] **Best Paper Runner-up.**

4. **Mariana O. Silva**, *et al.* Overpricing Analysis in Brazilian Public Bidding Items. *Journal on Interactive Systems*, p. 130–142, 2024. [DOI]

5. **Mariana O. Silva**, *et al.* LiPSet: A Comprehensive Dataset of Labeled Portuguese Public Bidding Documents. *Journal of Information and Data Management*, p. 196–205, 2024. [DOI]

6. **Mariana O. Silva**, Gabriel P. Oliveira, Lucas G. L. Costa, and Gisele L. Pappa. GovBERT-BR: A BERT-Based Language Model for Brazilian Portuguese Governmental Data. In *Brazilian Conference on Intelligent Systems (Bracis 2024)*, p. 19-32, 2024. [DOI]

Additional Datasets

1. **Mariana O. Silva**, Clarisse Scofield, Gabriel P. Oliveira, Danilo B. Seufitelli, and Mirella M. Moro. BraCID: Brazilian Cultural Identity Information Through Reading Preferences. 2021. [DOI]

2. **Mariana O. Silva**, *et al.* LiPSet: Um conjunto de Dados com Documentos Rotulados de Licitações Públicas. 2022. [DOI]

3. Danilo B. Seufitelli, Gabriel P. Oliveira, **Mariana O. Silva**, Clarisse Scofield, and Mirella M. Moro. MGD+: An Enhanced Music Genre Dataset with Success-based Networks. 2023. [DOI]

4. Gabriel P. Oliveira, **Mariana O. Silva**, Lucas G. L. Costa, Marco Túlio Dutra, and Gisele L. Pappa. ICPSet: Um Conjunto de Dados Estruturado de Itens de Compras Públicas. 2024. [DOI]

References

[1] M. C. Albrecht. The Relationship of Literature and Society. *American Journal of Sociology*, 59:425–436, 1954. doi:10.1086/221388.

[2] Terry Eagleton. *Literary Theory: An Introduction*. John Wiley & Sons, 2011.

[3] Elaine Showalter. Towards a Feminist Poetics. In *Women writing and writing about women*, pages 22–41. Routledge, 2012.

[4] André Botelho and Maurício Hoelz. Sociologias da Literatura: Do Reflexo à Reflexividade. *Tempo social*, 28(3):263–287, 2016.

[5] Antonio Candido. *Literatura e Sociedade*. Todavia, 2023.

[6] Elaine Showalter. Feminist criticism in the wilderness. *Critical inquiry*, 8(2):179–205, 1981.

[7] Rita Felski. *Beyond Feminist Aesthetics: Feminist Literature and Social Change*. Harvard University Press, 1989.

[8] Antonio Candido. *Iniciação à Literatura Brasileira*. Todavia, 2023.

[9] Larissa Andrade Tavares and Luciana Martins Quixadá. Sexismo e Representações do Feminino nos Romances Literários: Influências na Idealização de Relacionaentos Amorosos. *Psicologia & Sociedade*, 36:e277936, 2024.

[10] Cláudia Freitas and Flávia Martins. Bela, Recatada e do Lar: O Que a Mineração de Textos Literários Nos Diz Sobre a Caracterização de Personagens Femininas e Masculinas. *Fórum Linguístico*, 20(3):9118–9138, 2023.

[11] Madeleine Pownall and Nathan Heflick. Mr. Active and Little Miss Passive? The Transmission and Existence of Gender Stereotypes in Children's Books. *Sex Roles*, 89(11):758–773, December 2023. ISSN 1573-2762. doi:10.1007/s11199-023-01409-2.

[12] Tenghao Huang, Faeze Brahman, Vered Shwartz, et al. Uncovering Implicit Gender Bias in Narratives through Commonsense Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873. ACL, November 2021. doi:10.18653/v1/2021.findings-emnlp.326.

[13] Yawen Tsao. Gender Issues in Young Children's Literature. *Reading Improvement*, 45:108, 2008.

[14] Kexin Luo, Yue Mao, Bei Zhang, and Sophie Hao. Reflecting the Male Gaze: Quantifying Female Objectification in 19th and 20th Century Novels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13803–13812, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1205.

[15] Oscar Stuhler. The Gender Agency Gap in Fiction Writing (1850 to 2010). *Proceedings of the National Academy of Sciences*, 121(29):e2319514121, 2024. doi:10.1073/pnas.2319514121.

[16] David DeFranza, Himanshu Mishra, and Arul Mishra. How Language Shapes Prejudice Against Women: An Examination Across 45 World Languages. *Journal of personality and social psychology*, 119(1):7, 2020.

[17] Franco Moretti. *Distant Reading*. Verso Books, 2013.

[18] Matthew Jockers and Gabi Kirilloff. Understanding Gender and Character Agency in the 19th Century Novel. *Journal of Cultural Analytics*, 2(2), December 2016. doi:10.22148/16.010.

[19] Daniel Schulz and Štěpán Bahník. Gender Associations in the Twentieth-century English-language Literature. *Journal of Research in Personality*, 81:88–97, 2019. ISSN 0092-6566. doi:10.1016/j.jrp.2019.05.010.

[20] Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna M. Wallach, Isabelle Augenstein, et al. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 1706–1716. Association for Computational Linguistics, 2019. doi:10.18653/V1/P19-1167.

[21] Renata Vieira, Helena Cameron, Fernanda Olival, Fatima Farrica, et al. PLN e Humanidades Digitais. 2024.

[22] Frédérique Mélanie-Becquet, Jean Barré, Olga Clément Seminck, Clément Plancq, et al. BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature. In *Conference on Computational Literary Studies (CCLS 2024)*, 2024.

[23] Ted Underwood, David Bamman, and Sabrina Lee. The Transformation of Gender in English-Language Fiction. *Journal of Cultural Analytics*, 3(2), February 2018. doi:10.22148/16.019.

[24] Mayank Kejriwal and Akarsh Nagaraj. Quantifying Gender Disparity in Pre-Modern English Literature using Natural Language Processing. *Journal of Data Science*, 22 (1):77–96, 2023. ISSN 1680-743X. doi:10.6339/23-JDS1100.

[25] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7), April 2024. ISSN 0360-0300. doi:10.1145/3616865.

[26] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, et al. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29 of *NIPS*, 2016. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[27] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Procs ACM Conference on Fairness, Accountability, and Transparency*, FAccT, pages 610–623. ACM, 2021. doi:10.1145/3442188.3445922.

[28] Yufei Guo, Muzhe Guo, Juntao Su, et al. Bias in Large Language Models: Origin, Evaluation, and Mitigation. *arXiv preprint*, November 2024. doi:10.48550/arXiv.2411.10915. URL http://arxiv.org/abs/2411.10915.

[29] Roberto Navigli, Simone Conia, and Björn Ross. Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM Journal of Data and Information Quality*, June 2023. doi:10.1145/3597307.

[30] YiTian Ding, Jinman Zhao, Chen Jia, Yining Wang, Zifan Qian, Weizhe Chen, and Xingyu Yue. Gender Bias in Large Language Models across Multiple Languages: A Case Study of ChatGPT. In *TrustNLP*, pages 552–579. ACL, May 2025. URL https://aclanthology.org/2025.trustnlp-main.36/.

[31] Luiz Fernando F. P. de Lima and Renata Mendes de Araujo. A Call for a Research Agenda on Fair NLP for Portuguese. In *STIL*, pages 187–192. SBC, September 2023. doi:10.5753/stil.2023.233763.

[32] Cláudia Freitas and Diana Santos. Gender Depiction in Portuguese. In *Conference Reader: 2nd Annual Conference of Computational Literary Studies (CCLS2023)*, pages 4–30, 2023. doi:10.5281/zenodo.8093598.

[33] Mariana Silva and Mirella Moro. NLP Pipeline for Gender Bias Detection in Portuguese Literature. In *Anais do LI Seminário Integrado de Software e Hardware (SEMISH)*, pages 169–180. SBC, 2024. doi:10.5753/semish.2024.2914.

[34] Mariana O. Silva and Mirella M. Moro. Evaluating Pre-training Strategies for Literary Named Entity Recognition in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese, PROPOR 2024, Santiago de Compostela, Galicia/Spain, 12-15 March, 2024*, pages 384–393. Association for Computational Lingustics, 2024. URL https://aclanthology.org/2024.propor-1.39.

[35] Mariana O. Silva and Mirella Moro. Aprimorando o Reconhecimento de Entidades Nomeadas em Textos Literários em Português com Modelos Adaptativos. *Linguamática*, 17(1):77–94, Jun. 2025. doi:10.21814/lm.17.1.443.

[36] Mariana O. Silva and Mirella M. Moro. PPORTAL_ner: An Annotated Corpus of Portuguese Literary Entities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2024*, pages 12927–12937. European Language Resources Association (ELRA), 2024. URL https://aclanthology.org/2024.lrec-main.1132.

[37] Mariana O. Silva, Luiza Melo-Gomes, and Mirella Moro. Gender Representation in Literature: Analysis of Characters' Physical Descriptions. In *Anais do XI Symposium on Knowledge Discovery, Mining and Learning*, pages 17–24, Porto Alegre, RS, Brasil, 2023. SBC. doi:10.5753/kdmile.2023.232571.

[38] Mariana O. Silva, Luiza de Melo-Gomes, and Mirella M. Moro. From Words to Gender: Quantitative Analysis of Body Part Descriptions within Literature in Portuguese. *Information Processing & Management*, 61(3):103647, 2024. ISSN 0306-4573. doi:10.1016/j.ipm.2024.103647.

[39] Mariana O. Silva, Michele A. Brandão, and Mirella M. Moro. Rewriting Stories with LLMs: Gender Bias in Generated Portuguese-language Narratives. *Journal of the Brazilian Computer Society*, 31(1):1120–1136, Oct. 2025. doi:10.5753/jbcs.2025.5799.

[40] Mariana Silva, Michele Brand ao, and Mirella Moro. Gender Bias in Portuguese Literary Texts: A Masked Language Model Approach. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 407–419, Porto Alegre, RS, Brasil, 2025. SBC. doi:10.5753/stil.2025.37842.

[41] Jayde Pryzgoda and Joan C Chrisler. Definitions of Gender and Sex: The Subtleties of Meaning. *Sex roles*, 43:553–569, 2000. doi:https://doi.org/10.1023/A:1007123617636.

[42] Judith Butler and Gender Trouble. Feminism and the Subversion of Identity. *Gender trouble*, 3(1):3–17, 1990.

[43] Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. Gender Bias in Text: Origin, Taxonomy, and Implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 34–44, Online, August 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.gebnlp-1.5.

[44] Karolina Stanczak and Isabelle Augenstein. A Survey on Gender Bias in Natural Language Processing. *CoRR*, abs/2112.14168, 2021. URL https://arxiv.org/abs/2112.14168.

[45] Jean Barré and Yoann Dupont. Gender Bias in French Literature. In *Conference on Computational Humanities Research CHR2023*, December 2023. URL https://hal.science/hal-04447642.

[46] Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. In *Proceedings of the Tenth International Conference on Web and Social Media*, pages 112–120. AAAI Press, 2016. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13112.

[47] B. Smith. What Was "Close Reading"?: A Century of Method in Literary Studies. *Minnesota Review*, 87:57–75, 2016. doi:10.1215/00265667-3630844.

[48] David Bamman, Ted Underwood, and Noah A. Smith. A Bayesian Mixed Effects Model of Literary Character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi:10.3115/v1/P14-1035.

[49] Uri Margolin. Character. In *The Cambridge Companion to Narrative*, pages 66–79. Cambridge University Press, Cambridge, July 2007.

[50] Diana Santos, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, et al. Identifying Literary Characters in Portuguese - Challenges of an International Shared Task. In *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022*, volume 13208, pages 413–419. Springer, 2022. doi:10.1007/978-3-030-98305-5_39.

[51] Vincent Labatut and Xavier Bost. Extraction and Analysis of Fictional Character Networks: A Survey. *ACM Comput. Surv.*, 52(5):89:1–89:40, 2019. doi:10.1145/3344548.

[52] Diana Santos, Cristina Mota, Emanoel Pires, Marcia Langfeldt, et al. DIP - Desafio de Identificação de Personagens: Objectivo, Organização, Recursos e Resultados. *Linguamática*, 15(1):3–30, Jul. 2023. doi:10.21814/lm.15.1.399.

[53] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179, September 2024. doi:10.1162/coli_a_00524.

[54] Kate Crawford. The Trouble with Bias, 2017. URL `https://neurips.cc/virtual/2017/invited-talk/8742`. Keynote at NeurIPS.

[55] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.485.

[56] Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-3802.

[57] Dirk Hovy and Shrimai Prabhumoye. Five Sources of Bias in Natural Language Processing. *Lang. Linguistics Compass*, 15(8), 2021. doi:10.1111/LNC3.12432.

[58] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1630–1640. Association for Computational Linguistics, 2019. doi:10.18653/V1/P19-1159.

[59] Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. Assessing Gender Bias in Machine Translation: A Case Study with Google Translate. *Neural Comput. Appl.*, 32(10):6363–6381, 2020. doi:10.1007/S00521-019-04144-6.

[60] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing (EMNLP 2018), pages 4847–4853, Brussels, Belgium, 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1521/.

[61] Hadas Kotek, Rikker Dockum, and David Sun. Gender Bias and Stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701139. doi:10.1145/3582269.3615599.

[62] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain, April 2017. Association for Computational Linguistics. doi:10.18653/v1/W17-1609.

[63] Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, et al. Quantifying Gender Bias Towards Politicians in Cross-lingual Language Models. *Plos one*, 18(11):e0277640, 2023.

[64] Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-2031.

[65] Shikha Bordia and Samuel R. Bowman. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-3002.

[66] Yusu Qian. Gender Stereotypes Differ between Male and Female Writings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 48–53. Association for Computational Linguistics, July 2019. doi:10.18653/v1/P19-2007.

[67] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334): 183–186, 2017.

[68] Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 5034–5050, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.411.

[69] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1063.

[70] Wei Guo and Aylin Caliskan. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi:10.1145/3461702.3462536.

[71] Tommaso Dolci, Fabio Azzalini, and Mara Tanelli. Improving Gender-Related Fairness in Sentence Encoders: A Semantics-Based Approach. *Data Sci. Eng.*, 8(2): 177–195, 2023. doi:10.1007/S41019-023-00211-0.

[72] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, et al. Measuring and Reducing Gendered Correlations in Pre-trained Models. Technical report, Google Research, 2020. URL `https://arxiv.org/abs/2010.06032`.

[73] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-3823.

[74] Jaimeen Ahn and Alice Oh. Mitigating Language-Dependent Ethnic Bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.42.

[75] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.154.

[76] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.416.

[77] Masahiro Kaneko and Danushka Bollegala. Unmasking the Mask – Evaluating Social Biases in Masked Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11954–11962, Jun. 2022. doi:10.1609/aaai.v36i11.21453.

[78] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online, August 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.151.

[79] Percy Liang et al. Holistic Evaluation of Language Models. *Trans. Mach. Learn. Res.*, 2023, 2023. URL `https://openreview.net/forum?id=iO4LZibEqW`.

[80] Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.acl-long.84.

[81] Debora Nozza, Federico Bianchi, and Dirk Hovy. HONEST: Measuring Hurtful Sentence Completion in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, June 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.191.

[82] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, et al. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi:10.1145/3442188.3445924.

[83] Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. Sociodemographic Bias in Language Models: A Survey and Forward Path. In

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 295–322, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.gebnlp-1.19.

[84] Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. Gender Bias in Natural Language Processing and Computer Vision: A Comparative Survey. *ACM Comput. Surv.*, 57(6), February 2025. ISSN 0360-0300. doi:10.1145/3700438.

[85] Hila Gonen and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy, August 2019. Association for Computational Linguistics. URL `https://aclanthology.org/W19-3621/`.

[86] Andrea C Tricco, Erin Lillie, Wasifa Zarin, Kelly K O'Brien, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of internal medicine*, 169(7):467–473, 2018. doi:10.7326/M18-0850.

[87] Anil Ramakrishna, Victor R. Martínez, Nikolaos Malandrakis, Karan Singla, et al. Linguistic Analysis of Differences in Portrayal of Movie Characters. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1678. Association for Computational Linguistics, July 2017. doi:10.18653/v1/P17-1153.

[88] Dima Kagan, Thomas Chesney, and Michael Fire. Using Data Science to Understand the Film Industry's Gender Gap. *Palgrave Communications*, 6(1):1–16, 2020.

[89] Muhammad Junaid Haris, Aanchal Upreti, Melih Kurtaran, Filip Ginter, et al. Identifying Gender Bias in Blockbuster Movies through the Lens of Machine Learning. *Humanities and Social Sciences Communications*, 10(1):94, March 2023. ISSN 2662-9992. doi:10.1057/s41599-023-01576-3.

[90] Yixue Wang and Emoke-Ágnes Horvát. Gender Differences in the Global Music Industry: Evidence from MusicBrainz and The Echo Nest. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM*, pages 517–526. AAAI Press, 2019. doi:10.1609/icwsm.v13i01.3249.

[91] Avriel Epps-Darling, Henriette Cramer, and Romain Takeo Bouyer. Artist Gender Representation in Music Streaming. In *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, pages 248–254, 2020.

[92] Mariana O. Silva, Gabriel Oliveira, and Mirella Moro. Data Insights on Gender Representation: Analyzing the Book and Music Industries. In *Anais Estendidos*

do XXXIX Simpósio Brasileiro de Bancos de Dados, pages 338–347. SBC, 2024. doi:10.5753/sbbd_estendido.2024.243743.

[93] Scott E. Friedman, Sonja Schmer-Galunder, Anthony Chen, Robert P. Goldman, et al. Gender Gaps Correlate with Gender Bias in Social Media Word Embeddings. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society*, pages 2587–2593. cognitivesciencesociety.org, 2020.

[94] Eduard Fosch-Villaronga, Adam Poulsen, Roger Andre Søraa, and B. H. M. Custers. A Little Bird Told Me Your Gender: Gender Inferences in Social Media. *Inf. Process. Manag.*, 58(3):102541, 2021. doi:10.1016/j.ipm.2021.102541.

[95] Carole M. Kortenhaus and Jack Demarest. Gender Role Stereotyping in Children's Literature: An Update. *Sex Roles*, 28(3):219–232, February 1993. ISSN 1573-2762. doi:10.1007/BF00299282.

[96] Gabrielle Poeschl, Aurora Silva, and Alain Clémence. Representações da Masculinidade e da Feminilidade e Retratos de Homens e Mulheres na lLiteratura Portuguesa. *PSICOLOGIA*, 18(1):31–46, Jun. 2004. doi:10.17575/rpsicol.v18i1.410.

[97] Regina Dalcastagnè. A personagem do Romance Brasileiro Contemporâneo:: 1990-2004. *Estudos de literatura brasileira contemporânea*, (26):13–71, 2005. URL `https://periodicos.unb.br/index.php/estudos/article/view/9077`.

[98] Lucia Osana Zolin. A Portrait of Contemporary Brazilian Novels by Female Writers – Um Retrato do Romance Brasileiro Contemporâneo de Autoria Feminina. *Revista Artemis*, 31(1):295–322, 2021. ISSN 18078214. URL `https://link.gale.com/apps/doc/A669313704/AONE?u=anon~2d3e8c57&sid=googleScholar&xid=a741ecce`.

[99] Mariana O. Silva, Clarisse Scofield, Luiza de Melo-Gomes, and Mirella M. Moro. Cross-collection Dataset of Public Domain Portuguese-language Works. *Journal of Information and Data Management*, 13(1), Aug. 2022. doi:10.5753/jidm.2022.2349.

[100] Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S Verykios. A Survey on Challenges and Advances in Natural Language Processing with a Focus on Legal Informatics and Low-Resource Languages. *Electronics*, 13(3):648, 2024.

[101] Peter B. Crabb and Dawn Bielawski. The Social Representation of Material Culture and Gender in Children's Books. *Sex Roles*, 30(1):69–79, January 1994. ISSN 1573-2762. doi:10.1007/BF01420740.

[102] Stuart Oskamp, Karen Kaufman, and Lianna Atchison Wolterbeek. Gender Role Portrayals in Preschool Picture Books. *Journal of social behavior and personality*, 11(5):27–39, 1996.

[103] Diane M. Turner-Bowker. Gender Stereotyped Descriptors in Children's Picture Books: Does "Curious Jane" Exist in the Literature? *Sex Roles*, 35(7):461–488, October 1996. ISSN 1573-2762. doi:10.1007/BF01544132.

[104] Angela M. Gooden and Mark A. Gooden. Gender Representation in Notable Children's Picture Books: 1995–1999. *Sex Roles*, 45(1):89–101, July 2001. ISSN 1573-2762. doi:10.1023/A:1013064418674.

[105] David A. Anderson and Mykol Hamilton. Gender Role Stereotyping of Parents in Children's Picture Books: The Invisible Father. *Sex Roles*, 52(3):145–151, February 2005. ISSN 1573-2762. doi:10.1007/s11199-005-1290-8.

[106] Mykol C Hamilton, David Anderson, Michelle Broaddus, and Kate Young. Gender Stereotyping and Under-representation of Female Characters in 200 Popular Children's Picture Books: A Twenty-first-century Update. *Sex roles*, 55:757–765, 2006. doi:10.1007/s11199-006-9128-6.

[107] Janice McCabe, Emily Fairchild, Liz Grauerholz, Bernice A. Pescosolido, et al. Gender in Twentieth-Century Children's Books: Patterns of Disparity in Titles and Central Characters. *Gender & Society*, 25(2):197–226, 2011. doi:10.1177/0891243211398358.

[108] Anna Čermáková and Michaela Mahlberg. Gender Inequality and Female Body Language in Children's Literature. *Digital Scholarship in the Humanities*, 36 (Supplement_2):ii72–ii77, December 2021. ISSN 2055-7671. doi:10.1093/llc/fqaa051.

[109] Anna Čermáková and Michaela Mahlberg. Gendered Body Language in Children's Literature Over Time. *Language and Literature*, 31(1):11–40, 2022. doi:10.1177/09639470211072154.

[110] Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, et al. What We Teach About Race and Gender: Representation in Images and Text of Children's Books∗. *The Quarterly Journal of Economics*, 138(4):2225–2285, 08 2023. ISSN 0033-5533. doi:10.1093/qje/qjad028.

[111] E Poynter. My Girl, Old Girl: Keywords, Collocations and Gender in British Children's Fiction. *Journal of Language and Literature*, 24(1):27–40, May 2024. doi:10.24071/joll.v24i1.7497.

[112] E Hilton Hubbard. A Functional-syntactic Perspective on Gender Stereotyping in Popular Fiction. *Stellenbosch Papers in Linguistics Plus*, 1994(26):221–232, 1994.

[113] Amanda B. Diekman and Sarah K. Murnen. Learning to Be Little Women and Little Men: The Inequitable Gender Equality of Nonsexist Children's Literature. *Sex Roles*, 50(5):373–385, March 2004. ISSN 1573-2762. doi:10.1023/B:SERS.0000018892.26527.ea.

[114] Dunlop Ochieng. Sexism in Language: Do Fiction Writers Assign Agentive and Patient Roles Equally to Male and Female Characters? *Journal of Language and Linguistic Studies*, 8(2):0–47, 2012.

[115] Satoko Suzuki. Nationalism and Gender in the Representation of Non-Japanese Characters' Speech in Contemporary Japanese Novels. *Pragmatics*, 28(2):271–302, 2018.

[116] Eve Kraicer and Andrew Piper. Social Characters: The Hierarchy of Gender in Contemporary English-Language Fiction. *Journal of Cultural Analytics*, 3(2), 2019.

[117] Jonathan Cheng. Fleshing out Models of Gender in English-language Novels (1850–2000). *Journal of Cultural Analytics*, 5(1):11652, 2020. doi:10.22148/001c.11652.

[118] Houliang Chen and Qianwen Xu. Gender Politics and Victorian Literary Representation of the Body: A Distant Reading of the Body in Charles Dickens's Works. *Humanities and Social Sciences Communications*, 11(1):815, June 2024. ISSN 2662-9992. doi:10.1057/s41599-024-03366-x.

[119] Hari Das and Mallika Das. Gender Stereotyping in Contemporary Indian Magazine Fiction. *Asian Studies Review*, 33(1):63–82, 2009. doi:10.1080/10357820802713593.

[120] Li Lucy and David Bamman. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual, June 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.nuse-1.5.

[121] Lori Baker-Sperry and Liz Grauerholz. The Pervasiveness and Persistence of the Feminine Beauty Ideal in Children's Fairy Tales. *Gender & Society*, 17(5):711–726, 2003. doi:10.1177/0891243203255605.

[122] Scott Weingart and Jeana Jorgensen. Computational Analysis of the Body in European Fairy Tales. *Literary and Linguistic Computing*, 28(3):404–416, 05 2012. ISSN 0268-1145. doi:10.1093/llc/fqs015.

[123] Alba Morollon Diaz-Faes, Carla Murteira, and Martin Ruskov. Explicit References to Social Values in Fairy Tales: A Comparison between Three European Cultures. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational*

Linguistics for Uralic Languages, pages 62–75. Association for Computational Linguistics, December 2023. URL https://aclanthology.org/2023.nlp4dh-1.8.

[124] Shlomo Argamon, Charles Cooney, Russell Horton, Mark Olsen, et al. Gender, Race, and Nationality in Black Drama, 1950-2006: Mining Differences in Language Use in Authors and their Characters. *Digit. Humanit. Q.*, 3(2), 2009. URL http://www.digitalhumanities.org/dhq/vol/3/2/000043/000043.html.

[125] Huimin Xu, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. The Cinderella Complex: Word Embeddings Reveal Gender Stereotypes in Movies and Books. *PloS one*, 14(11):e0225385, 2019. doi:10.1371/journal.pone.0225385.

[126] Chengyue Zhang and Ben Wu. Characterizing Gender Stereotypes in Popular Fiction: A Machine Learning Approach. *Online Journal of Communication and Media Technologies*, 13(4):e202349, 2023.

[127] Aditya Narayan Sankaran, Vigneshwaran Shankaran, Sampath Lonka, and Rajesh Sharma. Revisiting the Classics: A Study on Identifying and Rectifying Gender Stereotypes in Rhymes and Poems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14092–14102, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1228.

[128] Roger Clark, Jessica Guilmain, Paul Khalil Saucier, and Jocelyn Tavarez. Two Steps Forward, One Step Back: The Presence of Female Characters and Gender Stereotyping in Award-Winning Picture Books Between the 1930s and the 1960s. *Sex Roles*, 49(9):439–449, November 2003. ISSN 1573-2762. doi:10.1023/A:1025820404277.

[129] Emily Sigalow and Nicole S. Fox. Perpetuating Stereotypes: A Study of Gender, Family, and Religious Life in Jewish Children's Books. *Journal for the Scientific Study of Religion*, 53(2):416–431, 2014. doi:10.1111/jssr.12112.

[130] Lucy Roberts, Karolina Bačová, Tigist Llaudet Sendín, and Marek Urban. Cultural Differences in the Construction of Gender: A Thematic Analysis of Gender Representations in American, Spanish, and Czech Children's Literature. *Human Affairs*, 33(1):34–50, 2023. doi:doi:10.1515/humaff-2022-2030. URL https://doi.org/10.1515/humaff-2022-2030.

[131] Lynda M. Nilges and Albert F. Spencer. The Pictorial Representation of Gender and Physical Activity Level in Caldecott Medal Winning Children's Literature (1940-1999): A Relational Analysis of Physical Culture. *Sport, Education and Society*, 7 (2):135–150, 2002. doi:10.1080/1357332022000018832.

[132] Brendan Luyt, Sheh Lee Seng, and Yung Yong Ng. Gender Representations and Stereotypes in Singaporean Picture Books: 1970 to 2008. *Malaysian Journal of Library and Information Science*, 16(3):49–65, 2011.

[133] Peter B Crabb and Deb L Marciano. Representations of Material Culture and Gender in Award-winning Children's Books: A 20-year Follow-up. *Journal of Research in Childhood Education*, 25(4):390–398, 2011. doi:https://doi.org/10.1080/02568543.2011.605209.

[134] Alexandra Vidyaeva. The Influence of the Author's Background on the Representation of Gender Stereotypes in Soviet Children's Literature. In *Digital Transformation and Global Society*, pages 574–581, Cham, 2019. Springer International Publishing. ISBN 978-3-030-37858-5.

[135] Kennedy Casey, Kylee Novick, and Stella F Lourenco. Sixty Years of Gender Representation in Children's Books: Conditions Associated with Overrepresentation of Male Versus Female Protagonists. *PLoS One*, 16(12):e0260566, 2021. doi:10.1371/journal.pone.0260566.

[136] Katarina Filipović. Gender Representation in Children's Books: Case of an Early Childhood Setting. *Journal of Research in Childhood Education*, 32(3):310–325, 2018. doi:10.1080/02568543.2018.1464086.

[137] Arsenio Jesús Moya-Guijarro and Roberto Martínez Mateo. Challenging Masculine Gender Stereotypes in Children's Picture Books: A Social Semiotic and Multimodal Analysis. *Journal of the Spanish Association of Anglo-American Studies*, 44:164–185, 2022.

[138] Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, et al. Analyzing Gender Bias within Narrative Tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217. Association for Computational Linguistics, November 2020. doi:10.18653/v1/2020.nlpcss-1.23.

[139] David Bamman, Sejal Popat, and Sheng Shen. An Annotated Dataset of Literary Entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2138–2144. Association for Computational Linguistics, 2019. doi:10.18653/v1/n19-1220.

[140] Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, et al. Rule-based Coreference Resolution in German Historic Novels. In *Proceedings of the Fourth*

Workshop on Computational Linguistics for Literature, pages 98–104. The Association for Computer Linguistics, 2015.

[141] David Bamman, Olivia Lewke, and Anya Mansoor. An Annotated Dataset of Coreference in English Literature. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 44–54. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.6/.

[142] Donghong Ji, Jun Gao, Hao Fei, Chong Teng, et al. A Deep Neural Network Model for Speakers' Coreference Resolution in Legal Texts. *Inf. Process. Manag.*, 57(6): 102365, 2020. doi:10.1016/j.ipm.2020.102365.

[143] Eckhard Bick. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press, 2000.

[144] Eckhard Bick. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. *Working with Portuguese corpora*, pages 279–302, 2014.

[145] Lincoln Mullen. *gender: Predict Gender from Names Using Historical Data*, 2021. URL https://github.com/lmullen/gender. R package version 0.6.0.

[146] SL Bem. Bem Sex Role Inventory–Short Form. *Personality and Individual Differences*, 1978.

[147] Arjun M. Kumar, Jasmine Y. Q. Goh, Tiffany H. H. Tan, and Cynthia S. Q. Siew. Gender Stereotypes in Hollywood Movies and Their Evolution over Time: Insights from Network Analysis. *Big Data and Cognitive Computing*, 6(2), 2022. ISSN 2504-2289. doi:10.3390/bdcc6020050.

[148] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining Gender Bias in Languages with Grammatical Gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5275–5283. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1531.

[149] Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. Fairness in Language Models Beyond English: Gaps and Challenges. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2061–2074. Association for Computational Linguistics, 2023. doi:10.18653/V1/2023.FINDINGS-EACL.157.

[150] Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. Gender Bias in Large Language Models across Multiple Languages. *CoRR*, abs/2403.00277, 2024. doi:10.48550/ARXIV.2403.00277.

[151] Daniela Barreiro Claro, Joaquim Santos, Marlo Souza, Renata Vieira, et al. Extração de Informação. In *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 17. BPLN, 2023. ISBN 978-65-00-80693-9. URL https://brasileiraspln.com/livro-pln/1a-edicao/parte8/cap17/cap17.html.

[152] Diana Santos, Eckhard Bick, and Marcin Wlodek. Avaliando Entidades Mencionadas na Colecao ELTeC-por. *Linguamática*, 12(2):29–49, December 2020. ISSN 1647-0818. doi:10.21814/LM.12.2.336.

[153] Eckhard Bick. Extração de Informação sobre Personagens Literários em Português. *Linguamática*, 15(1):31–40, 2023.

[154] Dalia Andrea Rodríguez, Julia Diaz-Escobar, Arnoldo Díaz-Ramírez, and Leonardo Trujillo. Domain-adaptive Pre-training on a BERT Model for the Automatic Detection of Misogynistic Tweets in Spanish. *Soc. Netw. Anal. Min.*, 13(1):126, 2023. doi:10.1007/S13278-023-01128-2.

[155] Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, et al. Named Entity Recognition for Distant Reading in ELTeC. In *CLARIN Annual Conference 2020*, Virtual Event, France, October 2020. URL https://hal.science/hal-03160438.

[156] Eckhard Bick. Functional Aspects in Portuguese NER. In *Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006, Proceedings*, volume 3960 of *Lecture Notes in Computer Science*, pages 80–89. Springer, 2006. doi:10.1007/11751984_9. URL https://doi.org/10.1007/11751984%5F9.

[157] Rafael Dias and Ivandré Paraboni. Cross-domain Author Gender Classification in Brazilian Portuguese. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1227–1234. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.154/.

[158] João Pedro Moreira de Morais and Luiz Henrique de Campos Merschmann. A Cascade Approach for Gender Prediction from Texts in Portuguese Language. In *WebMedia '22: Brazilian Symposium on Multimedia and Web*, pages 142–149. ACM, 2022. doi:10.1145/3539637.3557057.

[159] Haneen Tamim Abd Ali and Dhamyaa A. Nasrawi. Gender Classification Based on Linguistic Analysis: A Review. *AIP Conference Proceedings*, 3220(1):040006, October 2024. ISSN 0094-243X. doi:10.1063/5.0234674.

[160] Sudeshna Das and Jiaul H. Paik. Context-sensitive Gender Inference of Named Entities in Text. *Inf. Process. Manag.*, 58(1):102423, 2021. doi:10.1016/j.ipm.2020.102423.

[161] Sudeshna Das and Jiaul H. Paik. Gender Tagging of Named Entities Using Retrieval-assisted Multi-Context Aggregation: An Unsupervised Approach. *J. Assoc. Inf. Sci. Technol.*, 74(4):461–475, 2023. doi:10.1002/ASI.24735.

[162] Rosana C. B. Rego and Veronica M. L. Silva. Predicting Gender of Brazilian Names Using Deep Learning. *CoRR*, abs/2106.10156, 2021. URL `https://arxiv.org/abs/2106.10156`.

[163] *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition, 2024. ISBN 978-65-00-95750-1. URL `https://brasileiraspln.com/livro-pln/2a-edicao/`.

[164] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8342–8360. Association for Computational Linguistics, 2020. doi:10.18653/V1/2020.ACL-MAIN.740.

[165] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[166] Mariana O. Silva, Clarisse Scofield, and Mirella M. Moro. PPORTAL: Public Domain Portuguese-language Literature Dataset. In *Anais do III Dataset Showcase Workshop*, pages 77–88, Rio de Janeiro, Brazil, 2021. SBC. doi:10.5753/dsw.2021.17416.

[167] Mariana O. Silva, Clarisse Scofield, and Mirella M. Moro. PPORTAL: Public Domain Portuguese-language Literature Dataset, August 2021.

[168] Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer, 2020. doi:10.1007/978-3-030-61377-8_28.

[169] Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*. European Language Resources Association (ELRA), 2018. URL http://www.lrec-conf.org/proceedings/lrec2018/summaries/599.html.

[170] Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. Portuguese Named Entity Recognition using BERT-CRF. *CoRR*, abs/1909.10649, 2019. URL http://arxiv.org/abs/1909.10649.

[171] Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, pages 1986–1991. European Language Resources Association (ELRA), 2006. URL http://www.lrec-conf.org/proceedings/lrec2006/summaries/59.html.

[172] Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. Evaluating Pretraining Strategies for Clinical BERT Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 410–416. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.43.

[173] Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Domain Adaptation: Challenges, Methods, Datasets, and Applications. *IEEE Access*, 11: 6973–7020, 2023. doi:10.1109/ACCESS.2023.3237025.

[174] Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz, et al. The BDCamões Collection of Portuguese Literary Documents: a Research Resource for Digital Humanities and Language Technology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 849–854. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.106/.

[175] Hidelberg Oliveira Albuquerque, Ellen Souza, Carlos Gomes, Matheus Henrique de C. Pinto, et al. Named Entity Recognition: a Survey for the Portuguese Language. *Proces. del Leng. Natural*, 70:171–185, 2023.

[176] Mariana O. Silva and Mirella M. Moro. PPORTAL_ner: An Annotated Corpus of Portuguese Literary Entities, March 2024.

[177] George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May*

26-28, 2004, Lisbon, Portugal. European Language Resources Association, 2004. URL http://www.lrec-conf.org/proceedings/lrec2004/summaries/5.htm.

[178] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEvalNAACL-HLT 2013*, pages 341–350. The Association for Computer Linguistics, 2013. URL https://aclanthology.org/S13-2056/.

[179] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.*, 34(1):50–70, 2022. doi:10.1109/TKDE.2020.2981314.

[180] Diana Santos, Cláudia Freitas, and Eckhard Bick. OBras: A Fully Annotated and Partially Human-revised Corpus of Brazilian Literary Works in Public Domain. *quot; In CorLex 24 de setembro de 2018*, 2018.

[181] Diana Santos. Portuguese Novel Corpus (ELTeC-por): April 2021 Release, April 2021.

[182] Marcos Zampieri and Martin Becker. Colonia: Corpus of Historical Portuguese. *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, 5, 2013.

[183] Diana Santos and Eckhard Bick. Providing Internet Access to Portuguese Corpora: the AC/DC Project. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece*. European Language Resources Association, 2000. URL http://www.lrec-conf.org/proceedings/lrec2000/html/summary/85.htm.

[184] Elizabeth Marcuschi and Amanda Cavalcante de Oliveira Ledo. Representações de Gênero Social em Livros Didáticos de Língua Portuguesa. *Revista Brasileira de Linguística Aplicada*, 15(1):149–178, January 2015. ISSN 1984-6398. doi:10.1590/1984-639820155926.

[185] Chenhe Dong, Yinghui Li, Haifan Gong, et al. A Survey of Natural Language Generation. *ACM Comput. Surv.*, 55(8):173:1–173:38, December 2022. ISSN 0360-0300. doi:10.1145/3554727.

[186] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, et al. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *PNAS*, 115(16):E3635–E3644, April 2018. doi:10.1073/pnas.1720347115.

[187] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[188] Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT‘s Gender Bias. In *GeBNLP*, pages 1–16. ACL, December 2020. URL https://aclanthology.org/2020.gebnlp-1.1/.

[189] Brenda Salenave Santana, Vinicius Woloszyn, and Leandro Krug Wives. Is there Gender Bias and Stereotype in Portuguese Word Embeddings?, October 2018. URL http://arxiv.org/abs/1810.04528. arXiv:1810.04528.

[190] Fernanda Tiemi de S. Taso, Valéria Q. Reis, and Fábio V. Martinez. Sexismo no Brasil: Análise de um Word Embedding por Meio de Testes Baseados em Associação Implícita. In *STIL*, pages 53–62. SBC, September 2023. doi:10.5753/stil.2023.233845.

[191] Flavio Carvalho, Rafael Rodrigues, Gabriel Santos, et al. Avaliação da Versão em Português do LIWC Lexicon 2015 com Análise de Sentimentos em Redes Sociais. In *BRASNAM*, pages 24–34. SBC, 2019. doi:10.5753/brasnam.2019.6545.

[192] Flavio Carvalho, Fabio Paschoal Junior, Eduardo Ogasawara, et al. Evaluation of the Brazilian Portuguese version of linguistic inquiry and word count 2015 (BP-LIWC2015). *Language Resources and Evaluation*, 58(1):203–222, March 2024.

[193] Lucelene Lopes, Magali Duran, Paulo Fernandes, et al. PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model. In *LREC*, pages 6635–6643. ELRA, June 2022. URL https://aclanthology.org/2022.lrec-1.715/.

[194] Mariana Silva, Michele Brandão, and Mirella M. Moro. Gender Bias in Portuguese Literary Texts: A Masked Language Model Approach, August 2025.

[195] Tuhin Chakrabarty, Vishakh Padmakumar, He He, et al. Creative Natural Language Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, EMNLP, pages 34–40. ACL, December 2023. doi:10.18653/v1/2023.emnlp-tutorial.6.

[196] Arwa I. Alhussain and Aqil M. Azmi. Automatic Story Generation: A Survey of Approaches. *ACM Comput. Surv.*, 54(5):103:1–103:38, May 2021. ISSN 0360-0300. doi:10.1145/3453156.

[197] Fernanda Malheiros Assi and Helena de Medeiros Caseli. 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, November 2024. doi:10.5753/stil.2024.245358.

[198] Hugo Gonçalo Oliveira. Automatic Generation of Creative Text in Portuguese: An Overview. *Language Resources and Evaluation*, 58(1):7–41, March 2024. ISSN 1574-0218. doi:10.1007/s10579-023-09646-3.

[199] Mariana Silva, Michele Brandão, and Mirella M. Moro. Rewriting Stories with LLMs: Gender Bias in Generated Portuguese-language Narratives, June 2025.

[200] Hugo Touvron et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*, February 2023. doi:10.48550/arXiv.2302.13971. URL http://arxiv.org/abs/2302.13971.

[201] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint*, 2023. doi:10.48550/arXiv.2310.06825.

[202] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, et al. Is Temperature the Creativity Parameter of Large Language Models? *arXiv preprint*, 2024. doi:10.48550/arXiv.2405.00492. URL https://arxiv.org/abs/2405.00492.

[203] Freda Shi, Mirac Suzgun, Freitag Markus, et al. Language Models are Multilingual Chain-of-Thought Reasoners. *arXiv preprint*, 2022. doi:10.48550/arXiv.2210.03057. URL https://arxiv.org/abs/2210.03057.

[204] Haoyang Huang, Tianyi Tang, Dongdong Zhang, et al. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. *arXiv preprint*, 2023. doi:10.48550/arXiv.2305.07004. URL https://arxiv.org/abs/2305.07004.

[205] Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1331–1354, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.73/.

[206] Tomas Mikolov, Ilya Sutskever, Kai Chen, et al. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://papers.nips.cc/paper%5Ffiles/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[207] Piotr Bojanowski, Edouard Grave, Armand Joulin, et al. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi:10.1162/tacl_a_00051.

[208] Mariana O. Silva, Gabriel Oliveira, and Mirella Moro. Analyzing Character Networks in Portuguese-language Literary Works. In *Anais do XII Brazilian Workshop on Social Network Analysis and Mining*, pages 115–126, Porto Alegre, RS, Brasil, 2023. SBC. doi:10.5753/brasnam.2023.230585.

[209] Mariana O. Silva, Clarisse Scofield, Luiza de Melo-Gomes, Juliana E. Botelho, et al. Brazilian Reading Preferences in Goodreads: Cross-state and Cross-region Analyses. *iSys - Brazilian Journal of Information Systems*, 15(1):25:1–25:20, Dec. 2022. doi:10.5753/ISYS.2022.2411.

[210] Clarisse Scofield, Mariana O. Silva, Luiza de Melo-Gomes, and Mirella M. Moro. Book Genre Classification Based on Reviews of Portuguese-Language Literature. In *15th International Conference Computational Processing of the Portuguese Language - PROPOR*, pages 188–197. Springer, 2022. doi:10.1007/978-3-030-98305-5_18.

[211] Mariana O. Silva, Clarisse Scofield, Gabriel Oliveira, Danilo Seufitelli, et al. Exploring Brazilian Cultural Identity Through Reading Preferences. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 115–126, Porto Alegre, RS, Brasil, 2021. SBC. doi:10.5753/brasnam.2021.16130.

Appendix A

Systematic Mapping Study (SMS)

This appendix presents the full results of a Systematic Mapping Study (SMS) conducted to explore the research landscape on gender bias in literary texts. This study is designed to identify trends, methodologies, and research gaps in an emerging interdisciplinary field that intersects gender studies, literary analysis, and computational linguistics.

The mapping is structured as follows. Section A.1 outlines the mapping protocol adopted for compiling and charting relevant research on gender bias in literary texts. Next, Section A.2 provides an overview of how gender bias is conceptualized and manifested in the included literature. Then, Section A.3 investigates the main approaches and tools used in gender bias measuring. Finally, Section A.4 presents the resulting map of the field, discussing identified gaps and directions for future research.

A.1 Mapping Protocol

This section details the protocol adopted to identify, select, and chart relevant studies on gender bias in literary texts. Following the PRISMA Extension for Scoping Reviews (PRISMA-ScR) guidelines [86], the study ensures a systematic and replicable process. The protocol consists of: defining research questions (Section A.1.1), implementing a systematic search strategy (Section A.1.2), applying inclusion and exclusion criteria (Section A.1.3), screening and selecting studies (Section A.1.4), and charting the data (Section A.1.5).

A.1.1 Research Questions

The first step defines the research questions that guide this mapping study. Our primary goal is to provide a comprehensive overview of gender bias measuring in literary texts by mapping and categorizing the different approaches used for measuring and analysis. This study does not aim to test specific hypotheses; instead, it provides a descriptive map of the research landscape, identifying trends, methodological strategies, and potential gaps for future investigation. Through such a systematic analysis, the goal is to establish a foundation highlighting existing methods, tools, and research gaps, ultimately guiding future investigations. To do so, we address the following Research Questions (**RQs**):

RQ1. *What is the current* ***landscape*** *of research on gender bias in literary texts?*

RQ2. *Which methodologies have been employed to* ***measure*** *and* ***analyze*** *gender bias in literary texts?*

RQ3. *What are the* ***limitations*** *and* ***gaps*** *in existing literature, and what* ***perspectives*** *should future research?*

A.1.2 Search Strategy

We employ a keyword-based search strategy across multiple academic databases to compile relevant studies. We consider three primary search engines: SCOPUS,[1] DBLP,[2] and SciELO.[3] SCOPUS and DBLP are selected for their extensive coverage of

[1]SCOPUS: https://www.scopus.com/
[2]DBLP: https://dblp.org/
[3]SciELO: https://www.scielo.br/

Table A.1: Search engines and query strings used in the search-engine-based strategy.

Engines	String	Total
SCOPUS	gender `AND` (bias* `OR` stereot* `OR` represent*) `AND` (litera* `OR` novel `OR` book `OR` fiction)	306
DBLP	gender & (bias* | stereot* | represent*) & (litera* | novel* | *book* | *fiction*)	28
SciELO	gênero `AND` (ti:(represent* `OR` viés `OR` preconceito* `OR` estereótipo* `OR` caracteriza*)) `AND` (ti:(liter* `OR` livro* `OR` romanc* `OR` personage*))	14

academic publications through multiple disciplines, including computational linguistics, digital humanities, and literary analysis. SciELO is included to provide comprehensive access to research published in Latin America, specifically in Portuguese.

The search strategy involved crafting tailored search strings for each engine, combining keywords related to gender bias, literature, and literary analysis to maximize the capture of relevant studies. Table A.1 presents the search engines, the specific search strings applied, and the total number of studies retrieved. Overall, the search identified 348 studies (SCOPUS: 306, DBLP: 28, SciELO: 14), which resulted in a solid foundation for the subsequent stages of screening and analysis.

A.1.3 Inclusion and Exclusion Criteria

We apply inclusion and exclusion criteria to ensure a focused and rigorous selection of studies. Eligible studies must explicitly address gender bias in literary texts through quantitative or mixed-method analysis. Studies relying solely on qualitative analysis are excluded to maintain the focus on computational and quantitative methodologies, in line with the exploratory scope of this mapping study.[4] Additionally, only studies published in peer-reviewed journals or conference proceedings are included to ensure research quality. Finally, works that investigate diverse cultural contexts are encouraged, with a particular emphasis on the literature in Portuguese to align with our research focus.

Conversely, studies focused on non-literary texts—such as journalistic articles, textbooks, films, or digital media—are excluded, as are works that do not explicitly address gender bias in literary analysis. We further detail our exclusion criteria (ECs) as follows:

EC1. Studies focusing exclusively on non-literary texts or other forms of media;

EC2. Studies lacking a clearly defined methodological approach for analyzing gender bias;

EC3. Studies with only qualitative approaches, as well as opinion pieces or reviews;

EC4. Studies written in languages other than English or Portuguese;

EC5. Studies that have not been published in conference proceedings or journals;

EC6. Studies published outside the designated time frame (1990 to June 30, 2024).

A.1.4 Screening and Selection Process

After compiling studies through our search strategy, the next step is the systematic screening and relevant studies selection. In total, 348 records were identified across all sources (SCOPUS: 306, DBLP: 28, SciELO: 14). To ensure a transparent and reproducible process, we follow the PRISMA-ScR guidelines, illustrated in Figure A.1. The process consists of three steps: duplicate removal, title and abstract screening, and full-text review.

Duplicate Removal. The initial step involves removing duplicate entries from the databases, as some studies may be indexed in multiple sources.

[4] Qualitative-only studies were excluded to maintain the scope of the map, which is focused on the computational methodologies central to this study. It is important to note, however, that while these works were not formally charted as part of this mapping process, several were used to provide essential context and inform the broader discussion in other sections of this PhD dissertation.

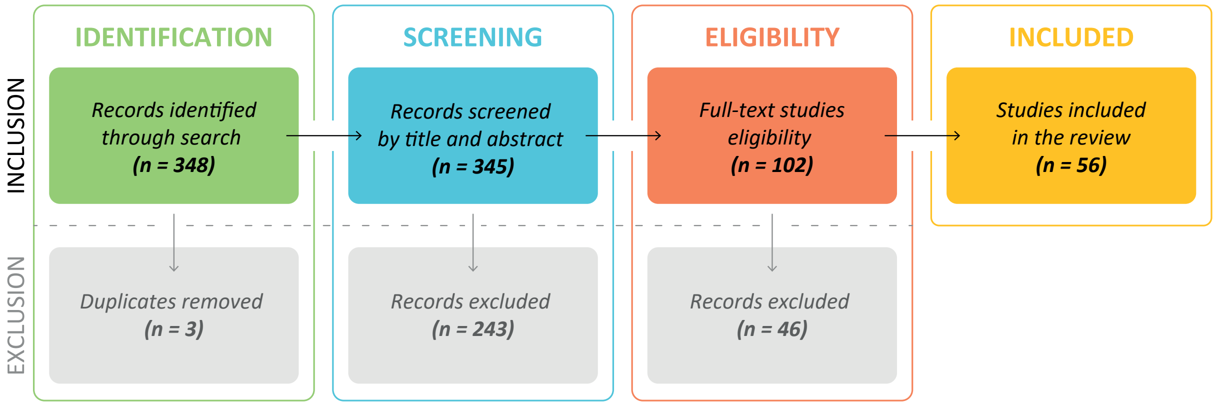


Figure A.1: Screening and selection process based on the PRISMA-ScR guidelines.

After eliminating duplicates, 345 unique studies remain for further analysis.

Title and Abstract Screening. Next, we screen titles and abstracts to assess whether each study aligns with our inclusion criteria (Section A.1.3). During this stage, studies that fall outside the scope of gender bias in literary texts, such as those focused solely on non-literary media or lacking a methodological framework, are excluded. This screening process reduces our selection to 102 studies.

Full-Text Review. Finally, the remaining studies undergo a full-text review to confirm that they meet all inclusion criteria, including methodological rigor, publication quality, and explicit focus on gender bias in the literature. Studies that do not meet these standards, such as opinion pieces or those lacking empirical data, are excluded at this stage. After this full-text review, a final set of 56 studies is retained for in-depth analysis and charting in our map.

Overall, the majority of included studies are published in academic journals (44 studies, 79%), as illustrated in Figure A.2(**A**). This prevalence of journal publications may reflect the methodological rigor and depth required for gender bias research in literary texts. Figure A.2(**B**) further highlights the top journals and conference proceedings represented, showcasing the main venues where gender bias research in the literature is most frequently disseminated. The most common venues are *Sex Roles* (10 studies, 18%) and the *Journal of Cultural Analytics* (4 studies, 7%). While *Sex Roles* primarily focuses on psychology, the *Journal of Cultural Analytics* centers on digital humanities, reflecting the interdisciplinary approaches to analyzing gender bias through literary texts.

A.1.5 Data Charting

For each study, relevant data are systematically extracted and charted based on the defined research questions (Section A.1.1). To address the first one (**RQ1**), we focus on key information, including publication year, the language of the literary works, literary genre, sample size, gender definition (i.e., whether the study adopts a binary or non-binary understanding of gender), type of analysis (quantitative or mixed-method), and reading approach (close or distant). For the second research question (**RQ2**), we extract data related to the methodologies and computational tools employed in each study, as well as the main types of gender bias analyses. Finally, we report the perspectives and future research opportunities for the third research question (**RQ3**). From these charted data, we construct our exploratory map of the field through three complementary analyses:

Descriptive Mapping of Studies (RQ1). This first analysis maps the state of gender bias within the included literary texts, exploring patterns and trends over time (Section A.2).

Analytical Mapping of Measuring Methods (RQ2). We analyze and categorize the various methodologies employed across studies to assess gender bias in the literature (Section A.3).

Perspectives and Research Gaps (RQ3). This section outlines emerging trends and suggests av-

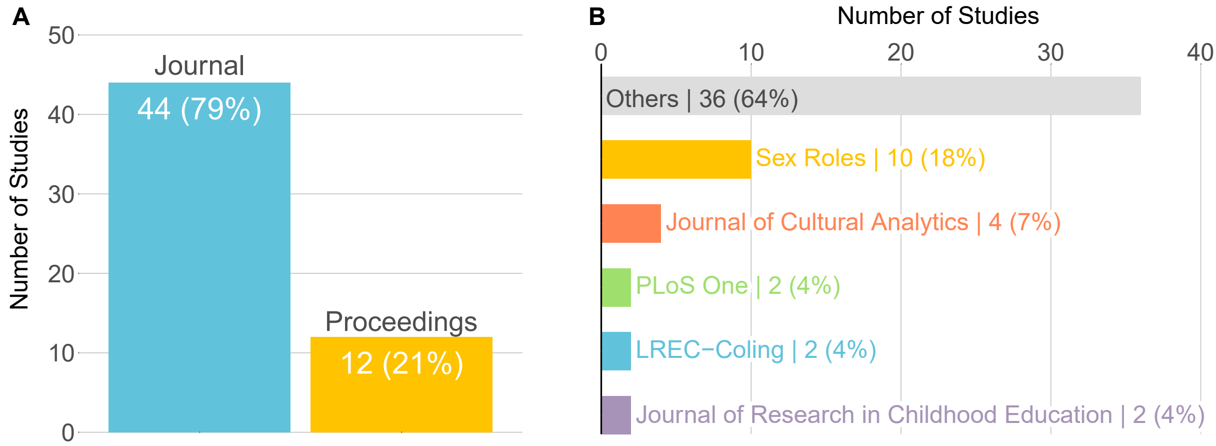


Figure A.2: Publication sources and venues of included studies. **(A)** Distribution of studies by venue type. **(B)** Top publication sources.

enues for future research based on the gaps identified in our map (Section 2.4.1).

A.2 Descriptive Mapping of Studies

Gender bias analysis has gained significant attention across various domains, including film, music, and social media. Regardless of the media, most studies have focused on gender representation through stereotypes, character roles, and other forms of bias. In the film industry, researchers have investigated gender representation in movie scripts, uncovering disparities in dialogue, character traits, and roles between male and female characters [87–89]. Similarly, studies on gender bias in the music industry have highlighted the underrepresentation of female artists [90–92]. On social media platforms, researchers have analyzed how inferential analytics may reinforce gender stereotyping [16, 93, 94].

The literary domain is no exception. Gender bias in literary texts has been increasingly explored since the 1990s, as shown in Table A.2,[5] which summarizes the main characteristics extracted from the mapped studies (as outlined in Section A.1.5). Based on such information, in this section, we discuss trends in gender bias research in literature by investigating five key aspects: temporal trends (Section A.2.1), language (Section A.2.2), literary genres (Section A.2.3), gender definition (Section A.2.4), and types of analysis and reading approaches (Section A.2.5).

Table A.2: Summary of key characteristics in gender bias studies.

Ref.	Year	Lang.	Genre	Sub-Genre	Size	Binary*	Analysis	Reading
[95]	1993	en	Fiction	Children's Literature	150	✓	Quantitative	Close
[112]	1994	en	Fiction	Novels	3	✓	Quantitative	Close
[101]	1994	en	Fiction	Children's Literature	220	✓	Quantitative	Close
[102]	1996	en	Fiction	Children's Literature	22	✓	Quantitative	Close
[103]	1996	en	Fiction	Children's Literature	30	✓	Mixed-Method	Close
[104]	2001	en	Fiction	Children's Literature	83	×	Quantitative	Close
[131]	2002	en	Fiction	Children's Literature	57	✓	Quantitative	Close
[121]	2003	en	Fiction	Fairy Tales	250	✓	Quantitative	Close
[128]	2003	en	Fiction	Children's Literature	84	✓	Mixed-Method	Close
[96]	2004	pt	Fiction	Novels	15	✓	Quantitative	Close

**Binary indicates if the study adopts a binary framework (male/female).*

[5]The complete annotation data is available at: https://tinyurl.com/gender-sms.

Table A.2 continued from previous page

Ref.	Year	Lang.	Genre	Sub-Genre	Size	Binary*	Analysis	Reading
[113]	2004	en	Fiction	Novels	20	✓	Quantitative	Close
[97]	2005	pt-br	Fiction	Novels	258	×	Quantitative	Close
[105]	2005	en	Fiction	Children's Literature	200	✓	Quantitative	Close
[106]	2006	en	Fiction	Children's Literature	200	✓	Quantitative	Close
[124]	2009	en	Drama	Plays	1,200	✓	Quantitative	Distant
[119]	2009	en	Fiction	Short Stories	32	✓	Quantitative	Close
[132]	2011	sg	Fiction	Children's Literature	624	×	Quantitative	Close
[107]	2011	en	Fiction	Children's Literature	5,618	×	Mixed-Method	Close
[133]	2011	en	Fiction	Children's Literature	85	✓	Quantitative	Close
[122]	2012	en	Fiction	Fairy Tales	233	✓	Quantitative	Close
[114]	2012	en	Fiction	Novels, Short Stories	100	✓	Quantitative	Close
[129]	2014	en	Fiction	Children's Literature	30	✓	Quantitative	Close
[46]	2016	en	Fiction	-	600,000	✓	Quantitative	Distant
[18]	2016	en	Fiction	Novels	3,329	✓	Quantitative	Distant
[136]	2018	en	Fiction	Children's Literature	15	✓	Mixed-Method	Close
[115]	2018	ja	Fiction	Novels	31	✓	Quantitative	Close
[23]	2018	en	Fiction	-	104,000	×	Quantitative	Distant
[19]	2019	en	Fiction	-	8,000,000	✓	Quantitative	Distant
[66]	2019	en	Fiction	Novels	3,036	✓	Quantitative	Distant
[116]	2019	en	Fiction	Novels	1,333	✓	Quantitative	Distant
[134]	2019	ru	Fiction	Children's Literature	57	✓	Quantitative	Close
[20]	2019	en	Fiction	-	3,500,000	✓	Mixed-Method	Distant
[125]	2019	en	Fiction | Drama	- | Movie Scripts	7,226	✓	Quantitative	Distant
[138]	2020	en	-	Tropes	15,495	✓	Mixed-Method	Distant
[117]	2020	en	Fiction	Novels	13,000	✓	Quantitative	Distant
[120]	2021	en	Fiction	Short Stories	402	✓	Quantitative	Distant
[108]	2021	en	Fiction	Children's Literature	15	✓	Mixed-Method	Distant
[135]	2021	en	Fiction	Children's Literature	3,280	✓	Mixed-Method	Close
[98]	2021	pt-br	Fiction	Novels	151	✓	Mixed-Method	Close
[12]	2021	en	Fiction	Short Stories	98,161	✓	Quantitative	Distant
[137]	2022	en	Fiction	Children's Literature	5	✓	Mixed-Method	Close
[109]	2022	en	Fiction	Children's Literature	669	✓	Mixed-Method	Distant
[10]	2023	pt-br	Fiction	-	284	✓	Mixed-Method	Distant
[126]	2023	en	Fiction | Drama	- | Movie Scripts	842	✓	Quantitative	Distant
[130]	2023	multi	Fiction	Children's Literature	99	✓	Quantitative	Close
[123]	2023	en	Fiction	Fairy Tales	90	✓	Mixed-Method	Distant
[45]	2023	fr	Fiction	Novels	2,942	×	Quantitative	Distant
[32]	2023	pt	Fiction	-	669	✓	Quantitative	Distant
[24]	2023	en	Fiction	-	3,036	✓	Quantitative	Distant
[11]	2023	en	Fiction	Children's Literature	81	✓	Quantitative	Close
[110]	2023	en	Fiction	Children's Literature	1,895	✓	Quantitative	Distant
[127]	2024	multi	Poetry	Poems	661	✓	Quantitative	Distant
[118]	2024	en	Fiction	Novels	33	✓	Quantitative	Distant
[111]	2024	en	Fiction	Children's Literature	41	✓	Quantitative	Distant
[14]	2024	en	Fiction	Novels	79	✓	Quantitative	Distant
[15]	2024	en	Fiction	-	87,531	✓	Quantitative	Distant

**Binary indicates if the study adopts a binary framework (male/female).*

A.2.1 Temporal Trends

Over the years, research interest in gender bias within literary studies has grown significantly. Figure A.3(**A**) shows a steady increase in published studies, especially from the early 2000s onward. Such an increase may be attributed to several fac-

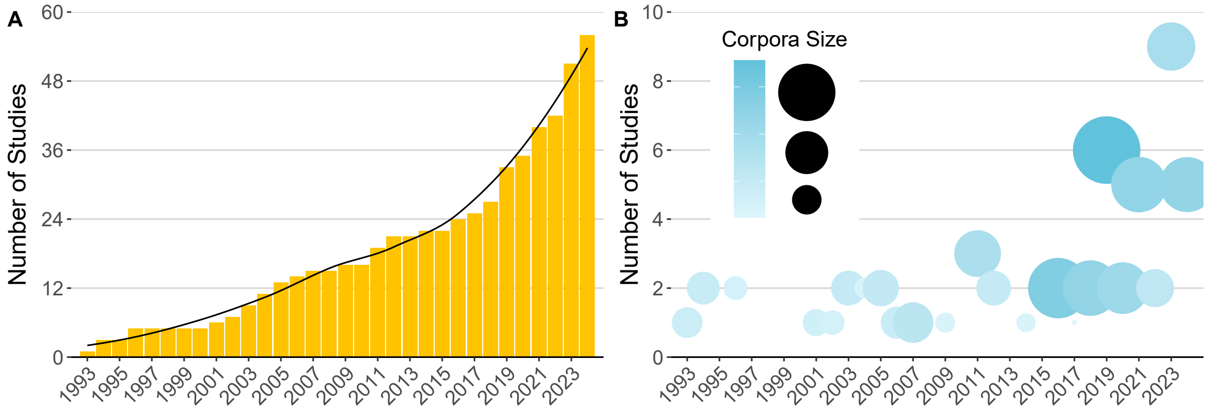


Figure A.3: Temporal trends. (**A**) Distribution of gender bias studies in literature over the years. (**B**) Growth in the average size of corpora analyzed over time.

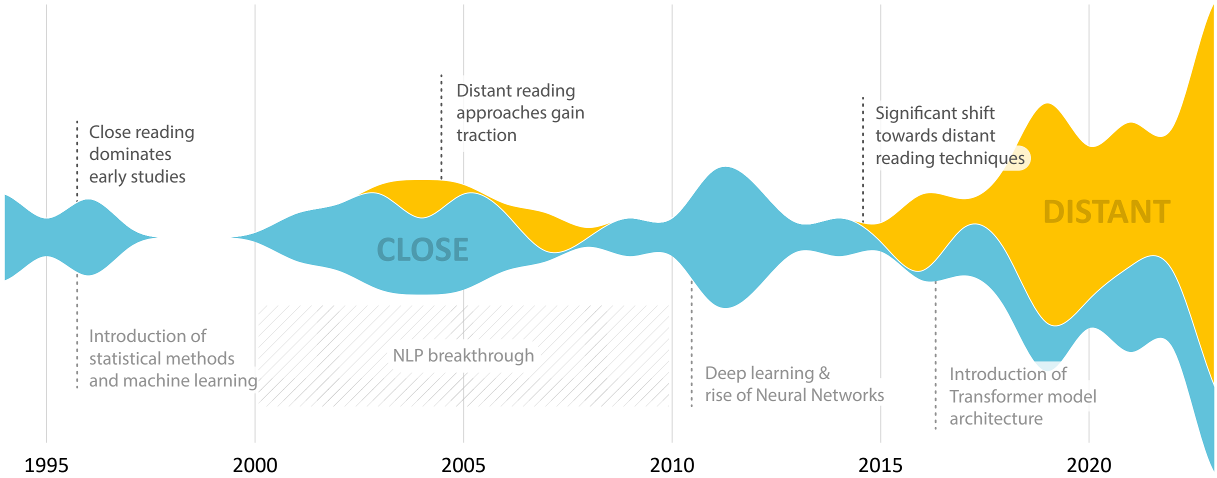


Figure A.4: Evolving distribution of close and distant reading approaches.

tors, including a growing societal awareness of gender issues, the emergence of digital humanities as a field, and the development of advanced computational tools enabling large-scale literary analysis. Moreover, the accessibility of various digital collections [99] has enabled researchers to analyze larger volumes of literary text, further fueling this trend.

The development of gender-focused research has seen notable changes in the scale and scope of data analyzed. Early studies often relied on smaller, manually curated datasets due to the labor-intensive nature of close reading. However, the advent of digital humanities—an interdisciplinary field combining computational methods with traditional humanities scholarship—has facilitated the adoption of automated text analysis tools. This fact enables researchers to analyze larger corpora and generate statistically robust findings. Figure A.3(**B**) highlights this shift, illustrating how the average corpus size has grown significantly over time, particularly in the last decade. This growth reflects technological advancements and a methodological shift toward distant reading approaches.

Figure A.4 illustrates the evolving distribution of close and distant reading approaches in gender bias studies, mapping this shift in parallel with the evolution of NLP techniques [100]. Early research predominantly relied on close reading, a method well-suited to detailed, qualitative analysis of individual texts. However, as NLP tools have become more sophisticated and accessible, distant reading has gained prominence, enabling researchers to uncover broader patterns of gender bias that span multiple works and genres. This methodological shift mirrors the evolution of NLP technologies, which have made it possible to automate the analysis of large text corpora, providing insights that would be difficult to obtain through manual analysis alone.

In summary, these temporal trends in gender bias research in literature reflect the field's growing interest, expanding methodologies, and increasing reliance on large-scale data analysis. This evolution underscores the importance of interdisciplinary approaches that combine traditional literary techniques with advanced computational methods to deepen our understanding of gender representation through diverse literary landscapes.

A.2.2 Language

Although previous research has covered broader gender dynamics over various languages and cultures, most studies have primarily focused on English-written works. Of the 56 works listed in Table A.2, 80% (45 studies) address texts written in English, reflecting a significant gap in the study of gender bias in literature from non-English-speaking cultures. Such a focus on literature in English is understandable, given the widespread use of this language in academic research and the availability of extensive resources. However, it overlooks the rich literary traditions and unique cultural contexts of other languages, such as Portuguese.

With its deep historical roots and contemporary expressions, Portuguese-language literature presents a unique and understudied perspective for analyzing gender bias [32, 50]. Works by authors from Portugal, Brazil, and other Lusophone countries offer a distinct perspective shaped by different social, cultural, and historical influences compared to their English-speaking counterparts. Exploring gender bias in Portuguese-language literary texts can reveal how gender norms and stereotypes manifest in different linguistic and cultural contexts, providing a more comprehensive understanding of global gender dynamics.

Research on gender bias in Portuguese-language literature has gained traction, but it remains somewhat limited (5 studies, 9%). Studies have highlighted persistent gender disparities in the representation of male and female characters, similar to findings in English literature. However, the specific ways these biases are articulated and the cultural factors that influence them can differ significantly [16]. For example, portraying gender roles in Portuguese literature may reflect the social and historical developments unique to Portuguese-speaking countries, such as colonial histories, political movements, and cultural traditions [212–214].

Poeschl et al. [96] analyzed how men and women are described in literary works designated as mandatory reading in Portuguese secondary education (from 1991 to 1997). Their results revealed that words about appearance, beauty, sweetness, suffering, and gift are more frequently associated with females, while words related to strength, passion, work context, and social life are more often associated with males. The authors concluded that the few transformations that occurred in the images presented might be the indispensable concessions to adjust the old portraits to the social realities of the time, but they may also reflect the difficulty of women in defining their identity in a way that is positive and distinct from masculine identity.

Dalcastagnè [97] explored the presence of characters from different social groups in Brazilian novels published between 1990 and 2004 by the publishers *Companhia das Letras*, *Record*, and *Rocco*. The study revealed that the majority of authors are male (72.7%) and white (93.9%), with a large majority having higher education (78.8%). Additionally, the research identified a predominance of authors with other professional activities besides writing, such as screenwriters, teachers, and journalists. These results raise important questions about representation and diversity within the contemporary Brazilian literary field. The study also highlighted the significant absence of characters from marginalized groups, such as blacks, the poor, women, and homosexuals, pointing to the need for democratization and inclusion in literary production.

Zolin [98] investigated the portrait of contemporary novels written by women in Brazil, highlighting the main thematic and ideological characteristics, with emphasis on feminist subversions. In total, 151 novels written by women and published between 2000 and 2015 were mapped, showing the representation of women in the narratives and their intersections with the socio-cultural context. The analysis of the characters revealed a predominance of female protagonists, representing different age groups and occupations, emphasizing life trajectories and general characterization. Contemporary literature by women challenges gender stereotypes by portraying female characters in situations of displacement and occupying public spaces.

Most of these previous studies have relied on manual analysis due to the limited availability of computational resources tailored to the Portuguese language [97, 98]. The scarcity of dedicated NLP tools and datasets challenges the analysis. Unlike English, which benefits from numerous NLP tools and extensive digital resources, Portuguese has fewer dedicated resources, making it challenging to conduct large-scale computational analyses. This limitation has led to a more labor-intensive approach in the literary context, often involving close readings and manual coding of texts to identify patterns of gender representation. While this method provides detailed insights, it is time-consuming and restricts the scope of studies to smaller datasets.

The growing interest in digital humanities and recent advancements in NLP present promising opportunities to overcome these challenges. By developing NLP tools specifically for Portuguese, researchers can automate the analysis of large corpora, allowing for more comprehensive and nuanced studies of gender bias. For example, Freitas and Martins [10] explores public domain Brazilian literature titles, compiled in a corpus with approximately five million words, semantic and morphosyntactically annotated. The proposed methodology identifies how male and female characters are portrayed in those texts, creating a general view of how women and men are built through language. The study happens on two fronts: first, by observing the predicates used in describing characters and the actions these characters take, comparing the male and female results; and second, by critically analyzing them.

Likewise, Freitas and Santos [32] study the words that describe human characters using extensive Portuguese corpora. They classify the studied words into four types: social (professions, occupations, and social status), emotional, character (personality traits), and physical (appearance). In summary, their results reveal: (*i*) female human characters have a higher chance of being described by their appearance compared to male ones; (*ii*) words associated with *character* and *social* attributes are frequently used to describe male humans, whereas those based on *appearance* and *character* are more common for female ones; and (*iii*) regarding preferred depiction words, differences between both genders become pronounced, then highlighting that specific attributes are often associated with one gender over the other.

Overall, gender bias research in Portuguese-language literature reveals specific challenges and insights about this linguistic context. Although limitations in computational tools have historically constrained the scope of research, recent advancements enable more comprehensive analysis, enhancing our understanding of how gender norms are reinforced or challenged in Portuguese literature. Continued research in this area contributes valuable perspectives to global discussions on gender representation in literary traditions.

A.2.3 Literary Genres

To provide a structured analysis of the mapped studies, we classify the literary genres based on three primary literary forms (*Fiction*, *Drama*, and *Poetry*) and their corresponding sub-genres. Figure A.5 illustrates the distribution of research across these categories, revealing a significant imbalance in scholarly focus. Research on gender bias is overwhelmingly concentrated on *Fiction*, which accounts for 52 of the 56 mapped studies (93%). This focus is likely attributed to the genre's well-defined narrative structures, such as characters, plots, and dialogue, which offer rich material for analyzing representational patterns. In contrast, only three studies (5%) address *Drama*, and a single study (2%) focuses on *Poetry*.

A closer look at sub-genres within *Fiction* reveals further patterns. Within the *Fiction* category, *Children's Literature* is the most studied sub-genre, comprising 23 studies (41%), followed by *Novels* with 13 studies (23%). This reflects sustained scholarly interest in the cultural influence of widely consumed literary forms. Other narra-

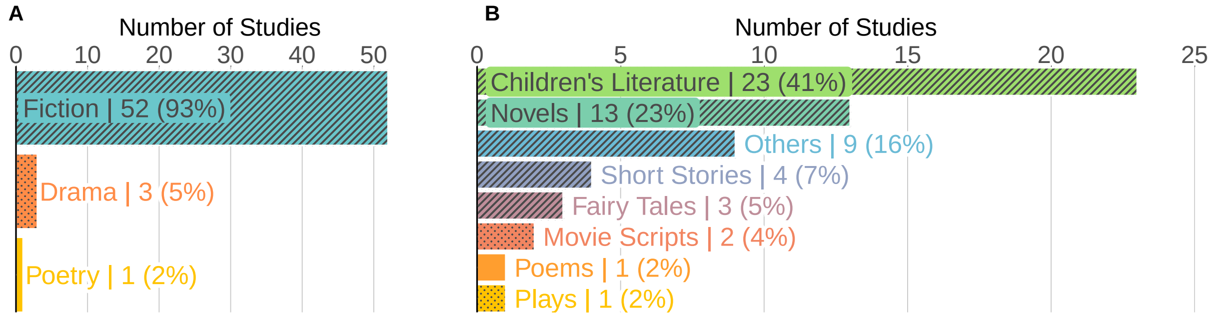


Figure A.5: Distribution of literary genres analyzed in the mapped studies, grouped by **(A)** primary literary genre and **(B)** specific sub-genres.

tive forms, including *Short Stories* (4 studies, 7%) and *Fairy Tales* (3 studies, 5%), also receive attention. A significant share of studies (9 studies, 16%) is grouped under "Others", likely encompassing analyses of general fiction corpora without clear sub-genre attribution.

The emphasis on *Children's Literature* reflects the genre's formative role in shaping gender perceptions from an early age [95, 101–103]. For example, Kortenhaus and Demarest [95] analyzed 150 children's picture books to assess shifts in gender representation. While they observed a more balanced distribution of male and female characters compared to previous decades, stereotypical portrayals persisted: male characters were often independent and active, whereas female characters appeared more passive and dependent.

Similarly, Turner-Bowker [103] and Oskamp et al. [102] analyzed Caldecott award-winning books from different periods. Turner-Bowker [103] investigated the prevalence of characters in titles, pictures, and central roles. Their results showed more males than females in titles and pictures, although there was no significant difference in central roles. Males were described as more powerful, potent, and active than females, often depicted with adjectives emphasizing positive, sweet, and beauty-related traits. On the other hand, Oskamp et al. [102] showed that although significant male-female differences in the depiction of gender-stereotypic traits have diminished, subtle traditional expectations are still present in many of these books.

Such a focus on *Children's Literature* does not stop at the 1990s; it also extends into contemporary times. Over the last years, researchers have continued investigating gender bias in children's books, recognizing the enduring significance of early literary experiences in shaping perceptions of gender roles and identities [13]. Studies have focused on diverse elements, from the prevalence of male and female characters [104–111], nurturing behavior [105, 106], occupations and roles [106, 130]. This ongoing research not only highlights the persistent presence of gender bias in children's literature but also underscores the importance of continued efforts to promote gender equity and diversity in literary representations for young readers [13].

Beyond children's books, gender bias has also been studied in *Novels* [14, 18, 45, 66, 96–98, 112–118], *Short Stories* [12, 114, 119, 120], *Fairy tales* [121–123], and general fiction corpora [10, 15, 19, 20, 23, 24, 32, 46, 125, 126, 138], revealing ongoing disparities in gender representation. For instance, researchers have uncovered that female characters are frequently underrepresented in terms of mentions and visibility compared to their male counterparts [23, 24, 116]. Additionally, gendered language and narrative structures often reinforce traditional stereotypes and expectations, limiting the diversity of female experiences portrayed in literature [20, 24, 66, 114, 120, 122].

Underrepresentation is even more pronounced in both *Drama* and *Poetry* categories. Within *Drama*, only two studies explore modern *Movie Scripts* [125, 126], while a single study addresses traditional *Plays* [124]. *Poetry* is represented by just one study [127], which considers a dataset of rhymes and poems to identify gender stereotypes. The structural and linguistic complexities of these genres may present challenges for computational analysis, contributing to their marginalization in gender bias research. Addressing this methodological gap represents a promising direction for future work.

In summary, the genre-specific distribution of studies highlights both the progress and limitations in understanding gender bias in literature. While *Fiction*, especially *Children's Literature* and *Novels*, has received substantial attention, other literary genres remain largely unexplored. As literary studies move toward more inclusive frameworks, there is a pressing need for intersectional and genre-diverse approaches that better reflect the complexity of gender identity and representation in literary and cultural narratives.

Literary Corpora. The mapped studies consider a wide array of literary corpora, as summarized in Table A.3. However, most studies, particularly those employing close reading methodologies, rely on samples of physical copies of books. Such a traditional approach usually involves selecting specific texts based on curated lists. Most works on children's literature relied on award-winning book lists, including the *Caldecott Award Books* [95, 101–103, 105–107, 128, 131, 133] and *Notable Children's Books* [104]. These lists are often used as benchmarks due to their cultural relevance and influence on shaping young readers' perceptions.

Other close reading works also consider samples or complete collections by specific authors [11, 96–98, 112, 113, 115, 119, 122, 129, 130, 132, 134, 136, 137]. For example, Baker-Sperry and Grauerholz [121] analyzed the complete collection of the Brothers Grimm's fairy tales to investigate the emphasis on "feminine beauty" and its implications for gender dynamics. Similarly, Poeschl et al. [96] explored gender representation in compulsory literary works for secondary education in Portugal (1991 to 1997), highlighting the implicit messages conveyed through selected texts. These targeted approaches allow researchers to delve into specific cultural or historical contexts.

Digital literary corpora are also widely used, offering larger datasets and enabling computational methods for gender bias analysis. For instance, the Project Gutenberg corpus, with over 70,000 texts, provides a multilingual repository spanning various genres, including classics and novels, facilitating large-scale comparative studies [14, 24, 125]. Similarly, corpora such as the Oxford Children's Corpus (OCC) and the CLiC Corpus focus specifically on children's literature, enabling detailed investigations into gender representations in texts intended for young audiences [108, 109]. These digital resources allow researchers to move beyond close reading and apply natural language processing techniques to uncover subtle and systemic patterns of gender bias across vast collections.

Some digital corpora are genre-specific, such as the *Black Drama Corpus*, which focuses on plays by African American authors, and the *Moreever Corpus*, which compiles 90 English fairy tales [123, 124]. Others, such as the *Google Books Ngram Corpus*, aggregate linguistic data from millions of books over centuries, enabling temporal analyses on the evolution of gender representation [19]. These resources provide valuable opportunities to analyze trends over time, often inaccessible through traditional physical sampling.

Although most resources focus on English, as indicated in Table A.3, resources are also dedicated to other languages. For example, the *Chapitres* corpus comprises 2,942 French novels from 1811 to 2020. This corpus has been used to explore gender representation in French literature from 1800 to the present, aiming to assess the prevalence of gender stereotypes in the description of fictional characters [45]. Similarly, Portuguese-language studies have leveraged corpora such as *Literateca* and *Obras* to explore gender dynamics in classic Portuguese literature [10, 32].

A.2.4 Gender Definition

Beyond the languages and literary genres analyzed, another key consideration is how gender is defined within the studies. As shown in Table A.2, most studies adopt a binary framework (51 out of 56, 91%), classifying characters strictly as *male* or *female* (or using equivalent binary terms, such as *girl* and *boy*). Such a binary approach may oversimplify the complexities of gender identity and expression, overlooking the experiences of non-binary, genderqueer, and transgender characters. As a re-

Table A.3: Overview of literary corpora used in the reviewed studies.

Corpus	Lang.	Sub-Genre	#Works	Link	Refs.
Contemporary American Corpus	en	Novels and Short Stories	25M*	[215]	[114]
Oxford Children's Corpus	en	Children's Literature	9M*	[216]	[108, 109]
Google Books Ngram	en	Novels	8M	[217]	[19]
Google Books Syntactic Ngram	en	Fiction	3,5M	[218]	[20]
ROCStories Corpora	en	Short Stories	98,161	[219]	[12]
NovelTM Datasets	en	Novels	87,531	[220]	[15]
Project Gutenberg	multi	Fiction	70,000	[221]	[14, 24, 125]
Gala et al. [138]	en	Fiction	15,495	[222]	[138]
Jockers and Kirilloff [18]	en	Novels	13,000	[223]	[18]
Chicago Corpus	en	Novels	9,000	[224]	[117]
Casey et al. [135]	en	Children's Literature	3,280	[225]	[135]
Chapitres Corpus	fr	Novels	2,942	[226]	[45]
Internet Movie Script Database	en	Movie Scripts	1282	[227]	[125]
Black Drama	en	Plays	1,200	[228]	[124]
Literateca	pt	Fiction	669	[229]	[32]
OBras	pt	Fiction	284	[230]	[10]
CLiC Corpus	en	Children's Literature	146	[231]	[108, 109]
moreever	en	Fairy Tales	90	[232]	[123]

*Million words

sult, the reliance on binary definitions can limit the scope of gender bias analysis.

Despite its limitations, the binary definition is frequently chosen for its simplicity. In terms of data collection and analysis, for example, classifying characters into only two variables can make statistical evaluations more straightforward. Also, the prevalence of binary definitions in gender bias studies may stem from the focus on traditional literature, which has historically reinforced binary gender norms [113]. Many classic texts reflect the societal attitudes toward gender of their times, which often constrained characters to strictly defined roles based on their biological sex [214]. As a result, using a binary framework aligns with the content of these works, albeit at the cost of overlooking the complexity of contemporary gender identities.

Only five of the 56 studies (9%) adopt a more inclusive definition of gender that goes beyond the binary framework. Dalcastagnè [97] and Underwood et al. [23] introduce a simplified yet broader approach by incorporating an additional "other" category to encompass identities that do not fit neatly within the male-female dichotomy. This category allows for some recognition of non-binary, genderqueer, and transgender identities, although it still risks oversimplifying the diverse range of gender experiences. Other studies, such as Gooden and Gooden [104], McCabe et al. [107], and Barré and Dupont [45], adopt a more nuanced perspective by including neutral and plural categories, respectively. By recognizing these additional gender dimensions, these studies offer a more comprehensive framework for understanding gender bias.

However, despite these attempts at inclusivity, most research's emphasis on binary definitions remains a significant limitation. This highlights the need for more diverse and intersectional approaches in gender bias studies. As research continues to evolve, it is crucial to prioritize frameworks that accurately reflect the complexities of gender identity and expression.

A.2.5 Analysis and Reading Approaches

The mapped studies employed a variety of methodological strategies to investigate gender bias in literary texts. These strategies can be broadly categorized along two axes: (i) reading strategy (close vs. distant) and (ii) type of analysis (quantitative vs. mixed-method). This dual-axis classification highlights differences in scale, analytical tools, and interpretive depth.

In this context, *quantitative approaches* rely predominantly on measurable data, often using computational tools and statistical models to identify patterns in large or small corpora. In contrast, *mixed-method approaches* combine quantitative techniques with qualitative analysis, typically interpreting statistical findings through humanistic or contextual lenses.

Regarding reading strategies, *close reading* involves detailed, manual examination of specific texts or excerpts [47]. This traditional literary technique focuses on fine-grained textual features, offering interpretive depth but limited coverage. *Distant reading*, as popularized by Moretti [17], employs computational analysis across large textual datasets, enabling identification of broader trends or systemic patterns that would be difficult to detect manually.

Figure 2.2 presents the distribution of the mapped studies across four quadrants derived from this dual-axis classification. The horizontal axis represents the reading approach, and the vertical axis the type of analysis, resulting in four quadrants: (top left) Quantitative-Close Reading, (top right) Quantitative-Distant Reading, (bottom left) Mixed-Method-Close Reading, and (bottom right) Mixed-Method-Distant Reading. Each one is described as follows.

Quantitative-Close Reading. This quadrant comprises studies that use quantitative techniques, often on a more limited selection of texts, focusing on specific textual features. Figure 2.2 represents the highest number of studies, accounting for 25 (45%). These focused literary analyses involve text coding, where researchers annotate texts to identify and quantify patterns related to gender representation. Typically, one or more coders analyze the texts to ensure accuracy and consistency in identifying relevant features. Regarding quantitative techniques, many studies employed raw score tabulation [95, 97, 98, 102, 104, 105, 107, 112, 114, 115, 121, 122, 128–130] or statistical methods [11, 101, 106, 113, 119, 131–135] to quantify stylistic differences between male and female authors or characters. This approach provides a more nuanced understanding of gender bias that may remain hidden in solely qualitative analyses. However, the reliance on manual coding limits the scalability of studies within this quadrant, often confining them to smaller corpora or specific textual samples.

Quantitative-Distant Reading. This quadrant contains large-scale computational studies that explore gender representation across extensive literary corpora, combining quantitative methods with distant reading. With advances in NLP technologies, this approach has gained significant traction, representing the second-largest category, with 21 studies (38%). Techniques such as text mining [19, 46, 66, 96, 111, 118, 124], NLP [12, 14, 15, 18, 23, 32, 45, 110, 116, 117, 120, 125–127], machine learning [18, 45, 124, 127], and statistical modeling [12, 14, 15, 19, 23, 46, 66, 96, 110, 116–118, 120, 125, 126] are used, often in combination, to uncover patterns of gender representation at scale. It is important to note that these methods are interrelated: text mining often incorporates NLP components (e.g., named entity recognition, co-reference resolution), and machine learning models typically rely on features derived from both text mining and NLP pipelines. These techniques allow researchers to uncover patterns in gender representation that would be difficult to detect through close reading alone. By processing large text collections, this approach provides a broad understanding of gender roles and stereotypes in literary history. However, while computational analysis reveals overarching patterns, it may miss context-specific nuances better captured through close reading and qualitative approaches.

Mixed-Method-Close Reading. This quadrant includes studies that combine quantitative and qualitative methods within a close reading framework. Representing a smaller portion of the research, with three studies (5%), this approach involves analyzing small sets of texts in detail, often using coding frameworks to identify and quantify gendered language, roles, or character traits. Studies in this category may start with quantitative coding to decide the frequency of specific gendered terms or roles, then move to qualitative interpretation, providing context and meaning behind the statistics [103, 136, 137]. This mixed-method approach allows researchers to gain both a numerical overview and an understanding of how gender dynamics unfold in particular works, although it remains time-intensive and typically limited to smaller corpora.

Mixed-Method-Distant Reading. This quadrant represents a common methodological approach in Digital Humanities, applying computational techniques across large corpora while incorporating qualitative interpretation of key results. Representing seven studies (13%), these studies often leverage digital tools to reveal gender patterns across extensive collections while incorporating qualitative insights to contextualize and interpret findings. Typically, researchers employ NLP [10, 20, 24, 123, 138], text mining [10, 108, 109], statistical modeling [20, 24], or machine learning methods [138] to identify and analyze gender bias within literary texts. These quantitative findings are then analyzed through selected qualitative readings of specific passages or character analyses, which help to validate computational results. This mixed-method approach offers the advantages of scale and interpretive depth, allowing researchers to capture both macro-level trends and micro-level contexts.

In summary, studies on gender bias in literary texts span various methodologies that reflect varying balances of depth and breadth in their approach. Close reading, whether quantitative or mixed-method, allows for detailed textual insights but often limits analysis to smaller samples due to its time-intensive nature. In contrast, distant reading expands the analytical scope through computational tools, uncovering broad patterns across large datasets but sometimes missing text-specific subtleties. The choice of approach ultimately depends on the research goals–close reading is ideal for projects that require in-depth, nuanced analysis, while distant reading is better suited for identifying systemic trends across extensive literary corpora.

A.3 Analytical Mapping of Measuring Methods

Measuring gender bias in literary texts is a multifaceted task that spans literary interpretation, linguistic analysis, and computational modeling. Despite the methodological variability across prior studies, it is possible to identify a coherent set of analytical stages that underpin most approaches. Based on a systematic mapping of methodologies found in previous studies, we organize these stages into a unified *framework* comprising four core components (Figure 2.3): (i) character identification (Section A.3.1); (ii) gender classification (Section A.3.2); (iii) dependency analysis (Section A.3.3); and (iv) gender bias measuring (Section A.3.4).

A.3.1 Character Identification

The first step in measuring gender bias in literary texts is to identify characters, the entities whose representation will be analyzed. This task is not merely technical but methodological: the very definition of what constitutes a "character" directly shapes the scope of gender bias analysis. A broad definition that includes any textual reference to an entity (real, fictional, or personified) captures peripheral voices, while a narrower definition focusing only on plot-driving agents highlights the dominant actors in the narrative [50–52].

This conceptual divergence has practical consequences. Adopting a broad definition may reveal systemic gender imbalances in background or minor characters, whereas a restrictive one emphasizes the gendered structure of the core plot. Regardless of scope, character identification involves two subtasks: (*i*) occurrence identification, which detects each mention of characters in the text, and (*ii*) occurrence unification, which links different mentions that refer to the same underlying character [51]. Both subtasks are crucial, as errors at this stage propagate through subsequent analyses of gender classification, dependencies, and bias.

Occurrences Identification. The first subtask involves detecting each instance in which a character is mentioned in the text. Accurate identification is essential, as errors in this stage can spread throughout the analysis, potentially leading to skewed or invalid results. Identifier techniques vary in complexity and scale, ranging from manual annotation to automated methods. Table A.4 presents an overview of the identification methods used in the reviewed studies.

Manual annotation is the most common method, particularly in studies adopting a

Table A.4: Overview of character identification methods in literary texts.

Approach	Description	References
Manual annotation	Human annotators identify character mentions and resolve ambiguities.	[11, 95–98, 101–107, 112, 113, 115, 119, 121, 122, 127–137]
Dataset	Use of pre-annotated corpora containing character mentions.	[10, 20, 45, 114, 124, 125, 138]
NER	Automatically detects and classifies named entities.	[14, 15, 23, 24, 110, 116, 117, 120]
Pronouns	Uses gendered pronouns to detect and trace references to characters.	[18, 46, 108, 109, 118]
Nouns	Extracts mentions via gendered or descriptive nouns (e.g., "woman", "child").	[19, 66, 111]

close reading methodology, where annotators meticulously read the text and mark each occurrence of characters. This approach is well-suited for small datasets or focused studies, ensuring high accuracy and depth of analysis. However, the labor-intensive nature of manual annotation can limit its scalability, making it difficult to apply to larger datasets. As a result, studies relying on manual annotation often focus on specific authors or texts rather than broad genre or period analyses.

Most of these studies involve at least two or more annotators to enhance accuracy and reduce bias [11, 95–97, 101–107, 113, 119, 121, 127–133, 135, 137]. Annotators usually work independently and then compare their results to resolve discrepancies, sometimes involving a third party to reach consensus. This approach helps to improve the reliability of the results and mitigate individual bias. However, some studies rely on a single annotator [98, 112, 115, 122, 134, 136], typically the author of the study, which may introduce subjective bias. Single-annotator approaches are more feasible in studies with limited resources or smaller datasets, though they may compromise the robustness of the results due to the lack of cross-validation.

Another non-automated method involves using pre-existing **datasets** with annotated characters. These datasets, often compiled and validated by previous research or institutions, can significantly streamline the identification process by providing ready-made annotations covering various character data aspects [114]. For example, Argamon et al. [124] used the Black Drama Database,[6] which includes detailed character information such as name, race, age, gender, nationality, ethnicity, occupation, sexual orientation, performers, whether the character is a real person, and type.

Similarly, Hoyle et al. [20] used a dataset of syntactic n-grams to identify and analyze character mentions [233]. Xu et al. [125] employed the Project Gutenberg dataset[7] from the NLTK package,[8] which includes various classic literary texts with rich character data. Gala et al. [138] leveraged character lists from Goodreads,[9] a popular platform for book recommendations and reviews, to extract and analyze character information. Furthermore, Barré and Dupont [45] used the Chapitres corpus,[10] that provides annotated literary texts and character references to facilitate detailed character analysis. Finally, Freitas and Martins [10] used the corpus OBras,[11] which comprises a collection of works of Brazilian literature in the public domain annotated during the AC/DC project [183].

Such a data-based approach offers a practical alternative to labor-intensive manual annotation processes, particularly for studies requiring large-scale analyses. However, these resources are often scarce or may not cover the specific needs of all re-

[6] Black Drama Database: http://solomon.bld2.alexanderstreet.com/
[7] Project Gutenberg: https://www.gutenberg.org/
[8] NLTK: https://www.nltk.org/
[9] Goodreads: https://www.goodreads.com/
[10] Chapitres corpus: https://chapitres.hypotheses.org/
[11] OBras Corpus: https://www.linguateca.pt/OBRAS/OBRAS.html

search contexts. For example, pre-existing datasets might be limited to specific languages, genres, or types of literary works. Additionally, the quality and granularity of annotations can vary, and datasets may not always align perfectly with a given study's research goals or frameworks.

As an alternative, automated approaches have gained traction due to their scalability and efficiency. Such methods usually rely on algorithms or tools to identify characters within texts, which can significantly speed up the annotation process and handle larger datasets. The most common automated approach is to use Named Entity Recognition (NER) models, trained to identify and extract named entities based on linguistic patterns and context. Most studies have employed NER models [14, 15, 23, 24, 110, 116, 117, 120], including the widely used BookNLP tool [139],[12] which specializes in extracting character names and other relevant information from literary texts.

NER models are particularly advantageous for quickly processing large volumes of text, reducing the manual effort required for identification. They leverage pre-trained language models and can be fine-tuned on domain-specific datasets to improve accuracy [170]. However, while NER tools can handle extensive datasets, they may also face challenges with texts containing ambiguous or unconventional references, affecting their precision [139]. In such cases, additional post-processing or manual verification might be necessary to ensure the reliability of the identified entities.

In addition to NER models, other automated approaches use gendered **pronouns** and **nouns** to identify characters. Pronoun-based methods detect references such as "he", "she", "him", and "her", providing gender clues that aid in distinguishing characters and tracking occurrences [18, 46, 108, 109, 118]. The advantage of using pronouns is that they often provide direct clues about the gender and role of characters. However, the reliance on gendered pronouns can also introduce limitations, particularly in texts with ambiguous, fluid, or neutral pronouns or where characters are referred to indirectly.

Noun-based methods also contribute to identifying characters by detecting gendered nouns and titles, such as "man", "woman", "boy", and "girl" [18, 19, 111], which provide immediate context regarding the gender of the characters mentioned. For instance, identifying a character referred to as "grandmother" immediately conveys gender and a characteristic that may inform further analysis. Researchers have also used gendered word lists to analyze gender dynamics within texts [66], categorizing terms based on their gender connotations and enabling a deeper understanding of how gender is portrayed in literature [60].

Occurrences Unification. Once character mentions have been identified, the next subtask is unifying occurrences that refer to the same character. This step, often called coreference resolution [12], is crucial because characters can be mentioned in various ways throughout a text—by name, title, pronoun, or even by a descriptive phrase [51]. For example, a character might be introduced as "John Smith" but later referred to as "he", "the brave knight", or simply "the man". Accurately linking these varied references is essential for coherently understanding the character's role and significance within the narrative.

Notably, this subtask is optional, as not all analyses require a detailed unification of occurrences. In some cases, researchers may focus solely on the initial mentions of characters, especially in studies where the text context is not as complex or where a general overview suffices. Moreover, if the first subtask was performed manually, linking occurrences may not be as critical since the annotators may have already ensured a degree of coherence in their initial identification. However, coreference resolution becomes required for in-depth analyses to understand character dynamics, interactions, or development throughout the story.

The process of occurrence unification can range from manual techniques, where annotators link mentions, to automated methods. In automated approaches, coreference resolution models employ a combination of features, including semantic context and syntactic structures, to connect different references to the same character across the

[12] BookNLP: https://github.com/booknlp

Table A.5: Overview of gender classification methods in literary texts.

Approach	Description	References
Manual annotation	Human annotators assign gender labels based on contextual cues.	[11, 45, 95–98, 101–107, 112–115, 119, 121, 122, 127–137]
Pre-annotated datasets	Use of datasets with pre-defined gender labels.	[124]
Rule-based inference	Use of proper names, gendered pronouns, titles, or lexical cues to classify gender.	[12, 14, 18–20, 46, 66, 108–111, 118, 120, 123, 126, 138]
Computational tools	NLP tools for automatic gender inference using syntactic/semantic patterns.	[10, 15, 24, 32, 117, 125]

text [140–142]. These models are frequently built on advanced natural language processing frameworks [12, 117, 120], which provide scalable solutions for managing large datasets efficiently. Ultimately, effective occurrence unification enhances the robustness of analyses by ensuring that character references are accurately represented throughout the text.

A.3.2 Gender Classification

Once characters have been identified and unified, the next step is to classify their gender. This process is essential for analyzing gender dynamics within literary texts because, without a clear understanding of the gender associated with characters, it becomes challenging to assess how gender biases are represented or perpetuated in the narrative. Gender classification methods vary in scope and complexity, ranging from manual annotation to computational tools and rule-based inference. Table A.5 presents an overview of the methods used for gender classification in the mapped studies, grouped to distinguish human annotation, rule-based inference, computational tools, and annotated datasets.

Manual Annotation. Manual annotation is one of the most prevalent methods for gender classification in literary texts, typically employed after the manual identification of characters. In this approach, annotators carefully investigate contextual clues (e.g., gendered pronouns, names, and titles) to assign gender labels to each character [98, 112, 114, 115, 122, 134, 136]. This method can be valuable for handling complex representations of gender, including implied gender and non-binary identities, as it relies on human interpretation of nuanced contextual clues that automated methods might overlook. However, this method is highly labor-intensive and time-consuming, which limits its scalability to large datasets or extensive literary collections.

Although manual annotation is often more accurate, it is susceptible to individual biases, as annotators may impose subjective interpretations of gender roles and identities. Multiple annotators are often employed to mitigate these risks, followed by discussions to resolve discrepancies [11, 32, 45, 95–97, 101–107, 113, 119, 121, 127–133, 135, 137]. This collaborative effort helps improve the reliability of the classification and reduces the influence of subjective bias. Nevertheless, relying on human interpretation means that the results can vary depending on the annotators' cultural and historical contexts.

Datasets. Another approach to gender classification involves leveraging pre-existing datasets that contain annotated gender information. These datasets are particularly useful for automating the gender classification process, as they provide a reference point for matching names, pronouns, or other textual elements to gender labels. For instance, Argamon et al. [124] used the Black Drama Database, which comprised detailed character information, including their gender. By employing these datasets, researchers can streamline the process of gender classification, particularly in large-scale studies where manual annotation may be inefficient or prone to error. However, their availability is often limited, and their applicability is restricted to specific domains or languages.

Rule-based Inference. Rule-based inference is another widely used method for gender classification, leveraging predefined linguistic rules to classify gender based on gendered pronouns, nouns, or titles. These methods are particularly effective for large-scale processing, as they can rapidly scan and classify characters based on straightforward gender markers, such as proper names [110], honorifics [14], pronouns [12, 18, 46, 108, 109, 120], or gender-specific terms [19, 20, 66, 111, 123, 126, 138].

Regarding name-based inference, researchers usually rely on databases with gender associations for names, which are matched against the text. For example, Adukia et al. [110] used data from the U.S. Social Security Administration (SSA), which includes gender distribution statistics for names based on recorded births. Inference based on honorifics, on the other hand, classifies gender by identifying titles such as "Mr.", "Mrs.", and "Miss" or their equivalents in other languages, which are strongly associated with gender [14].

Pronoun-based inference leverages gendered pronouns such as "he" and "she" or their counterparts in other languages to classify gender. For instance, Huang et al. [12] focused on pronouns instead of first names to classify gender, believing this method to be more inclusive. They specifically used the pronouns "he/him/his" for males and "she/her" for females. These methods are particularly effective in languages where pronouns carry explicit gender information. However, pronoun-based approaches can struggle in contexts where pronouns are ambiguous or omitted.

Inference relying on gender-specific words identifies terms closely associated with a particular gender, such as "girl", "woman", "boy", "man", and others. Qian [66] and Gala et al. [138] considered a gender-balanced lexicon previously compiled by Zhao et al. [60], which comprises 222 male-female word pairs. While effective for texts where gendered words are explicitly used, this method can struggle with texts that intentionally avoid gender-specific terms or use abstract language. Additionally, it may not fully capture non-binary or fluid gender identities, as such representations often fall outside the binary structure of these lexicons.

Computational Tools. Another approach is using computational tools to classify or infer the gender of characters in literary texts. The most commonly used tool is BookNLP [10, 15, 23, 116], a powerful framework for analyzing narrative texts in English. BookNLP identifies characters and classifies their gender by associating them with pronouns and other contextual linguistic markers. Another example is the PALAVRAS parser [32], a constraint grammar-based parsing system for Portuguese [143, 144]. PALAVRAS applies syntactic rules to detect gendered references based on linguistic structure, making it particularly useful for texts in Portuguese.

Additional tools include the *gender* R package [145], which uses U.S. and North Atlantic census data to infer the gender based on first names and dates of birth using historical datasets [117]. Similarly, Python libraries like NLTK [125] and *Gender_Detector* [24] provide functionality for identifying gendered terms or inferring gender based on names. The computational tools' scalability and ability to process detailed linguistic features make them invaluable for large-scale studies.

A.3.3 Dependency Analysis

Dependency analysis assesses grammatical relationships within sentences, mapping how words are syntactically connected. Although it does not directly encode social relationships between characters, it provides structured insights into how agency, attributes, and experiences are distributed across gendered entities in the text. For example, several studies investigate whether male characters appear more frequently as subjects of active verbs, while female characters are more often positioned as objects or patients of actions [18]. In this sense, dependency analysis functions as a bridge between linguistic form and narrative representation.

This approach is not equally adopted across the mapped studies. It tends to be absent in works focused exclusively on metadata (e.g., authorship) or in those relying on manual annotation, where syntactic patterns are addressed qualitatively [95, 101, 102, 104, 107, 112, 116, 124, 128, 130–134, 136]. When employed, however, dependency parsing enables scalable, fine-grained analyses by systematically exposing structural asymmetries in character portrayal.

Syntactic Parsing. Dependency trees link verbs, subjects, objects, and modifiers, providing a formal basis for measuring who performs which actions and how. Widely used tools include the Stanford Dependency Parser [18], SpaCy [15], PALAVRAS for Portuguese [10, 32], and BookNLP, specifically adapted for literary texts [23, 45, 117, 120]. These parsers enable large-scale mapping of syntactic patterns that may reveal gendered differences in narrative agency.

Semantic Role Labeling (SRL). Going beyond syntax, SRL assigns semantic roles to entities—such as agent (doer), patient (receiver), or experiencer (affected). For instance, Luo et al. [14] applied SRL-BERT to identify whether gendered entities were more often portrayed as agents or patients. This level of analysis highlights systematic portrayals of gendered characters as active versus passive, contributing to the measurement of narrative bias.

Collocation Analysis. A complementary, more lightweight strategy involves extracting collocations of gendered nouns (e.g., "man", "woman") with verbs or adjectives [20, 66, 138]. Such analysis can reveal lexical tendencies, for example, verbs like "lead" or "protect" linked with male characters, versus "nurture" or "submit" with female ones. Relatedly, studies on gendered body part clusters (GBPCs) combine gendered pronouns with body part terms to expose stereotypical portrayals tied to physicality [108, 109].

A.3.4 Gender Bias Assessment

The final step in our analytical mapping concerns how gender bias is evaluated in literary texts. Studies adopt a wide range of approaches, which we organize here into two complementary dimensions: (i) *conceptual categories*—the what is being investigated (e.g., authorship, character traits, representation), and (ii) *quantitative metrics*—the how bias is operationalized and measured.

A.3.4.1 Conceptual Categories

Regarding the conceptual dimension, studies investigate gender bias from multiple perspectives, capturing both explicit and implicit forms of gender bias. Table A.6 summarizes the main findings from these studies, grouped into seven categories discussed as follows. Each category reflects a distinct aspect of how gendered dynamics are constructed and represented in literary texts, ranging from authorial influence to thematic concerns.

A.3.4.2 Quantitative Metrics

To complement the conceptual analyses, the mapped studies employ a variety of quantitative and computational metrics. These measures enable the systematic assessment of gender bias in literary texts, moving beyond purely qualitative interpretation. Across the studies, six main categories of quantitative approaches were identified (Table A.7): descriptive counts, hypothesis tests and traditional statistics, correlation/association, predictive modeling, representation-based methods, and composite or derived scores.

A.4 Overall Considerations

This systematic mapping study provides a structured overview of the research landscape on gender bias in literary texts, highlighting trends, methodological approaches, and open research gaps. By classifying studies according to measuring methods, analytical categories, and research perspectives, this mapping provides a comprehensive overview of the field, highlighting areas of concentration as well as underexplored domains. The findings help answer the Research Questions (**RQs**) posed at the outset of this study.

RQ1. *What is the current **landscape** of research on gender bias in literary texts?* The mapping shows that research on gender bias has diversified over time. Early studies focused on broad metrics such as male-to-female character ratios and differences between male and female authors. More recent work increasingly explores nuanced dimensions, including language use, character traits, inter-character relationships, thematic elements, and temporal shifts. There is also a growing interest in intersectional perspectives, considering social factors such as race and class. Overall, the

Table A.6: Overview of the main analyses of gender bias in literary works.

	Analysis	Description	References
	Authorship	Analyze how the author's gender influences character portrayal and narrative bias.	[14, 15, 23, 24, 32, 46, 66, 102, 103, 106, 116, 124, 134, 135, 138]
Charac. Traits	Personality	Analyze how characters' personality traits reflect gender stereotypes.	[102, 105, 113, 128]
	Appearance	Analyze how physical descriptions reinforce gendered portrayals.	[12, 14, 32, 108, 109, 111, 117–119, 121, 122]
Language	Adjectives	Analyze descriptive adjectives to assess biased character portrayals.	[10, 11, 19, 20, 23, 32, 45, 46, 98, 103, 111, 122, 125]
	Co-occurrence	Analyze word co-occurrence patterns to reveal gendered associations.	[96, 108, 109, 111, 118, 122, 125]
	Embeddings	Use word embeddings to capture semantic gender associations.	[123, 125–127]
	Lexicon	Apply gendered lexicons to measure biased word usage.	[10–12, 32, 46, 66, 120, 123, 138]
	Linguistics	Analyze linguistic features to expose gender bias in style or discourse.	[115]
	Valence	Measure sentiment polarity (positive/negative) linked to genders.	[12, 19, 20, 125]
	Verbs	Analyze action verbs and predicates to reveal gendered roles in agency.	[10, 11, 15, 18, 20, 23, 45, 46, 111, 114, 118, 125, 137]
Relationships	Antagonism	Analyze conflictual interactions to uncover gendered antagonism.	[116]
	Assortativity	Analyze gender-based clustering in character networks.	[116]
	Connectivity	Map social networks focusing on gendered connections.	[15, 116]
	Nurturing	Analyze nurturing roles and supportive behaviors by gender.	[98, 105, 106, 129]
	Sexual/Social	Analyze romantic and social bonds to expose gendered norms.	[97, 98, 119]
Representation	Age	Analyze character ages and stereotypes linked to gender.	[19, 97, 98, 103, 107, 110, 111, 119, 121, 122, 129, 135]
	Education	Analyze disparities in education levels by gender.	[98, 119]
	Gender	Quantify male vs. female characters to assess representation.	[11, 14, 15, 23, 24, 32, 45, 66, 95, 97, 98, 101–107, 110, 113, 116, 119, 120, 124, 128–133, 135, 136]
	Locations	Analyze spatial settings (indoor/outdoor) assigned by gender.	[98, 102, 106, 119, 132]
	Occupations	Analyze professions or activities by gender to measure stereotypes.	[66, 97, 98, 106, 113, 119, 129, 132]
	Roles	Analyze narrative roles assigned by gender (e.g., hero, caregiver).	[12, 15, 95, 97, 98, 101, 103, 105, 106, 112, 113, 119, 128, 131, 133, 136]
	Race	Analyze how race and gender intersect in representation.	[97, 103, 110, 124, 129]
	Religion	Analyze religious affiliations in relation to gender portrayals.	[129]
	Orientation	Analyze sexual orientation and its intersection with gender.	[97]
	Status	Analyze differences in social status by gender.	[97, 105, 113, 119]
	Temporal	Track changes in gender bias over time in literary texts.	[15, 23, 24, 101, 106, 107, 109, 110, 117, 128, 131–133, 135]
Theme	Genre	Analyze gender bias across literary genres.	[32, 135, 138]
	Topics	Analyze topics associated with male- vs. female-centered narratives.	[120, 130, 138]

Table A.7: Quantitative metrics used to measure gender bias in literary texts.

Category	**Description**	**References**
Descriptive Counts	Basic quantification of occurrences without inferential testing, including frequency counts, proportions, and percentages.	[10, 11, 15, 18–20, 24, 32, 45, 46, 66, 95–98, 101–119, 121–123, 128–137]
Hypothesis Tests & Traditional Statistics	Statistical tests to compare distributions or means across groups.	[11, 12, 14, 19, 20, 24, 46, 66, 95, 96, 101–107, 111, 113, 116, 119–121, 128, 131–135]
Correlation/ Association	Measures correlation or association between variables.	[11, 12, 18–20, 45, 46, 116–119, 121–123, 125, 134]
Predictive Modeling	Supervised learning models and performance metrics.	[12, 18, 23, 45, 46, 117, 124, 127]
Representation-based	Embedding- and vector-based metrics capturing semantic bias.	[12, 14, 120, 123, 125, 126]
Composite/ Derived Scores	Bias-specific indicators derived from multiple features.	[14, 15, 19, 45, 66, 138]

research landscape spans traditional literary analysis and computational approaches, covering a wide range of literary genres and periods.

RQ2. *Which methodologies have been employed to* ***measure*** *and* ***analyze*** *gender bias in literary texts?* Studies fall into a spectrum of methodological approaches. Manual annotation and close reading remain prevalent, particularly in small-scale studies, but computational techniques—including named entity recognition, embedding-based models, and predictive algorithms—are increasingly applied. These methods allow researchers to categorize gendered language, character roles, and interactions more systematically. Advances in machine learning, such as contextual embeddings and coreference resolution, have further enabled mapping of subtle patterns of gender representation, even when explicit gender markers are absent.

RQ3. *What are the* ***limitations*** *and* ***gaps*** *in existing literature, and what* ***perspectives*** *should future research?* The mapping reveals several recurring limitations. Most studies continue to rely on binary gender classifications, which do not capture the full spectrum of gender identities. Many detection methods face challenges with linguistic challenges, particularly in texts with archaic language, non-standard forms, or deliberately ambiguous references. Future research should prioritize the development of inclusive frameworks for non-binary and fluid gender identities, expand the size and diversity of literary corpora, and integrate interdisciplinary perspectives. These directions can strengthen both methodological rigor and the comprehensiveness of gender bias mapping in literature.

Limitations. It is important to acknowledge a limitation of this mapping study. The entire selection and categorization process, including both title/abstract screening and full-text review, was conducted by a single annotator. While practical constraints necessitated this approach, best practices recommend at least two independent reviewers to minimize subjectivity and reduce potential selection bias. Therefore, the results presented here should be interpreted with this caveat in mind.

Appendix B

Preprocessing and Excerpt Segmentation

In this appendix, we describe the preprocessing (Section B.1) and excerpt segmentation (Section B.2) intermediate steps of the framework, as they do not represent novel methodological contributions but are essential for the reproducibility of the overall workflow.

B.1 Preprocessing

The preprocessing step standardizes raw text data and prepares it for effective analysis. Our framework expects one or more files exclusively in plain text format; other formats, such as PDF or DOCX, are not supported, as they require additional parsing that could introduce errors and inconsistencies. This step consists of two main sub-steps, text cleaning and sentence segmentation, described as follows.

B.1.1 Text Cleaning

The text cleaning step addresses formatting irregularities and ensures the text is suitable for further processing. This process involves five operations, including removing unnecessary spaces, noisy headers, and special characters, as well as email addresses and website URLs. These operations use regular expressions to standardize the text and ensure cohesion. Additionally, we handle potential encoding issues, ensuring that texts in non-UTF-8 formats are correctly converted to a consistent character set. We further describe each operation involved in this step as follows.

Whitespace Normalization. Inconsistent whitespace can make text difficult to process, especially in cases with multiple spaces, tabs, or newline characters. By using regular expressions, we normalize all whitespace, replacing multiple spaces or newline characters with a single space to create a cohesive and uniform text structure.

Noisy Header Removal. Digital versions of literary texts often include headers or footers referencing the document source, website information, or other unrelated text. We remove these noisy elements by checking for specific phrases commonly found in headers and eliminating unwanted text up to that point.

Special Character Removal. Special characters (except hyphens, punctuation marks, and line breaks) can interfere with linguistic analysis and introduce noise. We remove any unnecessary special characters while retaining essential punctuation for maintaining sentence structure.

Summaries Removal. Summaries embedded in the text may interfere with analysis if they overlap with the main text content. We use regular expressions to identify and remove such sections when they follow consistent patterns.

Email and Website Removal. Email addresses and website URLs are irrelevant to literary analysis but are often found in digital versions of literary texts. To ensure that only the narrative content is retained, we systematically remove all occurrences of email addresses and website URLs.

B.1.2 Sentence Segmentation

Once cleaned, the text is segmented into individual sentences, an important step for entity recognition and excerpt segmentation tasks. Here, we use the spaCy sentence tokenizer, part of the spaCy NLP library.[1] By default, a transition-based dependency parser component performs the spaCy sentence segmentation. The dependency parser jointly learns sentence segmentation and labeled dependency parsing. The parser uses a variant of the non-monotonic transition system proposed by [234], adding a "break" transition to perform the sentence segmentation.

We chose the spaCy sentence tokenizer because it is designed to be language-specific, which is crucial for languages with specific sentence boundary markers or exceptions. Moreover, spaCy allows for customization of the sentence tokenizer, where it is possible to define new rules or adjust the model's behavior for particular use cases. We use the `pt_core_news_lg` model, trained on a diverse corpus of Portuguese-language texts. This substep can be easily adapted to other spaCy models, depending on the language of the text being analyzed and other sentence segmentation tools.

After segmentation, we exclude texts containing fewer than ten sentences. Texts shorter than this threshold are unlikely to provide sufficient narrative context for meaningful gender bias analysis. This filtering step ensures that the framework focuses on literary works with enough depth and content to yield reliable insights. It is important to note that, throughout this framework, the notion of a *sentence* is defined operationally as a unit produced by the spaCy sentence segmentation model, rather than by a purely orthographic criterion (e.g., punctuation such as periods). As a result, sentence boundaries are determined by syntactic and contextual cues learned by the dependency parser, which is particularly important in literary texts containing abbreviations, dialogue markers, or unconventional punctuation.

B.2 Excerpt Segmentation

In this step, we define text excerpts that capture the immediate context surrounding each identified `PERSON` entity to facilitate further analysis. Each excerpt is defined as a sliding window of three sentence units, where the first unit contains the `PERSON` entity mention, and the two following units are the immediately subsequent sentences in the segmented sequence. This approach ensures that each excerpt contains sufficient context to detect patterns in how entities are portrayed while minimizing the risk of working with fragmented or insufficiently contextualized excerpts. Algorithm 1 details the excerpt segmentation process.

[1] spaCy: `https://spacy.io/`

Algorithm 1: Excerpt Segmentation

```
Input: sentences: list of sentences, entities: dictionary of PERSON entities
Output: entities: dictionary of PERSON entities with text excerpts
text ← Join sentences with spaces
ranges ← empty list
foreach sentences_tuple in window(sentences, n = 3) do
    start ← position of sentences_tuple[0] within text
    end ← position of sentences_tuple[−1] within text + length(sentences_tuple[−1])
    append(start, end) to ranges
foreach entity in entities do
    start ← position of entity.person within text
    end ← start + length(entity.person)
    entity.start ← start
    entity.end ← end
    entity.excerpt ← empty string
foreach entity in entities do
    start ← entity.start
    end ← entity.end
    foreach (s, e) in ranges do
        if start ≥ s and end ≤ e then
            entity.excerpt ← text[s : e]
            break
return entities
```

The segmentation process initially groups sentences into windows of three, with each window corresponding to a range of text. The choice of a three-sentence window ensures that the immediate syntactic and discursive context surrounding each PERSON entity is preserved, providing a balanced excerpt for analysis. However, this window size can be adjusted depending on the specific needs of the analysis. For example, a larger window may be more appropriate for texts with complex structures or more intricate narrative styles. This flexibility allows the segmentation process to adapt to different text types and analytical goals.

Appendix C

PPORTAL_ner Corpus

Table C.1: Corpus main statistics.

Title	Author	Lang.	Year	#T	#S	#E
Menina e Moça	Bernardim Ribeiro	pt	1554	5,004	139	106
Os Lusíadas	Luís Vaz de Camões	pt	1572	5,000	118	188
Eurico, o Presbítero	Alexandre Herculano	pt	1844	5,000	147	174
Memórias de um Sargento de Milícias	Manuel Antônio de Almeida	pt-br	1854	5,000	166	212
Amor de Perdição	Camilo Castelo Branco	pt	1861	5,001	204	309
Iracema	José de Alencar	pt-br	1865	5,000	275	335
As Pupilas do Senhor Reitor	Júlio Dinis	pt	1867	5,000	264	199
A Morgadinha dos Canaviais	Júlio Dinis	pt	1868	5,000	245	242
Inocência	Visconde de Taunay	pt-br	1872	5,002	204	156
Helena	Machado de Assis	pt-br	1876	5,000	275	260
O Mandarim	Eça de Queirós	pt	1880	5,000	144	180
Memórias Póstumas de Brás Cubas	Machado de Assis	pt-br	1881	5,002	204	156
O Alienista	Machado de Assis	pt-br	1882	5,001	195	218
Casa de Pensão	Aluísio Azevedo	pt-br	1884	5,000	289	254
O Cortiço	Aluísio Azevedo	pt-br	1890	5,000	195	153
Quincas Borba	Machado de Assis	pt-br	1891	5,000	299	201
Dom Casmurro	Machado de Assis	pt-br	1899	5,003	256	183
Os Sertões	Euclides da Cunha	pt-br	1902	5,000	154	179
Esaú e Jacó	Machado de Assis	pt-br	1904	5,000	249	204
Cartas de Inglaterra	Eça de Queirós	pt	1905	5,000	134	299
Memorial de Aires	Machado de Assis	pt-br	1908	5,001	271	245
A Confissão de Lúcio	Mário de Sá-Carneiro	pt	1913	5,000	244	137
Alves & Companhia	Eça de Queirós	pt	1925	5,000	192	172
Capitães da Areia	Jorge Amado	pt-br	1937	5,001	225	306
Vidas Secas	Graciliano Ramos	pt-br	1938	5,001	345	230
	Total			**125,014**	**5,418**	**5,266**

Lang.: Language — #T: Total of tokens — #S: Total of sentences — #E: Total of entities

This appendix presents key statistics and characteristics of the *PPORTAL_ner* corpus, as well as a detailed description of its annotation process and evaluation. Table C.1 summarizes the main corpus statistics, including metadata about each literary work and the total number of tokens, sentences, and annotated named entities. Together, these elements provide a comprehensive overview of the corpus composition and its suitability for training and evaluating literary named entity recognition models.

Metadata Information. In addition to the textual content, the *PPORTAL_ner* corpus provides metadata for each literary work, including the author's name, title, language, and publication year. As presented in Table C.1, our corpus showcases a diverse array of literary works spanning several centuries and featuring authors from Portuguese and Brazilian literature. The metadata serves as a valuable resource for researchers, allowing them to explore the corpus with a historical and authorship perspective, further enriching the potential applications of our dataset.

Corpus Size. The corpus contains 25 individual literary works selected to offer a broad and diverse perspective on the landscape of Portuguese-written literature. In total, the corpus contains 125,014 tokens, 5,418 sentences, and 5,266 annotated entities.

Table C.2: Existing literary corpora annotated with named entities.

Reference	**Corpus**	**Language**	**Category**	**Size***
[139]	LitBank	English	Literary entities	100 (2,000 tokens each)
[235]	OWTO	English	Literary entities	40 (300 sentences each)
[141]	LitBank	English	Literary coreference	100 (2,000 tokens each)
[236]	RiQuA	English	Literary quotation	11 (full-length)
[237]	SLäNDa	Swedish	Literary quotation	8 (2–10 chapters each)
[181]	ELTeC-por	Portuguese	Literary entities	100 (full-length)
[238]	PDNC	English	Literary quotation	22 (full-length)
[239]	SLäNDa 2.0	Swedish	Literary quotation	19 (full-length)
[176]	***PPORTAL_ner***	Portuguese**	Literary entities	25 (5,000 tokens each)

*Total of literary works considered
**Brazil's and Portugal's

Entity Distribution. The *PPORTAL_ner* corpus exhibits a diverse distribution of annotated entities across five distinct categories: `PERSON`, `LOC`, `GPE`, `ORG`, and `DATE`. Table 4.4 provides a comprehensive breakdown of each entity category's frequency, expressed as a percentage of the total annotated entities, along with illustrative examples that offer a glimpse into the dataset's content.

As previously discussed, one notable characteristic of literary texts is their distinctive distribution of entity categories, which deviates from more news-centric datasets. In our corpus, there is a pronounced emphasis on entities related to individuals (`PER`) and vivid descriptions of places (`LOC`). Both categories collectively account for nearly 90% of all annotated entities in the dataset. This unique emphasis aligns with the thematic focus of Portuguese-language literary works, where character portrayal and immersive settings play a central role in storytelling.

C.1 Existing Literary Corpora

The growth of computational literary research [155] has increased demand for labeled datasets designed for literary NER. However, the availability of such resources remains limited, restricting model development and evaluation [34]. As a result, most works still rely on general-domain corpora that fail to capture the distinctive narrative structures, entity distributions, and stylistic features of literary texts [139].

To fill this gap, Bamman et al. [139] introduced LitBank, a dataset of 100 English literary texts from Project Gutenberg annotated with six ACE 2005 entity types: `PERSON`, `FAC`, `GPE`, `LOC`, `VEH`, and `ORG`. Their results demonstrated that models trained on in-domain literary data achieved substantial performance gains, improving F1-scores from 45.7 to 68.3, confirming the benefits of domain adaptation.

Another relevant dataset is OWTO [235], which focuses exclusively on the `PERSON` entity class. OWTO includes 40 English novels, providing a specialized resource for character name recognition. In addition, several other English-language datasets have addressed related tasks, such as quotation attribution [236, 238, 239] and coreference resolution [141]. While these resources are invaluable for advancing NLP tasks, they do not directly address the critical shortage of annotated data for NER tasks in literary works, particularly for a broader range of entity categories.

The scarcity of resources becomes even more evident in non-English contexts, particularly in Portuguese, despite its extensive literary heritage. Although some Portuguese-language corpora exist [174, 180, 182], few include explicit entity annotations. This absence limits the development of NER models sensitive to the linguistic and stylistic diversity of Portuguese-language literary texts, across both European and Brazilian varieties.

Table C.2 summarizes the main existing literary corpora, detailing their language coverage, entity categories, and size. Among them, only the ELTeC-por corpus [181], part of the European Literary Text

Collection (ELTeC),[1] includes annotated Portuguese-language literary texts. It contains 100 European Portuguese novels published between 1840 and 1920, annotated for six entity types (`PERS`, `ORG`, `LOC`, `EVENT`, `WORK`, and `BRAND`) using the PALAVRAS-NER system [156].

In contrast, *PPORTAL_ner* extends coverage to both Brazilian and European Portuguese, broadening the temporal and cultural scope. Although smaller in size, it encompasses diverse genres and literary movements, thereby capturing stylistic, cultural, and linguistic variation across centuries. This diversity enhances its utility for training and evaluating NER models that generalize across different varieties of literary writing.

C.2 Annotation Process

The annotation process of the *PPORTAL_ner* corpus follows a semi-automatic workflow combining automatic pre-annotation with manual human revision. The objective of this strategy is to balance annotation quality, consistency, and feasibility when working with literary texts, which are known for their linguistic complexity and stylistic variation.

Annotators and Review Process. All annotations were reviewed and corrected by a single human annotator with prior experience in literary text analysis and named entity annotation. Although the involvement of multiple annotators and the computation of inter-annotator agreement are recommended best practices, practical constraints led to the adoption of a single-annotator setup. To mitigate subjectivity, the annotation process followed detailed and consistent guidelines, and all texts were reviewed systematically from beginning to end. This limitation is explicitly acknowledged and discussed in the threats to validity of this dissertation.

Annotated Entity Classes. The corpus includes manual annotation of five named entity classes: `PERSON` (`PER`), `LOCATION` (`LOC`), `GEOPOLITICAL ENTITY` (`GPE`), `ORGANIZATION` (`ORG`), and `DATE`. While all these classes were annotated and evaluated in the NER experiments, the `PERSON` class plays a central role in the broader framework, as it constitutes the foundation for subsequent modules related to character identification, gender inference, and gender bias analysis. This analytical emphasis reflects the narrative nature of literary texts, where characters are the primary carriers of social representation.

Pre-annotation and Tooling. The initial annotation was automatically generated using spaCy pretrained Portuguese NER models. These pre-annotations were not assumed to be correct; instead, they served as a starting point to accelerate the manual annotation process. The Prodigy annotation tool was used to visualize, correct, remove, and add entity spans as needed. During this phase, the annotator systematically reviewed each sentence, correcting span boundaries, fixing misclassified entities, and adding missing entities not detected by the automatic models.

C.2.1 Annotation Principles

The annotation guidelines are designed to provide clear, structured principles that guide the annotator through the correction process, ensuring the reliability and consistency of the labeled data. The following principles are emphasized to help maintain high annotation standards throughout the dataset.

Accuracy. The annotator is instructed to prioritize accuracy at every stage of the annotation process. This means carefully reviewing each entity to ensure it is correctly identified and classified according to its category. Special attention is given to entities with ambiguous or overlapping references, using context from the surrounding text to disambiguate their meanings.

[1] https://distant-reading.net

Consistency. The annotator is encouraged to maintain consistent entity labeling conventions, following the annotation guidelines and label definitions adopted in this corpus, while maintaining compatibility with spaCy's NER annotation scheme where applicable.

Nested Entities. The annotator is directed to recognize and label nested entities as a flat structure in which entity labels cannot be embedded within each other. For example, in the following sentence, the annotator should recognize that "Capitu's parents" is a hierarchical entity consisting of two individuals, Senhor Pádua and Dona Fortunata. To maintain a flat structure, each individual is annotated separately without embedding entity labels within each other:

$\overbrace{\textit{Capitu's parents}}^{\texttt{PERSON}}$, $\overbrace{\textit{Sr. Pádua and Dona Fortunata,}}^{\texttt{PERSON}}$ *were concerned about their children.*

$\overbrace{\textit{Os pais de Capitu}}^{\texttt{PERSON}}$, $\overbrace{\textit{Sr. Pádua e Dona Fortunata,}}^{\texttt{PERSON}}$ *estavam preocupados com seus filhos.*

Distinguishing `GPE` and `LOC`. Clear guidance is provided for distinguishing between Geopolitical Entities (`GPE`) and Locations (`LOC`), with a focus on real-world geographic entities for `GPE` and a broader scope for `LOC`, considering both named and commonly imaginary locations. For instance, while "Lisbon" would be categorized as `GPE`, a reference to "the countryside" in a rural novel would fall under `LOC` due to its more generalized and less geographically fixed nature.

C.2.2 Annotation Format

Annotations are stored in JSON format, a widely-used format that ensures compatibility with various NLP tools and systems. Each annotation record contains the document ID (`doc_id`), document text (`doc_text`), and a list of annotated entities (`entities`). Each entity object in the list includes information such as entity ID (`entity_id`), entity text (`text`), entity label (`label`), start offset (`start_offset`), and end offset (`end_offset`). An example is presented as follows:

```
{
  "doc_id": 2550,
  "doc_text": "Se a lembrança de Iracema estivesse nalma do estrangeiro,
  ela não o deixaria partir." (Portuguese)
  "If the memory of Iracema was in the foreigner's mind,
  she wouldn't let him go." (English),
  "entities": [
    {
      "entity_id": 1,
      "text": "Iracema",
      "label": "PESSOA",
      "start_offset": 18,
      "end_offset": 25
    },
    {
      "entity_id": 2,
      "text": "estrangeiro",
      "label": "PESSOA",
      "start_offset": 45,
      "end_offset": 56
    }
  ]
}
```

Table C.3: Main characteristics of the considered pre-trained models.

Model	Pre-training Corpus	Domain	Entities Tags
pt_core_news_sm	WikiNER [240]	news, media	PER, LOC, ORG, MISC
pt_core_news_md	WikiNER [240]	news, media	PER, LOC, ORG, MISC
pt_core_news_lg	WikiNER [240]	news, media	PER, LOC, ORG, MISC
BERT-CRF	HAREM (Selective) [171]	general	PER, LOC, ORG, VALUE, TIME

Although the example above includes an English translation for illustration purposes, only the original Portuguese text is annotated and used in all experiments. The JSON format is not exclusive to spaCy, making it versatile for integrating various NLP tools and platforms. Its structure enhances compatibility, facilitating seamless integration into different research workflows and applications.

C.3 Evaluation

Adopting general-domain language models in specialized domains often leads to suboptimal performance due to significant domain-specific variations in vocabulary, syntax, and context [173]. Therefore, researchers have explored strategies to create domain-specific models by either pre-training from scratch on domain-relevant corpora or using techniques such as continuous domain-adaptive pre-training and fine-tuning on domain-specific data [139, 173]. These approaches adapt existing models to target domains, often achieving more accurate and contextually appropriate results.

In this section, we fine-tune four pre-trained language models to evaluate how our corpus can enhance domain-specific language modeling and NLP tasks. We compare these models' original and fine-tuned versions to assess how domain-specific adaptation enhances their performance in recognizing named entities in a literary context. This comparison allowed us to measure how much domain adaptation, specifically through fine-tuning with our corpus, enhanced each model's ability to recognize entities in this literary context.

C.3.1 Pre-trained Models

Table C.3 outlines the primary characteristics of each considered pre-trained model. In particular, we consider three pre-trained models from spaCy's library. SpaCy offers pre-trained NER models for Portuguese in different sizes, including *large*, *medium*, and *small*. For this study, all three sizes are considered. Note that these spaCy models were initially trained on the WikiNER annotation [240], which does not contain all the entity classes in our corpus. To align the spaCy annotations with our gold standard, we made adjustments, considering `GPE` instances as `LOC` and `DATE` instances as `MISC`.

In addition to spaCy models, we assess a BERT-based model introduced in [170]. Souza et al. [170] proposed the BERT-CRF model that combines a BERT-based embedding model with a Conditional Random Fields layer. Built upon the BERTimbau [168], a Portuguese-tailored BERT-based embedding model, BERT-CRF was initially trained on the HAREM corpus [171]. The HAREM corpus offers two versions, and we consider the "selective" version, featuring five classes: Person, Organization, Location, Value, and Time. In adapting BERT-CRF to our standard, we also merge `GPE` instances with `LOC` and `DATE` instances with `TIME` for consistency across entity labels.

C.3.2 Experimental Setup

The outcome of the annotation process results in a collection of 6,965 annotated sentences originating from 25 distinct literary works. To facilitate the evaluation of the four language models, we employ a

Table C.4: NER models evaluation results on different training data.

Model	Training Data	Precision	Recall	F1 Score
pt_core_news_sm	WikiNER	0.44	0.22	0.29
	PPORTAL_ner	0.67	0.49	0.56
pt_core_news_md	WikiNER	0.49	0.24	0.32
	PPORTAL_ner	0.66	0.52	0.58
pt_core_news_lg	WikiNER	0.47	0.23	0.31
	PPORTAL_ner	0.69	0.60	0.64
BERT-CRF	HAREM	0.79	0.27	0.41
	PPORTAL_ner	0.77	0.77	0.77

sentence-level stratification approach, partitioning the annotated sentences into training, development, and test sets. This stratification involves allocating 80% of the sentences to the training set, equivalent to 5,572 sentences. Furthermore, 10% of the sentences, amounting to 696 sentences, are designated for the validation set, while the remaining 10%, totaling 697 sentences, constitute the test set.

All four pre-trained models are fine-tuned on the NER downstream task using our literary annotated corpus (*PPORTAL_ner*).[2] During fine-tuning, all models are trained for a fixed number of 10 epochs. No extensive hyperparameter search is performed, as the primary objective is to evaluate the impact of domain adaptation rather than to achieve state-of-the-art performance. We employ a narrow hyperparameter search with predefined parameters to evaluate the models fairly. In these fine-tuning sessions, the models are trained until they converge regarding the validation set loss.

C.3.3 Results and Discussion

Table C.4 presents the performance metrics for the evaluated NER models. We evaluate four different models, varying the training data to verify whether fine-tuning on domain-specific corpora significantly enhances the performance of the pre-trained models. The goal is to ascertain whether adapting to a domain could significantly enhance the models' capability to recognize named entities in literary texts.

Overall Performance. Our evaluation of spaCy models indicates significant variations in their performance depending on the training data source. When the spaCy models are exclusively trained on the WikiNER dataset, the performance is suboptimal, with all models achieving an F1-Score of less than 35%. In particular, the small-sized spaCy model exhibited the weakest performance (P: 0.44, R: 0.22, F1: 0.29), whereas the large one exhibited the best performance (P: 0.47, R: 0.23, F1: 0.31).

In contrast, the BERT-CRF model trained on the HAREM dataset achieves a high precision (0.79) but a relatively low recall (0.27) and F1-Score (0.41), albeit higher when compared to spaCy models trained on WikiNER. Such a result has already been previously observed, but in 18th-century medical texts written in Portuguese [241]. The notable difference in results between spaCy and BERT-CRF can be attributed to various factors, such as the model architecture, training data, and entity category harmonization.

The notable low recall prompts further investigation into its underlying causes as it evaluates a model's capability to identify all relevant entities. This may stem from three factors: (i) differences in entity classes across the pre-training data and our corpus, causing misalignment; (ii) the complex language of literary texts, which use varied and creative references for entities; and (iii) literary texts may contain

[2]There is no comparison with the existing corpora for Portuguese-language literature because most of them provide only raw text, without annotations. Regarding the ELTeC-por corpus, a direct comparison with our results is left for future work due to differences in format and annotation classes.

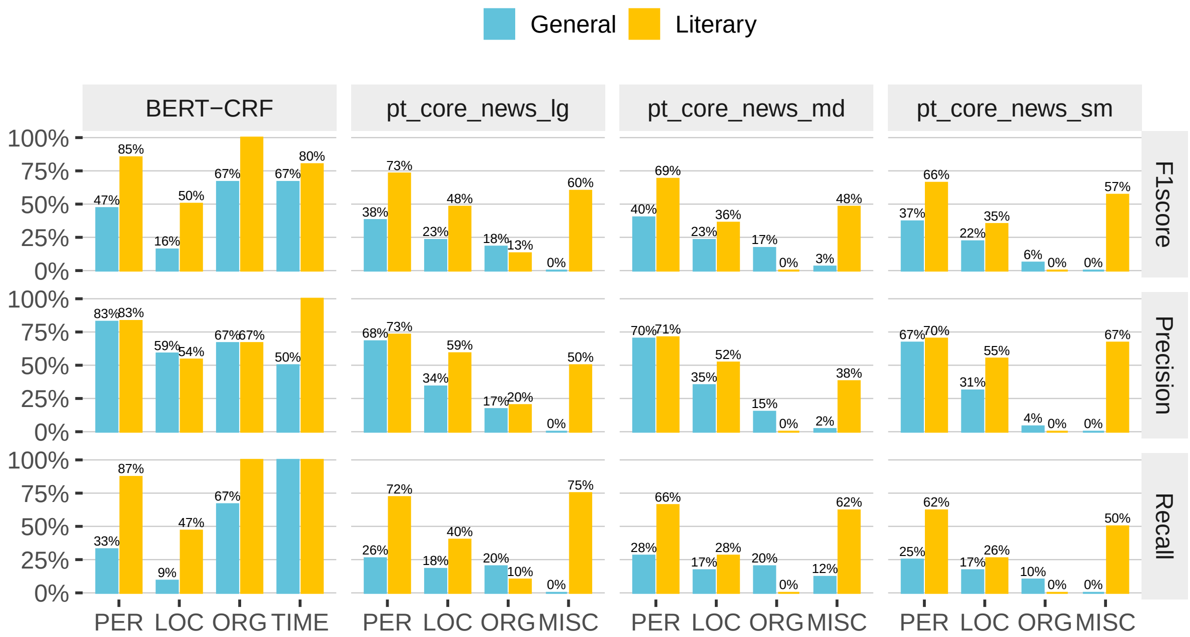


Figure C.1: Evaluation metrics by entity categories and domains.

complex hierarchical entities or multi-word expressions that are challenging to capture accurately (e.g., "Capitu's parents" or "Sr. Págua and Dona Fortunata").

Regarding the fine-tuned models, overall, fine-tuning on our domain-specific *PPORTAL_ner* corpus led to a substantial and consistent improvement in the performance of all evaluated pre-trained models. Domain-specific fine-tuning nearly doubled the F1-Score for all spaCy models, highlighting its significant impact. The BERT-CRF model significantly improved in all three performance metrics, achieving an F1-Score of 0.77 while maintaining high precision and recall.

Entity Analysis. Figure C.1 depicts the results of the evaluation metrics of each model, categorized by entity type and training domain. Regarding recall, which indicates how well the model is capturing all real entities in the data, the results reveal that across different models on the literature domain, entities falling into the categories of `TIME`, `PERSON`, and the harmonized `MISC` class yield the highest recall scores, i.e., these entities are well-captured by the models. Conversely, both `LOC` and `ORG` categories show lower recall rates. The limited success in correctly identifying location and organization entities could be attributed to the creative variations in how those categories are referred to or the potential complexities presented by hierarchical or multi-word expressions.

In contrast to the recall metric, the precision scores offer an alternative perspective on the models' performance in correctly identifying entities. Precision indicates the quality of the model's predictions, measuring how much we can trust the entities the model identifies as correct. For all three spaCy models, the entity categories `PERSON` and `LOC` exhibit relatively higher precision rates across the general and literary domains.

The `MISC` entity category significantly improves when the models are applied to the literary domain, indicating an enhanced ability to correctly identify miscellaneous entities within literary texts. In contrast, the challenge of identifying `ORG` entities persists across both training domains. Such a challenge can be attributed to the intricacies of recognizing organizational names within literary language, where creative expressions and variations in organization names are common. For instance, within Eça de Queirós's work "Cartas de Inglaterra", references to organizations such as *Mollie Maguire*, a 19th-century Irish secret society, and *the Fenians*, a sister organization of the Irish Republican Brotherhood, are frequently encountered. These organizations may feature non-standard names, posing difficulties for the models in correctly identifying them.

In contrast to the spaCy models, the BERT-CRF model stands out for its overall strong performance, particularly excelling in identifying entities within the `ORG` category. This better performance may be attributed to the robustness of the BERT-based model, its capacity to capture context and dependencies in the text, and the advantages of pre-training on the HAREM corpus, which aligns more closely with our gold standard.

Overall, the F1-Score results highlight the unique challenges posed by different entity categories and the impact of domain-specific training. The spaCy models demonstrate significant improvements in F1-Score when applied to the literary domain, most notably for the `MISC` category. This suggests that domain adaptation, even with limited specific-domain training data, can positively impact the models' ability to correctly identify entities in literary texts.

However, the `ORG` category remains a challenge for the spaCy models, with lower F1-Scores across both domains, emphasizing the need for further improvements in identifying these entities in literary contexts. The BERT-CRF model, in turn, maintains a higher F1-Score across most entity categories (except for the `LOC` entity) and both domains, showcasing its superior performance and adaptability to literary NER tasks.

Discussion. Overall, our results underscore the potential for enhancing NER model performance through domain-specific fine-tuning. Improvements were consistent across models, regardless of size or architecture, demonstrating the efficacy of this approach. However, variations among entity categories, particularly the persistent challenges in recognizing `ORG` entities, highlight the complexities of literary language and the need for further research. Future work could focus on refining model architectures or developing targeted pre-training strategies to better capture complex entity expressions.

Appendix D

NER Error Analysis

This appendix analyzes the errors made by the models in recognizing named entities in Portuguese-language literary texts. The goal is to identify major sources of errors and recurring patterns that could inform future improvements. The analysis is organized into three main categories: missing entities (Section D.1), spurious entities (Section D.2), and incorrect or partial entities (Section D.3). Finally, we discuss the implications of these errors and potential mitigation strategies (Section D.4).

Error analysis is important in evaluating NER models, as it highlights specific areas for improvement and sheds light on the models' limitations. As shown in Figure D.1, errors are classified into four primary categories: Missing (M), Spurious (S), Incorrect (I), and Partial (P). Incorrect (I) entities are assessed under both **Strict** and **Type**-based criteria, while Partial (P) entities are assessed exclusively in the **Partial** evaluation scenario, where boundary overlap is allowed.

Missing (M). The rate of missing entities reflects a model's ability to identify all relevant entities in a text. A high rate suggests that models struggle to capture named entity instances, potentially due to insufficient training data or the unique characteristics of entities in literary texts.

Spurious (S). Spurious entities refer to instances where the model incorrectly identifies terms as entities. These errors may arise from linguistic ambiguities or limitations in contextual modeling. For example, in the sentence *"A rosa vermelha floresceu sob o luar"*, the model might incorrectly classify "rosa" as a person entity. A high frequency of spurious entities, especially in precision-critical settings, highlights the need for fine-tuning and more robust preprocessing techniques.

Incorrect (I). Incorrect entities occur when the model identifies an entity but assigns the wrong category or an incorrect span that does not correspond to the gold annotation. For example, in the sentence *"As obras de Shakespeare têm sido amplamente estudadas."*, the model might label "Shakespeare" as an organization rather than a person. Such errors often occur with complex categories such as `ORG` or `TIME`, where expressions may vary significantly.

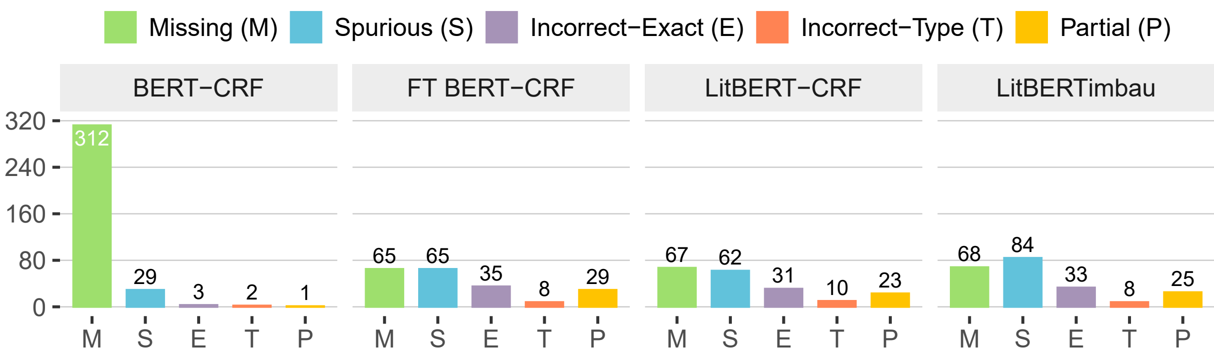


Figure D.1: Distribution of errors in the evaluated models.

Table D.1: Examples of *missing* entities by different models.

Model	Actual	Predicted
LitBERTimbau	*Iam ser da Igreja, como seu dote.*	*Iam ser da Igreja, como seu dote.*
LitBERTimbau	*Nesta época eu era contratado pela Secretaria de Educação de Mato Grosso para dar aulas de teatro naquele estado, [...]*	*Nesta época eu era contratado pela Secretaria de Educação de Mato Grosso para dar aulas de teatro naquele estado, [...]*
BERT-CRF	*Era a hora que toda a Espanha dormia no verão.*	*Era a hora que toda a Espanha dormia no verão.*
BERT-CRF	*Sinhá Vitória acomodou os filhos, que arriaram como trouxas, cobriou-os com molambos.*	*Sinhá Vitória acomodou os filhos, que arriaram como trouxas, cobriou-os com molambos.*
BERT-CRF	*Nossa Senhora, com o Menino Jesus em seus braços, resolveu descer à Terra e visitar um mosteiro.*	*Nossa Senhora, com o Menino Jesus em seus braços, resolveu descer à Terra e visitar um mosteiro.*

ORG — LOC — PER — TIME

Partial (P). Partial entities occur when only part of an entity is correctly identified. This error is common in scenarios with demanding entity boundaries, where segmentation inaccuracies lead to suboptimal results. For example, in *"O grande poeta Fernando Pessoa é conhecido"*, the model might correctly identify "Fernando" as part of a person's name but fail to capture the full entity "Fernando Pessoa".

D.1 Missing Entities

As shown in Figure D.1, the BERT-CRF model exhibits the highest number of missing entities (312), indicating significant difficulties in identifying named entities. In contrast, fine-tuned models substantially reduce this type of error. The high number of missing entities in the baseline BERT-CRF model can be attributed to the lack of adaptation to the specific domain of literary texts, suggesting that incorporating more relevant and domain-specific training data significantly improves the model's performance.

Moreover, as shown in Figure 4.3, based on the recall metric, the fine-tuned models, FT BERT-CRF and LitBERT-CRF, were able to capture all the few existing `ORG` and `TIME` entities in the test set. However, the pure BERT-CRF and LitBERTimbau models struggled to achieve the same performance level. The three fine-tuned models significantly outperformed the baseline in terms of `PERSON` and `LOC` entities. Table D.1 illustrates five examples that highlight these challenges.

In the first two examples, LitBERTimbau fails to correctly recognize the `ORG` entities. In the case of "Igreja", the model incorrectly classifies it as a `PERSON` rather than as an organization. In the second example, LitBERTimbau partially recognizes the entity *"Secretaria de Educação de Mato Grosso"* but incorrectly segments *"Mato"*, only identifying *"Secretaria de Educação"* as an entity. These misclassifications could stem from insufficient examples in the training data for such contexts or from an imbalance in the entity classes during training, where some categories may have been over-represented.

In the third example, BERT-CRF fails to detect the `TIME` entity *"verão"* (summer), suggesting that the generic model lacks the capability to capture specific temporal expressions commonly found in literary texts. The last two examples show that BERT-CRF also fails to recognize the `PERSON` entities *"os filhos"* (the children) and *"o Menino Jesus"* (the baby Jesus), as well as the `LOC` entity *"um mosteiro"* (a monastery). These errors indicate that the generic model faces challenges in identifying named entities that may be referenced implicitly or in less conventional ways in literary texts. For instance, the failure to

Table D.2: Examples of *spurious* entities by different models.

Model	Actual	Predicted
FT BERT-CRF LitBERTimbau	*Seu Tomás fugira também, com a seca, a bolandeira estava parada.*	*Seu Tomás fugira também, com a seca, a bolandeira estava parada.*
LitBERT-CRF LitBERTimbau	*Baleia jantara os pés, a cabeça, os ossos do amigo, e não guardava lembrança disto.*	*Baleia jantara os pés, a cabeça, os ossos do amigo, e não guardava lembrança disto.*
FT BERT-CRF LitBERTimbau	*Mas chegando aos juazeiros, encontrou os meninos adormecidos e não quis acordá-los.*	*Mas chegando aos juazeiros, encontrou os meninos adormecidos e não quis acordá-los.*

ORG — LOC — PER — TIME

detect the entity *"os filhos"* points to a possible limitation in recognizing expressions that do not feature proper names or are referred to more generically.

D.2 Spurious Entities

Regarding spurious entities, i.e., those that were incorrectly identified as entities by the model, the analysis reveals several important insights. Overall, compared to the baseline, the fine-tuned models showed a considerable number of spurious entities, suggesting a tendency to capture terms that should not be classified as named entities (see Figure D.1). Table D.2 presents examples of spurious entities detected by different models.

The first two examples illustrate how the model can make errors by identifying terms that, although contextualized, do not hold significance as named entities. For instance, in the first case, the FT BERT-CRF and LitBERTimbau models incorrectly identified the expression *"a bolandeira"*[1] as an entity, even though it does not refer to a specific organization, location, or person.

However, the last two examples illustrate borderline cases where the distinction between annotation decisions and model errors becomes less clear. In the first case, the models labeled the term "amigo" as a `PERSON` entity. While this follows the corpus annotation guidelines, such cases may appear spurious from a semantic perspective, as the term can function as a common noun depending on context. In the second case, the FT BERT-CRF and LitBERTimbau models correctly classified the term *"juazeiros"* as a `LOC` entity. However, since this term can refer to something more general, like a region or a point of reference, its classification as `LOC` may be interpreted broadly but accurately when in the correct context.

This type of confusion demonstrates how context and word polysemy can impact the accuracy of NER. Ambiguity in the use of language in literary texts is a significant challenge for machine learning models, which often rely on clear and well-defined patterns. Therefore, including more training data that represent the complexity and diversity of literary language can help reduce these errors, improving the model's accuracy and its ability to discern between different uses of words.

D.3 Incorrect and Partial Entities

To conclude the error analysis, we also discuss cases of incorrect entities, considering the **Strict**, **Type**, and **Partial** scenarios. Incorrect entities refer to instances where the model detects an entity but assigns

[1]In the novel *Vidas Secas*, by Graciliano Ramos, the *"bolandeira"* is an image that is part of the imagination of Fabiano, the protagonist of the story. The *"bolandeira"* is known as a "benefiting machine".

Table D.3: Examples of *incorrect* entities by different models.

Model	Actual	Predicted
FT BERT-CRF LitBERT-CRF	*Coitado, morrera* *na areia do rio* *[...]*	*Coitado, morrera na* *areia do rio* *[...]*
FT BERT-CRF LitBERT-CRF LitBERTimbau	*A cachorra Baleia* *foi enroscar-se junto dele.*	*A* *cachorra Baleia* *foi enroscar-se junto dele.*
FT BERT-CRF LitBERT-CRF LitBERTimbau	*Estavam no* *pátio de uma fazenda sem vida*.	*Estavam* *no pátio de uma fazenda* *sem vida.*

ORG — LOC — PER — TIME

an incorrect span and/or an incorrect category. This error category is crucial for understanding the limitations of the models and their generalization capabilities.

In the **Strict** scenario, where precision and recall must be rigorously observed, the rate of incorrect entities tends to be higher. This is because, in this context, entities must match exactly the true references in the text, and even the slightest discrepancy results in an error. For example, a model might correctly identify *"Universidade Federal"* (Federal University) as an organization, but by labeling *"Universidade"* as an entity on its own, it fails to recognize the full entity, leading to incorrect classification.

In the **Type** scenario, where some flexibility is allowed in the overlap of entities, the models may still confuse categories of entities. This is evident when, for instance, an entity like *"São Paulo"* is incorrectly classified as an organization rather than a location, due to the ambiguity that can arise in literary texts that mention cities in specific contexts. These cases highlight the need for more robust and contextualized training, with data representing variations in entity usage across different contexts.

Finally, in the **Partial** scenario, where boundary matching is more flexible regardless of entity type, models tend to show a higher detection rate but may still misclassify entities. In this scenario, a model might recognize a text snippet that contains part of an entity but fails to capture the complete entity or correctly associate it with its category. For example, with the expression *"Instituto Federal de Educação"* (Federal Institute of Education), a model may identify only *"Instituto"* as an entity, ignoring the rest of the expression needed for full identification.

Table D.3 provides examples of partially recognized entities, illustrating how models handle cases where identification is incomplete. The three examples show that most recognition issues are related to models' difficulty in capturing the entirety of an entity in the presented context. The models frequently fail to capture articles and prepositional phrases that are part of the annotated entity span, resulting in fragmented entity identification. These cases underscore the importance of training models with data that includes rich and varied examples of linguistic structures.

D.4 Discussion

Overall, our results revealed that the baseline BERT-CRF model had the highest number of missing entities. This suggests that the lack of adaptation to the specific domain of literary texts is a critical factor that can be mitigated by incorporating more relevant and domain-specific training data. On the other hand, the fine-tuned models (i.e., FT BERT-CRF, LitBERT-CRF, and LitBERTimbau) showed a significant reduction in missing entities, demonstrating the effectiveness of fine-tuning. However, there are still instances where contextually ambiguous words are misclassified, indicating the need for more rigorous training that accounts for the complexity and nuances of literary texts.

Appendix E

Body Parts Dictionary

Table E.1 presents the complete dictionary of the considered body parts and their synonyms. Researchers can refer to this table to gain insights into the specific body parts and their associated synonyms considered in our analysis.

Table E.1: Complete dictionary of considered body parts and synonyms. Terms with an asterisk (*) could not be properly translated into English.

Body part	Synonyms
cabeça (head)	crânio (skull), moleira*, cocuruto*, coco (coconut), testa (forehead), carola*
cabelo (hair)	cabelos (hairs), cabeleira*, cabeleiras*, juba (mane), jubas (manes), mecha (lock), mechas (locks), melena*, madeixa*, madeixas*, cabelame*, tufo (tuft), tufos (tufts)
pêlo (pelage)	penugem (fluff), pelagem (hair), fio (strand), fios (strands), pêlos (bristle), pelame (fur)
cílio (eyelash)	cílios (eyelashes)
cérebro (brain)	encéfalo (encephalon)
testa (forehead)	fronte (front)
fronte (front)	testa (forehead)
sobrancelha (eyebrow)	sobrecenho*, sobrolho*, supercílio*, celha*
pálpebra (eyelid)	pálpebras (eyelids)
face (face)	rosto (countenance), cara*, fronte (front), rostro (rostrum)
olho (eye)	olhos (eyes), vista (view), visão (vision), olhar (look)
pupila (pupil)	pupilas (pupils)
bochecha (cheek)	bochechas (cheeks)
nariz (nose)	narigão (big nose), bicanca*, narina (nostril), narinas (nostrils)
boca (mouth)	lábios (lips), cavidade bucal (buccal cavity)
orelha (ear)	ouvido (ear canal), orelhas (ears), ouvidos (ear canals)
tronco (torso)	
pescoço (neck)	
nuca (nape)	
barriga (belly)	abdômen (abdomen), center (womb), pança (paunch), pandulho*, bandulho*, bucho*, estômago (stomach), bojo*, panturra*
quadril (hip)	quadris (hips), cintura (waist), anca*, ilharga*
membros (members)	braços (arms), pernas (legs), mãos (hands), pés (feet)
ombro (shoulder)	ombros (shoulders), espádua*, espalda*
braço (arm)	braços (arms), ramo (branch)
antebraço (forearm)	antebraços (forearms), punho (fist), punhos (fists)
pulso (pulse)	pulsos (pulses)
mão (hand)	mãos (hands), garra (claw), garras (claws), palmo (palm), pata (paw), patas (paws)
joelho (knee)	joelhos (knees), articulação (articulation), geolho*, rótula (kneecap)
coxa (thigh)	coxas (thighs)
perna (leg)	pernas (legs), cambito*, gambito*, caniço*
pé (foot)	pés (feet), pata (paw), patas (paws), call (callus), calos (calluses)
calcanhar (heel)	calcanhares (heels)
dedo (finger)	dedos (fingers), dedito*
unha (nail)	unhas (nails), casco (hull), cascos (hooves), garra (claw), garras (claws), úngula (ungula)
osso (bone)	ossos (bones)
pele (skin)	cútis (cutis), tez*, espinhas (pimples)
junta (joins)	articulação (articulation), articulações (joint), juntas (joints)
costas (back)	espinha dorsal (backbone), espinha (spine)
bunda (butt)	bundas (butts), traseiro*, traseiros*, rabo (tail), nádegas (buttocks)
cotovelo (elbow)	cotovelos (elbows)
intestino (intestine)	
coração (heart)	
pulmão (lung)	
fígado (liver)	
pâncreas (pancreas)	
rim (rim)	
bexiga (bladder)	
cólon (colon)	
esôfago (esophagus)	
tornozelo (ankle)	tornozelos (ankles)
peito (chest)	peitos (breasts), colo (lap), busto (bust), seio (breast), seios (breasts), torso (torso), tórax (thorax)
coluna (spine)	espinha dorsal (backbone), espinha dorsais (backbones)
corpo (body)	
garganta (throat)	
dente (tooth)	dentes (teeth)
gengiva (gum)	gengivas (gums)
mandíbula (jaw)	mandíbulas (jaws)
língua (tongue)	línguas (tongues)
sorriso (smile)	sorrisos (smiles)

Appendix F

Text Generation Details

This appendix provides additional details on the text generation process, including model execution, inference settings, hardware specifications, and an illustrative example of input-output pairs.

Model Execution and Configuration. We employed two instruction-tuned LLMs for text generation: *Mistral-7B-Instruct* and *LLaMA 3.2-3B*. Both models were executed using *Ollama*, a framework optimized for running large language models locally with efficient resource management. Ollama provides a streamlined inference pipeline, allowing rapid text generation without requiring cloud-based APIs. The models were downloaded as pre-trained checkpoints and loaded into the Ollama framework. The inference process was conducted with the following configuration parameters:

- **Model:** Mistral-7B-Instruct / LLaMA 3.2-3B
- **Temperature:** 0.9 (for creative variation in responses)
- **Top-p:** 0.9 (to balance coherence and diversity)
- **Stop:** `\n` (to ensure single-sentence continuations)
- **Number of generations per input:** 1

Hardware Specifications. The text generation was conducted on a local machine with the following hardware specifications:

- **Processor:** 13th Gen Intel(R) Core(TM) i7-13650HX
- **RAM:** 32GB
- **GPU:** NVIDIA GeForce RTX 4050

Example of Text Generation. To illustrate the model's behavior, Table F.1 presents an example of an input-output pair for one of the processed sentences. This example highlights differences in narrative style between the two models, showcasing their ability to produce coherent and contextually relevant continuations.

Processing Time. The average inference time per sentence was approximately 4 seconds for *Mistral-7B-Instruct* and 2 seconds for *LLaMA 3.2-3B*. The total generation process for the entire dataset was completed within 47 hours.

Additional Results. This appendix provides supplementary results from the lexicon-based analysis, including detailed WEAT effect sizes for all category pairs and datasets. These results offer a more granular view of the gendered word associations in the original and generated texts.

Figure F.1 presents the WEAT effect sizes for all ten category pairs, comparing the original texts with the outputs of Mistral and Llama. The results reveal consistent patterns of gendered associations,

Table F.1: Example of input-output pairs for text generation.

Original sentence: "Capitu foi ao muro, e, com o prego, disfarçadamente, apagou os nossos nomes escritos."

Input Prompt:
Instructions: continue writing in pt-br the story from the following sentence, describing what the character did next. The character is `[MASK]`.
Sentence: "`[MASK]` foi ao muro, e, com o prego, disfarçadamente, apagou os nossos nomes escritos."

Model Output (Mistral-7B-Instruct):
"Depois de apagar sutilmente os nossos nomes do muro com o prego, `[MASK]` segurou a respiração para não desmascará-lo e olhou em volta da sala. Com um sorriso leve, ele voltou ao seu lugar no banco e voltou a ficar absorvido no livro. Aparentemente, ele estava tão envolvido com o romance de sua leitura que ninguém notou a mudança na sala ou percebeu que os nomes dos presentes haviam desaparecido."

Model Output (LLaMA 3.2-3B):
"Enquanto fazia isso, seus olhos se moviam rapidamente para garantir que ninguém estivesse assistindo, e sua mente estava em constante alerta para qualquer sinal de perigo. Com um lastimeiro gesto, apagou os nomes escritos no muro, deixando apenas uma marca escurecida onde havia uma vez as palavras: *Eles não estão mais aqui*. Sua mão esfregou o prego contra a camisa, para garantir que nenhum indício de sua atividade tivesse sido deixado atrás."

with male terms more strongly linked to cognitive processes, work, and risk, while female terms are associated with emotions, home, and health. These findings align with the broader trends discussed in Section 6.3.2.3.2 and highlight the persistence of gender stereotypes in LLM-generated narratives.

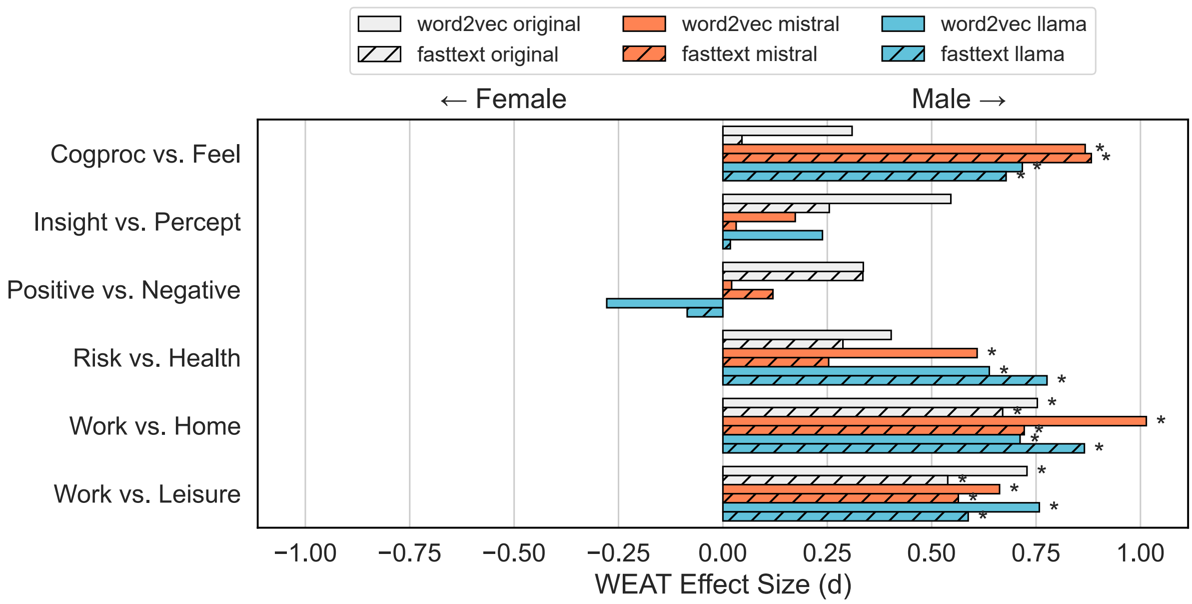


Figure F.1: WEAT effect sizes (d) for gendered word associations across all category pairs and datasets. Positive values indicate stronger associations with male terms, while negative values indicate stronger associations with female terms. Asterisks (*) denote statistically significant results ($p < 0.05$).

Local References

[212] Elis de Almeida Cardoso Caretta. Escolhas Lexicais: A Caracterização de Personagens Femininas no Discurso Literário. *Linha D'Água*, 26(1):15–28, jun. 2013. doi:10.11606/issn.2236-4242.v26i1p15-28.

[213] Soundarya Joseph. The Gender Discourse in the Novels of Margaret Laurence's and Alice Munro. In *The 2016 WEI International Academic Conference Proceedings*, pages 288–297, 2016.

[214] Andreia Filipa Nunes. Estereotipos de Xénero nos Libros Infantís Premiados. *Elos: Revista de Literatura Infantil e Xuvenil*, (6), Dec. 2019. doi:10.15304/elos.6.5867.

[215] Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/. [Accessed 10-07-2025].

[216] Oxford Children's Corpus. https://www.sketchengine.eu/oxford-childrens-corpus/0. [Accessed 10-07-2025].

[217] Google Books Ngram. https://storage.googleapis.com/books/ngrams/books/datasetsv2.html. [Accessed 10-07-2025].

[218] Google Books Syntactic Ngram. https://storage.googleapis.com/books/syntactic-ngrams/index.html. [Accessed 10-07-2025].

[219] ROCStories. https://cs.rochester.edu/nlp/rocstories/. [Accessed 10-07-2025].

[220] NovelTM Datasets. https://github.com/tedunderwood/noveltmmeta. [Accessed 10-07-2025].

[221] Project Gutenberg. https://www.gutenberg.org/. [Accessed 10-07-2025].

[222] Narrative Tropes. https://github.com/dhruvilgala/tvtropes. [Accessed 10-07-2025].

[223] Matthew Jockers. Understanding Gender and Character Agency in the 19th Century Novel, 2016.

[224] The Chicago Corpus. https://github.com/centre-for-humanities-computing/chicago_corpus. [Accessed 10-07-2025].

[225] Sixty Years of Gender Representation in Children's Books. https://osf.io/97gfk/. [Accessed 10-07-2025].

[226] Chapitres Corpus. https://chapitres.hypotheses.org/. [Accessed 10-07-2025].

[227] The Internet Movie Script Database (IMSDb). https://imsdb.com/. [Accessed 10-07-2025].

[228] Black Drama. https://www.libraries.rutgers.edu/databases/black_drama. [Accessed 10-07-2025].

[229] Literateca. https://www.linguateca.pt/Gramateca/Literateca/. [Accessed 10-07-2025].

[230] OBras. https://www.linguateca.pt/OBRAS/OBRAS.html. [Accessed 10-07-2025].

[231] Centre for Corpus Research. CLiC. https://clic.bham.ac.uk/. [Accessed 10-07-2025].

[232] moreever: Multicorpus Explorer for Explicit Value References. https://github.com/umilISLab/moreever/. [Accessed 10-07-2025].

[233] Yoav Goldberg and Jon Orwant. A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247. Association for Computational Linguistics, June 2013. URL https://aclanthology.org/S13-1035.

[234] Matthew Honnibal and Mark Johnson. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1162.

[235] Niels Dekker, Tobias Kuhn, and Marieke van Erp. Evaluating Named Entity Recognition Tools for Extracting Social Networks from Novels. *PeerJ Comput. Sci.*, 5:e189, 2019. doi:10.7717/peerj-cs.189.

[236] Sean Papay and Sebastian Padó. RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 835–841. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.104/.

[237] Sara Stymne and Carin Östman. SLäNDa: An Annotated Corpus of Narrative and Dialogue in Swedish Literary Fiction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 826–834. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.103/.

[238] Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.628.

[239] Sara Stymne and Carin Östman. SLäNDa version 2.0: Improved and Extended Annotation of Narrative and Dialogue in Swedish Literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5324–5333. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.570.

[240] Abbas Ghaddar and Philippe Langlais. WiNER: A Wikipedia Annotated Corpus for Named Entity Recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 413–422. Asian Federation of Natural Language Processing, 2017. URL https://aclanthology.org/I17-1042/.

[241] Leonardo Zilio, Maria José Finatto, and Renata Vieira. Named Entity Recognition Applied to Portuguese Texts from the XVIII Century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese*, volume 3128 of *CEUR Workshop Proceedings*, pages 1–10. CEUR-WS.org, 2022. URL https://ceur-ws.org/Vol-3128/paper10.pdf.