Gender Representation in Literature:
Analysis of Characters' Physical Descriptions

Mariana O. Silva[1], Luiza de Melo-Gomes[1], Mirella M. Moro[1]

Department of Computer Science
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil
{mariana.santos,luizademelo,mirella}@dcc.ufmg.br

Abstract. This study employs Natural Language Processing (NLP) techniques to quantitatively analyze the descriptions of male and female body parts in Portuguese literature. We investigate these descriptions' frequency, specificity, and objectification by examining a corpus of literary works. The results indicate distinct differences in how male and female bodies are portrayed, revealing evidence of gender bias in the choice of specific descriptors for body parts. This research contributes to the ongoing discourse surrounding gender representation in literature, shedding light on the potential biases in textual descriptions. Furthermore, it underscores the significance of NLP techniques in uncovering patterns within literary texts, providing valuable insights into data mining. Through this analysis, we deepen our understanding of gender dynamics within literary works and foster critical discussions on representation in literature.

CCS Concepts: • **Computing methodologies** → **Information extraction**; **Machine learning algorithms**; • **Information systems** → **Data mining**.

Keywords: data mining, natural language processing, gender representation, Portuguese literature

1. INTRODUCTION

The description of the human body is a central element in literary works as it helps readers visualize characters and immerse themselves in the story. However, how the male and female bodies are described can reflect and reinforce gender norms and biases. For example, female characters may be more objectified and described in their physical appearance, while male characters may be more described based on their actions and emotions. Furthermore, specific words used to describe different body parts can carry gender connotations and reinforce harmful stereotypes.

Although gender representation in literature has been addressed from different perspectives, including characterizations, gendered body language, and hierarchical [Adukia et al. 2022a; Cheng 2020; Jockers and Kirilloff 2016], the analysis of physical descriptions of characters' bodies has received limited attention [Hoyle et al. 2019]. This gap in the literature is particularly notable in less-spoken languages, such as Portuguese. Moreover, current works often focus on qualitative analyses, which can be subjective and difficult to generalize.

Therefore, there is a pressing need for quantitative investigations into the portrayal of male and female body parts in literary works, especially in less-spoken languages such as Portuguese. In this context, we propose a quantitative analysis of male and female body parts descriptions in Portuguese literary works. Using Natural Language Processing (NLP) techniques, we analyze literary works from different genres and periods to investigate how descriptions of the male and female bodies differ in frequency and objectification. Our analysis provides insights into how gender is represented in literature and sheds light on potential biases in descriptions of human body parts. The main contributions of this work are summarized as follows:

(1) We address the relatively unexplored field of analyzing descriptions of human body parts in literary works, particularly in less-spoken languages like Portuguese.
(2) Unlike previous works that heavily rely on qualitative analyses, we adopt a quantitative approach, enhancing the generalizability of the findings.
(3) We employ NLP techniques to conduct a quantitative analysis of male and female body part descriptions in Portuguese literary works, providing a more systematic and objective assessment of the frequency and objectification of body part descriptions.

2. RELATED WORK

Analyzing gender representation in literature is crucial as it plays a significant role in shaping societal perceptions and reinforcing gender stereotypes [Adukia et al. 2022b; Hoyle et al. 2019]. While previous studies have examined gender dynamics in various domains, such as films [Khadilkar et al. 2022], video games [Kohler et al. 2021], and scientific communities [Cordeiro et al. 2020; Pizzol et al. 2022], literature remains relatively unexplored in this context. Existing research has predominantly focused on qualitative analyses, delving into aspects like gendered body language [ermáková and Mahlberg 2020], characterization, binary notions [Cheng 2020], and gender hierarchy [Jockers and Kirilloff 2016]. However, there is a lack of comprehensive exploration regarding the physical description of male and female characters within the narrative.

Therefore, in this study, we aim to fill this research gap by adopting a quantitative approach and leveraging Natural Language Processing (NLP) techniques to analyze the physical descriptions of characters in Portuguese literature. This focus on Portuguese literature is particularly significant as it pertains to a less-studied language and offers insights specific to this linguistic and cultural context [Cristiani et al. 2020; Cardoso and Pereira 2020]. By employing NLP, we can conduct large-scale analyses encompassing diverse genres and periods, enabling us to identify patterns, tendencies, and potential biases in representing male and female bodies.

Overall, by shedding light on the portrayal of gender through the analysis of physical descriptions, our research contributes to a broader understanding of gender representation in literature. Furthermore, it provides a foundation for future studies in other languages and literary traditions, emphasizing the global significance of examining gender dynamics in literary works.

3. METHODOLOGY

This section presents the methodology used to quantitatively analyze the descriptions of male and female body parts in Portuguese literature. Sections 3.1 and 3.2 describe the data used and the process of extracting text from literary works. Section 3.3 presents the character detection stage, including identifying and unifying occurrences. Section 3.4 details the gender detection of character names. Finally, Sections 3.5 to 3.7 describe the detection of human body parts (BPs), the extraction of descriptive adjectives, and the assignment of BP to characters and their respective genders.

3.1 Dataset

In this work, we use *PPORTAL*, which stores metadata related to over 80,000 public domain works in the Portuguese language [Silva et al. 2021]. In particular, *PPORTAL* consists of three digital libraries of public domain works, mainly from Brazil and Portugal: Domínio Público,[1] Projeto Adamastor,[2]

[1] Domínio Público: `https://www.dominiopublico.gov.br`

[2] Projeto Adamastor: `https://projectoadamastor.org`

and Biblioteca Digital de Literatura dos Países Lusófonos (BLPL).[3] To obtain a representative sample of works from *PPORTAL*, a list of the top 100 literary works in Portuguese, ranked by users of the Goodreads platform, was used.[4] After cross-referencing this list with the books available in *PPORTAL*, we identified the books with download links, resulting in 60 works that met these criteria.

Character lists were obtained from external sources to facilitate the identification of characters in the analyzed works. A custom crawler was developed to automatically extract character lists from two sources: Wikipedia and the "Todo Estudo" website, which provides educational resources, including information about literary works. This approach extracted character lists from 34 of the 60 downloaded books. The complete list of the analyzed works, including title, authorship, number of identified characters, and the specific external source used, can be found in the project repository.[5]

3.2 Text Extraction

To extract the text from the downloaded PDF works, the *pdfplumber*[6] library was utilized. While effective in extraction, the library faces challenges in identifying text paragraphs due to the nature of PDF files, which often divide the text into lines instead of paragraphs. Consequently, additional text processing techniques were employed to segment narrative units like paragraphs and chapters. Since literary works typically consist of chapters, regular expressions were employed to identify chapter boundaries by matching textual elements such as "Chapter 1" or "Chapter I". However, not all downloaded files contained explicit chapter divisions. In such instances, text segmentation was performed in chunks of 100 lines, approximating a chapter. Although this approach was applied, limitations remain, and further improvements may be explored to enhance the accuracy of segmentation and extraction from PDF files.

3.3 Character Detection

The character detection stage can be divided into two tasks: (*i*) identifying occurrences of characters in the narrative, and (*ii*) unifying these occurrences, i.e., determining which occurrences correspond to the same character. Each task is described in more detail below.

Identification of character occurrences: Automating character detection in text poses challenges due to the different forms in which characters can appear, including *proper names* (e.g., "Capitu"), *pronouns* (e.g., "she"), and *nominal references*, which are anaphoric noun phrases referring to characters (e.g., "Bentinho's wife"). While existing methods handle proper names effectively, detecting pronouns and nominal references is more complex [Labatut and Bost 2019]. To address this, we adopted a semi-automated approach, utilizing external sources such as Wikipedia and the "Todo Estudo" website. A crawler was implemented to extract character lists from each literary work, but variations in formatting across Wikipedia pages required manual processing for cleaning and structuring the character name lists. The result of this step is a list for each work, providing character names and descriptions. For instance, the character Capitu from "Dom Casmurro" is described as "*Capitolina, also known as Capitu. She is the great love and wife of Bentinho. Unlike her husband, she comes from a poor family and shows intelligence and forward-thinking*".

Unification of occurrences Unifying character occurrences in text poses challenges due to the different forms of writing. Resolving coreferences is also necessary, where chains of references to the same character need to be identified. Approaches to this problem usually involve name clustering based on string similarity and gender compatibility and utilizing linguistic resources to associate

[3] BLPL: `https://www.literaturabrasileira.ufsc.br`

[4] Best Literature in Portuguese: `https://www.goodreads.com/list/show/366.Best_Literature_in_Portuguese`

[5] `https://marianaossilva.github.io/DSW2021/`

[6] *pdfplumber*: `https://github.com/jsvine/pdfplumber`

Table I. Comparison of the two gender detection heuristics.

Gender	Frequency (%)	Heuristic	Accuracy	Precision	Recall	F1-Score
Female	104 (30%)	*genderBR*	88%	86%	99%	92%
Male	228 (70%)	*Dependency-based*	89%	87%	96%	92%

nominal synonyms with the same character. In this study, we extract linguistic features from character descriptions using the *spaCy* library,[7,8] including Universal POS tags, to identify additional proper names related to a given name. By creating a list of proper names in each character's description and removing any matches with other character names, we obtain a list of nominal synonyms used for unifying character occurrences. For instance, the nominal synonyms of the character Capitu are "Capitolina" and "Capitu".

3.4 Gender Detection

Automatic gender detection of characters in literary works poses challenges due to the ambiguity of some names and the influence of cultural and historical context. Computational methods, such as Natural Language Processing (NLP) and Machine Learning, can assist in this task by extracting linguistic properties from the text. These properties, such as descriptions, pronouns, and gendered attributes, can provide clues about a character's gender. However, these approaches may be imprecise, particularly for non-traditional or ambiguous names, and require extensive training data.

Two approaches were tested to identify the gender of characters automatically. The first approach utilized the *genderBR*[9] package, which predicts gender based on Brazilian Census data. This method performs well for names clearly associated with a specific gender but can be inaccurate for non-traditional or ambiguous names. The second approach focused on analyzing descriptive text passages to determine gender. By performing dependency analysis (Section 3.6) on character descriptions, the most frequent gender of the associated words was assigned to the character. Based on gender-associated characteristics, this approach proved reliable for cases where the character's name was ambiguous or non-traditional.

To evaluate the accuracy of these approaches, manual annotation of the gender of each character was performed based on character lists. Two authors independently annotated the gender, with discrepancies resolved through discussion. The results of the manual annotation served as a reference for evaluating the approaches. Table I presents the comparative results between the two heuristics. Both methods showed satisfactory results, with F1-Scores above 90%. It is worth noting that the *genderBR* approach assigns names not classified by any gender to the majority class, which is Male, potentially introducing bias. However, despite this, both approaches demonstrate high F1 scores, indicating their effectiveness in accurately identifying character gender in the literary works.

3.5 Detection of Human Body Parts

To detect human body parts (BPs) in each literary work, a dictionary containing a set of words related to common BPs was manually compiled. A crawler was created to extract synonyms for each word from the Online Thesaurus[10] to increase the dictionary's coverage. In total, 55 BPs and 104 synonyms were considered.[11] From the resulting dictionary, we used regular expressions to identify and extract narrative units containing occurrences of BPs and their synonyms in each corpus work. For this purpose, we defined narrative units as chunks of three sentences within each works' chapter.

[7] *spaCy library:* `https://spacy.io/`

[8] We used the `pt_core_news_lg` Portuguese large model.

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

[10] Online Thesaurus: `https://www.sinonimos.com.br/`

[11] The resulting dictionary is available in the project repository: `https://marianaossilva.github.io/DSW2021/`

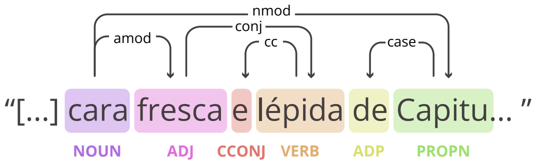

Fig. 1. Dependency analysis applied to a passage from Dom Casmurro.

3.6 Dependency Analysis

As the next step, we performed dependency analysis on the narrative units containing body parts. Specifically, this process extracts dependency graphs that represent the grammatical structure of a text. In dependency graphs, the words in the text are the nodes, and the grammatical relations are the edges. In our context, dependency analysis was applied to extract the nominal subject and adjectival/nominal modifiers related to body parts. We used the *spaCy* library for the dependency analysis. Figure 1 presents an example applied to a passage from the novel "Dom Casmurro" by Machado de Assis. A node in the graph represents each term in the passage, and the grammatical relations are indicated by edges connecting the nodes, with arrows indicating the direction of the relation. In the example, it can be observed how the character Capitu is described in terms of her physical appearance, with the noun "cara" (face) being modified by the adjective "fresca" (fresh). Although it is another modifying adjective, the word "lépida" (nimble) was not labeled as such but rather as a verb. This illustrates the challenge of using automated dependency analysis tools for literary texts, as they may not accurately capture all linguistic nuances.

3.7 Assignment

The final step in the methodology is to assign the respective adjectives and nominal subjects to the passages that contain body parts. Then, the assignments are filtered to keep only those in which the nominal subjects are related to a character in the literary work. This allows linking the gender of the character to the body part, which is, in turn, linked to the describing adjective. This process enables the analysis of how descriptions of body parts vary according to the gender of the characters. Unlike the previous step, the chapters were considered narrative units during the assignment process, allowing for a more comprehensive and contextualized analysis. That is, for each chapter, we collate all the passages containing body part descriptions of characters and proceed with the assignment of adjectives and nominal subjects to those passages.

4. RESULTS

In this section, we present the results of the quantitative analysis to investigate differences in physical descriptions of male and female characters in Portuguese literature. Firstly, we analyzed the frequency of body part mentions for each gender (Section 4.1). Then, we examined the most commonly used adjectives to describe the body parts of male and female characters, aiming to identify possible differences in specificity and objectification of the descriptions (Section 4.2).

4.1 Frequency of Descriptions

To analyze the frequency of descriptions regarding the gender of characters, we evaluated the gender bias for each description. Specifically, for each body part BP, two percentages are calculated for the female and male genders: $pct_F(BP)$ and $pct_M(BP)$, respectively. Each percentage is obtained by dividing the number of times a particular BP is assigned to characters of the corresponding gender by the total number of assignments of that BP to any gender. For a given BP, the formulas for calculating each percentage are as follows:

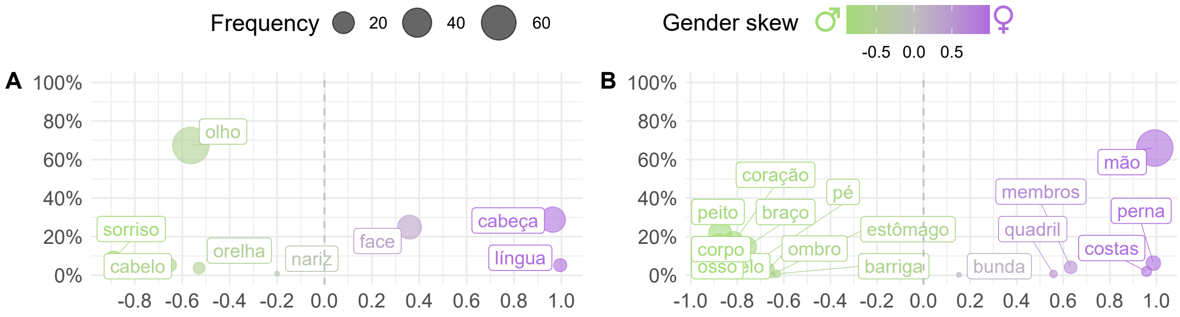

Fig. 2. Gender skew of body parts mentioned for characters in literature, separated into descriptions related only to (A) head and only to (B) body.

$$pct_F(BP) = \frac{Number\ of\ times\ the\ BP\ is\ assigned\ to\ female\ characters}{Total\ number\ of\ assignments\ of\ BP\ to\ female\ characters}$$

$$pct_M(BP) = \frac{Number\ of\ times\ the\ BP\ is\ assigned\ to\ male\ characters}{Total\ number\ of\ assignments\ of\ BP\ to\ male\ characters}$$

From the calculated percentages, the gender skew of each description $skew(BP)$ is calculated as the difference between $pct_F(BP)$ and $pct_M(BP)$ divided by the sum. This calculation can range from -1 to 1, allowing us to identify if a description is more frequently used for male characters (-1), female characters (1), or used equally for both genders (0). For example, if a BP is assigned to female characters 70% of the time and to male characters 30% of the time, the percentage $pct_F(BP)$ is 0.7, and the percentage $pct_M(BP)$ is 0.3. The gender $skew(BP)$ is then calculated as: $\frac{pct_F(BP)-pct_M(BP)}{pct_F(BP)+pct_M(BP)} = \frac{0.7-0.3}{0.7+0.3} = 0.4$. In this case, the gender skew is positive for this BP, indicating that it is more frequently used for female characters than male characters, but it is not exclusive to either gender.

Figures 2A and 2B present the analysis of gender skew for different body parts mentioned in the literature. The descriptions were manually separated into two categories: parts related only to the head and parts related only to the body, resulting in two separate plots. The bubbles in the plots indicate the frequency of the respective descriptions, with larger bubbles representing more frequent descriptions. The horizontal position of the bubbles indicates the gender skew for each description, with positive values indicating a higher frequency of descriptions for female characters and negative values indicating a higher frequency of descriptions for male characters.

In Figure 2A, most of the descriptions related only to the head show a more negative skew, indicating that female characters are less described regarding their facial physical characteristics than male characters. The same trend is observed for descriptions related only to the body (Figure 2B), where female characters are less described regarding their physical body characteristics than male characters. These results suggest a tendency in the literature to provide less detailed descriptions of female characters' physical features, especially regarding the head and body. These differences may indicate gender stereotypes and highlight the need to analyze gender representations in literary works critically.

Regarding gender, the descriptions most frequently used to describe facial physical characteristics of male characters are "eye" and "smile", while "head" and "face" are the most frequent descriptions for female characters. When analyzing physical body characteristics, terms such as "chest", "heart", "arm", "body", and "hair" are more likely to be used to describe male characters, while for female characters, the most frequent terms include "hand", "leg", "back", and "hip". These observations suggest a tendency for female characters to be described in terms of physical body characteristics that are more commonly associated with sexualization and objectification, while male characters are described in terms of more neutral or positive characteristics.

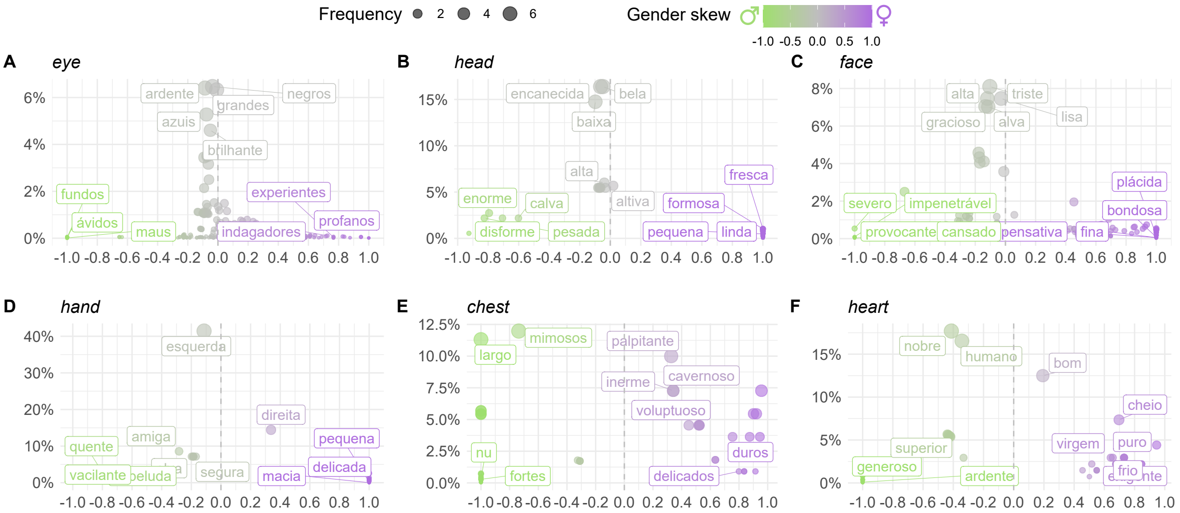

Fig. 3. Gender skew of the three most frequent adjectives describing (A-C) facial and (D-F) bodily physical characteristics mentioned for characters in literature.

4.2 Frequency of Adjectives

To gain a deeper understanding of the results, we also analyzed the most commonly used adjectives to describe the body parts of male and female characters. Similar to the previous analysis, we assessed the gender skew for each adjective that describes a body part. Specifically, for each BP (body part) and adjective adj, two percentages were calculated for both genders: $pct_F(BP, adj)$ and $pct_M(BP, adj)$. Each percentage is obtained by dividing the number of times a particular BP, described by the adjective adj, is attributed to characters of the given gender by the total number of attributions between BP, described by adjectives, and the gender in question. For each BP and adjective adj, the calculation of the percentages for the female and male genders is as follows:

$$pct_F(BP, adj) = \frac{\text{number of times } BP\text{, described by } adj\text{, is attributed to female characters}}{\text{total number of attributions of } BP\text{, described by adjectives, to female characters}}$$

$$pct_M(BP, adj) = \frac{\text{number of times } BP\text{, described by } adj\text{, is attributed to male characters}}{\text{total number of attributions of } BP\text{, described by adjectives, to male characters}}$$

Once again, the gender skew of each description is calculated as the difference between the percentage of attribution of the description to female characters, $pct_F(BP, adj)$, and the percentage of attribution to male characters, $pct_M(BP, adj)$, divided by the total attributions for that description, $pct_F(BP, adj) + pct_M(BP, adj)$. The result of this equation indicates whether the description is more frequently attributed to female characters (positive value), male characters (negative value), or used equally for both genders (0).

Figures 3(A-C) and 3(D-F) present, respectively, the analysis of gender skew for the three most frequent adjectives describing facial and bodily physical characteristics mentioned for characters in literature. Regarding **facial physical characteristics**, the most frequent descriptions for male characters include eyes *keen, deep, evil*; head *misshapen, heavy, huge, bald*; and face *stern, impenetrable, provocative, tired.* On the other hand, for female characters, the descriptions are eyes *inquisitive, experienced, profane*; head *beautiful, lovely, small, fresh*; and face *thoughtful, kind, serene, delicate.*

Regarding **bodily physical characteristics**, the most frequent descriptions for male characters include hand *hairy, warm, trembling*; chest *broad, strong, robust*; and heart *generous, burning, noble.* Meanwhile, for female characters, the descriptions are hand *delicate, small, soft*; chest *voluptuous, delicate, firm*; and heart *virgin, pure, demanding, cold.* These descriptions reflect stereotypes and sometimes objectification of both male and female bodies. Furthermore, how physical characteristics are described can reinforce beauty standards and harmful gender norms for both genders.

5. CONCLUSION

This study examined descriptions of body parts in Portuguese literature for male and female characters, utilizing Natural Language Processing techniques on a corpus of diverse Portuguese literary works. The analysis revealed a tendency for more detailed descriptions of male characters' body parts compared to females, with the latter often objectified while the former focused on actions and emotions. Significant differences in frequency, objectification, and gender bias in body part descriptors were identified. These findings emphasize the importance of gender representation and call for critical reflection on how physical attributes are portrayed. The study also showcases the potential of NLP techniques in identifying patterns and biases in literature, offering insights for future research.

Limitations and Future Work. Despite the relevant insights, some limitations of this work must be acknowledged. For instance, the analyzed corpus of literary works is rather restricted, which may introduce biases and limit the generalizability of the findings. Additionally, gender detection relied on character names, which may not accurately reflect the gender identity of some characters. Lastly, the objectification of characters can be assessed in various ways, and the methodology employed in this study may only capture some nuances of objectification. These limitations highlight the need for more comprehensive research to address these issues.

Acknowledgments. This work was partially funded by CNPq, CAPES, and FAPEMIG, Brazil.

REFERENCES

Adukia, A. et al. Portrayals of race and gender: Sentiment in 100 years of childrens literature. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*. Seattle, WA, USA, pp. 2028, 2022a.

Adukia, A. et al. Tales and tropes: Gender roles from word embeddings in a century of children's books. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022.* International Committee on Computational Linguistics, pp. 3086–3097, 2022b.

Cardoso, B. and Pereira, D. Evaluating an aspect extraction method for opinion mining in the portuguese language. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning.* SBC, pp. 137–144, 2020.

Cheng, J. Fleshing out models of gender in english-language novels (1850–2000). *Journal of Cultural Analytics* 5 (1): 11652, 2020.

Cordeiro, D. et al. Representativeness of women in postgraduate programs in computer science in brazil. In *Anais do XIV Women in Information Technology.* SBC, Cuiabá, pp. 110–119, 2020.

Cristiani, A., Lieira, D., and Camargo, H. A sentiment analysis of brazilian elections tweets. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning.* SBC, pp. 153–160, 2020.

Hoyle, A. et al. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL.* Vol. 1. pp. 1706–1716, 2019.

Jockers, M. and Kirilloff, G. Understanding gender and character agency in the 19th century novel. *Journal of Cultural Analytics* 2 (2), 12, 2016.

Khadilkar, K., KhudaBukhsh, A. R., and Mitchell, T. M. Gender bias, social bias, and representation: 70 years of bollywood. *Patterns* 3 (2): 100409, 2022.

Kohler, L. et al. A representatividade feminina nos jogos digitais. In *Anais do XV Women in Information Technology.* SBC, Evento Online, pp. 265–269, 2021.

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

Pizzol, N. D., Barbosa, E., and Musse, S. Gender representation in brazilian computer science conferences. In *Anais do XVI Women in Information Technology.* SBC, Niterói, pp. 67–76, 2022.

Silva, M. O. et al. PPORTAL: Public domain Portuguese-language literature Dataset. In *SBBD DSW.* SBC, Rio de Janeiro, Brazil, pp. 77–88, 2021.

ermáková, A. and Mahlberg, M. Gender inequality and female body language in childrens literature. *Digital Scholarship in the Humanities* 36 (Supplement_2): ii72–ii77, 12, 2020.