Gender Identification in Brazilian Portuguese Product Reviews: A Comparative Study of Classical Models, mBERT, and LLMs

Tiago de Melo and Carlos M. S. Figueiredo
LSI - Intelligent Systems Laboratory
UEA - State University of Amazonas
Manaus-AM
{tmelo, cfigueiredo}@uea.edu.br

Abstract

This study analyzes gender identification in Brazilian Portuguese using Amazon reviews drawn from ten product categories. Ten models were evaluated: three classical classifiers (Logistic Regression, Random Forest, and SVM), two BERT models for Portuguese and multilingual, and five LLMs (ChatGPT 4o, ChatGPT 3.5, DeepSeek, Sabia3, and Sabiazinho). Experiments show that mBERT achieved the best performance (macro-F1 = 0.634), outperforming ChatGPT 4o and Logistic Regression by less than one percentage point. Reviews authored by women reach an average F1 of 0.654—four points higher than those by men. Performance also varies by domain: books and automotive are easier, whereas baby and pets are more challenging.

1 Introduction

Author profiling is the task of inferring author characteristics, such as gender, age range, education level, social class, and cultural background. This task has received considerable attention in recent years due to its new applications and utility in the fields of marketing and forensic linguistics (Bsir and Zrigui, 2018; Suman et al., 2022). For example, author profiling can be applied to uncover cybercriminal acts and identify sexual harassers or pedophiles. It can also aid in the analysis of human behavior and the recognition of personality types.

Among the characteristics of the authors, a relevant task is the identification of the gender of the text, known as gender profiling. Application examples in the field of digital marketing include the analysis of user consumption behavior, the personalization of the interface, and the monitoring of informational campaigns (Suman et al., 2022). When considering applications involving sensitive data, such as in healthcare, assessing the possibility of gender identification can be a relevant tool to indicate the need for anonymization or bias elimination (Majeed and Lee, 2021).

Studies in English show a consistent advantage of neural architectures over traditional methods in the task of author profiling (Rangel et al., 2020). However, such results do not directly transfer to Portuguese due to differences in gender inflection, pronoun usage, and regional variants (Luegi et al., 2024). Furthermore, most studies focus on a single textual domain — for example, tweets or news — making it difficult to assess whether the model's difficulty varies across thematic contexts. Research in Spanish indicates that domain-specific terms can strongly influence performance (García-Díaz et al., 2024); however, there is a lack of systematic evidence in Brazilian Portuguese (PT-BR) to confirm or refute this effect in e-commerce environments.

The growth of e-commerce in Brazil has made consumer reviews a strategic resource for marketing, recommendation systems, and algorithmic bias audits (Veiga et al., 2024). Identifying the author's gender from text in Brazilian Portuguese remains a little-explored task, as it is dependent on the typical informality of these messages and the morphosyntactic particularities of the language (Dias and Paraboni, 2020). On digital platforms, it is common for users to use nicknames or keep their profiles anonymized, which prevents the direct acquisition of demographic attributes. Figure 1[1] illustrates a real example: no profile information accompanies the text, forcing the model to infer gender exclusively from the text.

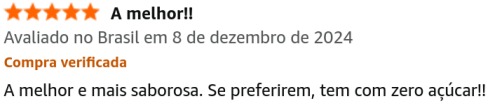

Figure 1: Example of an Amazon product review.

[1]The best and most delicious one. If you prefer, there's a sugar-free option!!

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 11–19

This work addresses this gap by exclusively analyzing the raw text of reviews from ten Amazon Brazil categories: *automobile*, *baby*, *toy*, *cellphone*, *food*, *game*, *laptop*, *book*, *fashion*, and *pets*, which were chosen because they are very popular, thematically diverse, and heavily populated with reviews. The study compares the following traditional algorithms: Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM); neural models, represented by multilingual mBERT (Devlin et al., 2019) and BERTimbau (Souza et al., 2019), which is specialized in the Brazilian Portuguese language; and recent Large Language Models (LLMs), including ChatGPT-4o, ChatGPT3.5, DeepSeek, Sabia3, and Sabiazinho. For this purpose, a vast set of user reviews was collected and manually labeled, organized by gender and category.

To the best of our knowledge, no previous work has comprehensively investigated the use of AI models for gender identification in Brazilian Portuguese from reviews across multiple product categories. To fill this gap, this study answers the following research questions (RQs):

- RQ1 – Which of the investigated models achieves the best performance on the gender identification task? To answer this question, each model was evaluated on a diverse set of metrics.
- RQ2 – Is one gender harder to predict? To answer this question, the F1 score was calculated per gender, using the best-performing model identified in RQ1.
- RQ3 – Does the difficulty vary across product categories? To answer this question, the F1 score was calculated per category, again using the best-performing model from RQ1.

The contributions of this work can be summarized as follows: 1) The construction of a multi-category dataset[2] in Brazilian Portuguese (PT-BR), which will be released to the scientific community. 2) A comprehensive comparison between classic models, neural networks, and recent LLMs in the gender identification task. 3) A detailed analysis of the performance variation by gender and category, revealing domains of high and low difficulty. 4) A discussion of limitations for future commercial and academic applications.

This paper is organized as follows: Section 2 reviews the related literature, Section 3 details the methodology, Section 4 presents the results, Section 5 presents and discusses the limitations of this study, and Section 6 concludes the paper, suggesting directions for future research.

2 Related Works

Gender prediction from text falls within the field of *author profiling*. This section presents the most relevant studies, organizing them into two main areas: research conducted in other languages and specific investigations in Brazilian Portuguese (PT-BR).

2.1 Gender Identification in Other Languages

Early works focused on English and Spanish, generally with short texts from social networks. Volkova et al. (Volkova et al., 2015) showed that lexico-syntactic features combined with word embeddings outperform approaches based solely on n-gram frequencies for gender and age on X (formerly Twitter). Rangel et al. (Rangel et al., 2020) reviewed eight editions of the international PAN challenge[3], a competition dedicated to author profiling and computational forensics tasks. Their analysis revealed consistent advances of neural network-based models, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), over classic methods, and also highlighted the negative impact of class imbalance on model performance.

Vashisth and Meehan (Vashisth and Meehan, 2020) evaluated fastText, BiLSTM, and BERT in English, concluding that contextualized language models are more robust to orthographic noise. Research in low-diffusion languages, such as Thai (Jintawatsakoon and Poonsawat, 2023) and Hausa (Onikoyi and Adamu, 2023), reinforces that longer texts improve accuracy and that incorporating metadata (profile description, location) can increase the F1 score by up to 6 percentage points.

The present work is inspired by these studies conducted in other languages to perform a novel analysis for a task in Brazilian Portuguese, simultaneously evaluating the performance of different models, such as BERT architectures or modern

[2]The dataset used in this study is publicly available on Zenodo: https://doi.org/10.5281/zenodo.18856652

[3]https://pan.webis.de

LLMs. Furthermore, this work aims to comparatively verify whether there are significant performance differences between multilingual models and models specifically trained in Portuguese.

2.2 Gender identification in Brazilian Portuguese

For PT-BR, the literature is still incipient and mostly limited to the X domain. Dias and Paraboni (Dias and Paraboni, 2020) investigated three corpora (X, Facebook, and forums) and demonstrated a performance loss in *cross-domain* scenarios when using only n-grams. The authors tested SVM, Random Forest, and Naïve Bayes, showing that char n-grams offer greater robustness than word n-grams. Lopes and Oliveira (Lopes and Oliveira, 2023) designed a neural network cascade that achieves an F1 score of 0.79 on tweets, but the gain vanishes when the model is applied to blog posts. In the line of large pre-trained models, Souza et al. (Souza et al., 2019) introduced BERTimbau, which improves the F1 score on a sentiment task but has not yet been evaluated for gender prediction in PT-BR.

Despite these advances, no study was found that uses e-commerce reviews or analyzes how difficulty varies across different product categories. This work seeks to fill these gaps by building a multi-category dataset from Amazon Brazil, balanced by gender, and by comparing ten models from different families, measuring performance by gender and thematic domain.

3 Methods

3.1 Dataset

The dataset comprises reviews posted on Amazon Brazil between 2021 and 2024, all in Portuguese. The user reviews were collected for ten categories — automobile, baby, toy, cell phone, food, game, laptop, book, fashion, and pets — to ensure thematic balance.

The gender annotation was based on the name provided by the review's author. The process consisted of two independent stages. First, a manual screening of profiles was performed, removing: (i) opaque nicknames (e.g., "usuarioXPTO", "comprador321"); (ii) ambiguous names that could indicate more than one gender (e.g., "Alex"); and (iii) profiles with only surnames (e.g., "Silva79"). Next, two annotators individually judged the gender of the remaining profiles, considering the stated name and the public profile information on the platform. Each sample was annotated by both annotators; when there was disagreement between them, the corresponding profile name was removed from the dataset. The user names were anonymized, and the dataset will be made available for future work.

When annotations rely solely on names, noise may be introduced due to differences between the gender assigned at birth and the one self-identified in the posted reviews. Furthermore, official Brazilian name lists, such as those provided by IBGE[4], are insufficient to ensure high accuracy, since many users employ diminutives, nicknames, or unisex names. In addition, in cyber investigation contexts, authors often conceal their identities, making simple name-to-list matching ineffective. These factors highlight the importance of exploring models that rely not only on precompiled lists but also on linguistic and contextual cues for gender prediction.

After the labeling process, a dataset balanced by construction was obtained, composed of 2,000 reviews labeled as female and 2,000 as male, uniformly distributed among the ten categories. This balancing was intentional, as it aims to prevent class imbalances from distorting the comparisons between models, even though the resulting distribution does not necessarily reflect the actual proportion of genders among the platform's users.

3.2 Machine Learning Models

The study compares ten models organized into three groups. The first group includes three traditional Machine Learning algorithms: LR (Cabrera et al., 1994), RF (Breiman, 2001), and SVM (Cortes and Vapnik, 1995), which are widely used in classification problems (de Oliveira and de Melo, 2021; Ahmed and Ghabayen, 2022; Almeida Neto and de Melo, 2023; Hanić et al., 2024). These algorithms were applied to vectors generated by Term Frequency-Inverse Document Frequency (TF-IDF), with n-grams of order 1 and 2 (unigrams and bigrams) and a limit of the 3,000 most frequent terms. This textual representation was chosen for its simplicity and effectiveness in text classification tasks, allowing the capture of relevant lexical patterns. For each model, the best hyperparameters were selected automatically using GridSearchCV with 3-fold internal validation, using F1-macro as the evaluation metric.

[4] https://www.ibge.gov.br/estatisticas/multidominio/genero.html

The second group includes two pre-trained neural models available on the Hugging Face platform[5]: multilingual mBERT (Devlin et al., 2019) and BERTimbau (Souza et al., 2019). mBERT was included for its wide adoption and to evaluate the impact of general BERT variants on tasks in Portuguese. BERTimbau, trained specifically on Brazilian Portuguese corpora, allows a direct comparison to determine whether language specialization brings performance gains on the task under analysis. Although pre-trained models allow for fine-tuning, we chose not to incorporate this capability within the scope of the current project. Fine-tuning is left for future work.

The third group includes LLMs. ChatGPT 4o, ChatGPT 3.5 (Kalyan, 2024), DeepSeek (Guo et al., 2025), and the Brazilian models Sabia3 and Sabiazinho (Pires et al., 2023). As these models are closed, they are only accessible through their official API to be used in a zero-shot way. To reduce the stochastic variability of LLMs, the temperature was set to 0 for all requests, improving comparable responses across different runs and models. This choice eliminates the noise introduced by random sampling, allowing the observed differences to reflect the inherent behavior of each LLM, rather than decoding variations. As these LLMs already produce textual output according to instructions, each review was provided to the model with a fixed prompt requesting the gender in the format `M` (male) or `F` (female). Figure 2 presents the prompt used in the experiments, with its structure divided into five blocks and the description of each component on the right. Figure 3 shows the same prompt originally written in Brazilian Portuguese.

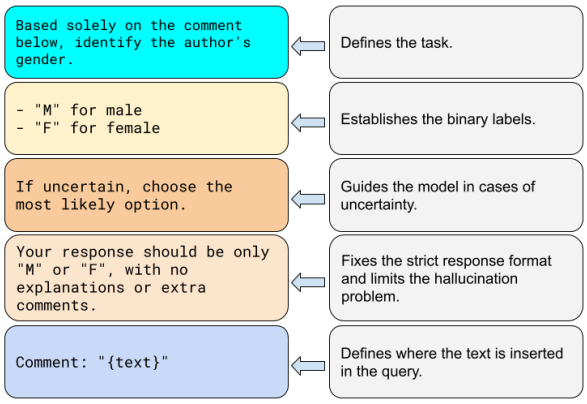

Figure 2: Prompt (translated) used in LLMs.

[5] https://huggingface.co/

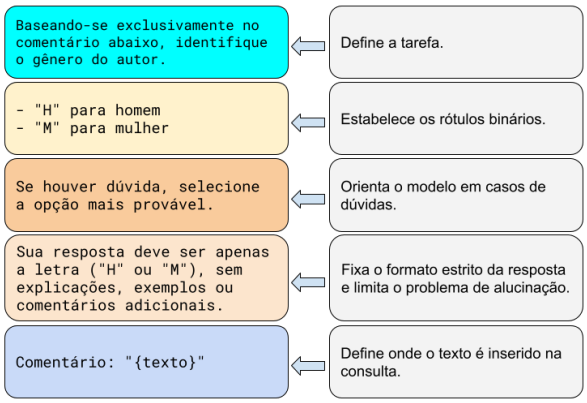

Figure 3: Prompt (original) used in LLMs.

3.3 Evaluation Metrics

The performance of the models is evaluated by four classic classification metrics applicable to text classification tasks (Cunha et al., 2025): Precision, Recall, F1-macro, and Accuracy. Let TP, FP, FN, and TN be true positives, false positives, false negatives, and true negatives, respectively. Precision measures the proportion of correct predictions among all instances labeled as positive, $Precision = TP/(TP + FP)$, while Recall measures the proportion of positive instances that were retrieved, $Recall = TP/(TP + FN)$. The F1 score is the harmonic mean of Precision and Recall. In this study, the macro version of the F1 score is used, calculated as the arithmetic mean of the F1 scores for each class, assigning equal weights to the male and female genders, regardless of their frequency. Accuracy corresponds to the fraction of correct predictions among all instances, $Accuracy = (TP + TN)/(TP + TN + FP + FN)$.

Since the dataset is balanced, Accuracy coincides with the F1-micro score, while the F1-macro score reveals performance differences between the classes. In addition to the averages, the standard deviation is reported across the five folds for the F1-macro metric, as an indicator of model stability. To investigate biases, the F1 score per class was calculated ($F1_M$ and $F1_F$), allowing a direct comparison of the prediction difficulty for each gender.

4 Results

4.1 Comparison of Models (RQ1)

To address Research Question 1 (RQ1), which asks which approach achieves the best average performance on the task, the ten models were evaluated using 5-fold stratified cross-validation. In each

fold, Logistic Regression, Random Forest, and SVM were trained on 80% of the user reviews (4 folds) and evaluated on the remaining 20% (1 fold). mBERT, BERTimbau, and the five LLMs (ChatGPT 4o, ChatGPT 3.5, DeepSeek, Sabia3, and Sabiazinho) do not require fine-tuning; therefore, in each fold, they were applied directly to the same test subset. mBERT operated in zero-shot mode through the text-classification pipeline from the Hugging Face library, while the LLMs received the prompt from Figure 3 and returned the predictions `F` or `M`. All models generated outputs on the same test sets, ensuring a fair comparison. The means and standard deviations calculated across the five folds are presented in Table 1.

mBERT achieved the highest mean F1-macro (0.634 ± 0.008), followed by ChatGPT 4o (0.627 ± 0.007) and Logistic Regression (0.624 ± 0.004). The difference between mBERT and ChatGPT 4o is less than one percentage point, indicating that zero-shot models based on medium-sized transformers and state-of-the-art LLMs exhibit very close performance in this domain. Among the trained models, SVM and LR were, respectively, 1.6 and 1.0 percentage points below mBERT, but they showed the lowest variability, signaling greater consistency.

In order to compare the models' performance, the Friedman test was applied, followed by the Nemenyi post-hoc test with a significance level of 0.05. The results indicated that mBERT, ChatGPT 4o, DeepSeek, LR, RF, and SVM form a statistically indistinct group; BERTimbau, ChatGPT 3.5, Sabia3, and Sabiazinho comprise a second group with lower performance. BERTimbau and Sabia3 exhibited the largest standard deviations, indicating strong instability across folds. Furthermore, it is observed that these models trained specifically on Portuguese language data underperformed the evaluated multilingual models, which contradicts the expectation that language specialization provides gains in linguistic tasks. A possible explanation for this result lies in the diversity and scope of the data used in the models' pre-training. Multilingual models like mBERT and DeepSeek were exposed to large volumes of data across multiple domains and registers, including Portuguese, which may have favored their generalization capability. Additionally, gender inference from short and often ambiguous reviews may demand greater semantic and contextual robustness, favoring larger-scale and more complex models, even if not exclusively trained in Portuguese.

Figure 4 illustrates the F1-score variability among the models. LR and SVM exhibit narrow violins, indicating that their results hardly oscillate across the folds. mBERT and ChatGPT 4o maintain high medians but with slightly wider ranges, indicating high performance that is, however, subject to fluctuations. DeepSeek shows the widest amplitude, revealing some instability among runs. Sabia3 and BERTimbau show two well-defined bumps, indicating a bimodal distribution and an alternation between good and bad folds. Finally, Sabiazinho concentrates almost all its values in the lower part of the plot, confirming a consistently low performance.

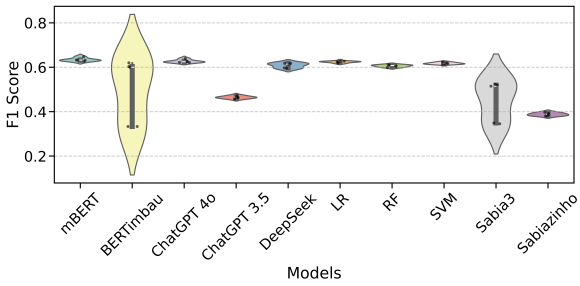

Figure 4: Variation of F1-Score by model.

In summary, the results show that: (i) the absolute gains of LLMs over traditional models are modest with adequate textual representations and a balanced corpus; (ii) a transformer in zero-shot mode, such as mBERT, offers the best cost-benefit ratio; and (iii) open-source portuguese-specialized LLMs, such as Sabia3, Sabiazinho, and BERTimbau, do not yet achieve competitive performance on the gender identification task probably because they are trained on less data.

4.2 Performance per Gender (RQ2)

RQ2 evaluates the difficulty of identifying one of the genders. mBERT was used due to its better F1-macro performance in RQ1. Figure 5 summarizes the F1-score per gender, where the left panel presents the values for each of the five cross-validation folds, while the right panel shows the mean and standard deviation.

The results reveal a consistent advantage for the female gender. The female F1 score ranges from 0.645 to 0.659, with a mean of 0.654 and a standard deviation of 0.006. The male F1 score oscillates between 0.607 and 0.636, with a mean of 0.614 and a standard deviation of 0.012. Therefore, male-authored text not only has a lower F1 score but

Table 1: Models' performance.

Model	Precision	Recall	F1-Score	Accuracy
mBERT	**0.637**	**0.635**	**0.634** ± 0.008	**0.635**
BERTimbau	0.465	0.565	0.498 ± 0.151	0.538
ChatGPT 3.5	0.583	0.538	0.465 ± 0.006	0.565
ChatGPT 4o	0.627	0.627	0.627 ± 0.007	0.627
DeepSeek	0.611	0.609	0.608 ± 0.011	0.609
LR	0.624	0.624	0.624 ± 0.004	0.624
RF	0.607	0.607	0.606 ± 0.005	0.607
SVM	0.618	0.618	0.618 ± 0.004	0.618
Sabia3	0.547	0.497	0.451 ± 0.094	0.573
Sabiazinho	0.606	0.517	0.388 ± 0.007	0.517

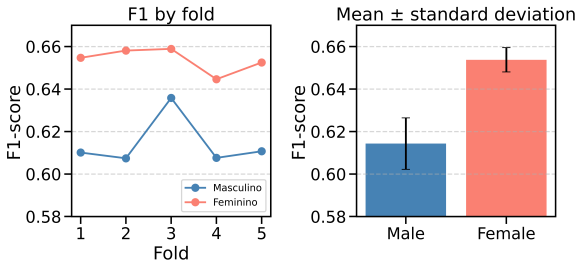

Figure 5: mBERT F1 by gender: on the left, results for each fold; on the right, mean and standard deviation.

also exhibits greater instability across folds. This difference appears to arise from two linguistic aspects. First, adjectives and participles inflected with "-a" appear more frequently in reviews by female authors, for example, "*estou encantada*" (I'm delighted) or "*fiquei satisfeita*" (I was satisfied). Such endings provide a direct morphological signal to the classifier. Second, female-authored texts include more varied affective vocabulary and intensifiers, e.g. "*amei demais*" (I loved it so much), and "*super recomendo*" (I highly recommend). Male-authored comments tend to be more objective and focus on specifications, reducing the number of explicit cues.

It is concluded that the male gender presents the greater challenge for mBERT, highlighting an asymmetry in the task's difficulty. Future research could mitigate this limitation by incorporating more significant male examples or by adjusting the prompt to explore linguistic cues beyond nominal agreement.

4.3 Performance per Category (RQ3)

To answer RQ3, the mBERT model, which led to the comparison in RQ1, was applied to each of the ten product categories. Figure 6 presents the mean F1-macro score (blue bars) and the corresponding standard deviation (horizontal lines) obtained across the five folds.

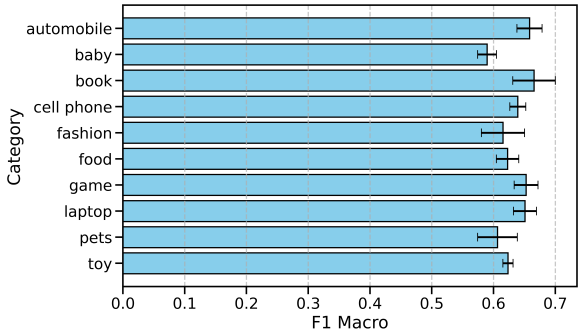

Figure 6: mBERT result by category.

The best scores are concentrated in the book (0.69 ± 0.03) and automobile (0.66 ± 0.02) categories. Reviews in these categories tend to be longer and more opinionated, with greater use of adjectives and participles that are inflected for gender, providing clearer cues to the classifier. In sentences like "*O carro entregou exatamente o que eu precisava para viajar* ***sozinho***" (The car delivered exactly what I needed to travel alone) or "*Fiquei* ***encantada*** *com a narrativa*" (I was delighted with the narrative), the decisive information lies in the morphological markers "sozinho" and "encantada", not in the pronouns "eu" (I) or "meu" (my). The frequent presence of gender agreement—adjectives ("satisfeit**o/a**"), participles ("empolgad**o/a**"), or marked nouns ("leit**or/ora**")—increases the F1 score, explaining why reviews about books and automobiles are more accessible to the model.

In the intermediate block are cell phone, game, and laptops (0.64). These texts combine technical specifications with subjective impressions; the specialized lexicon is useful, but the density of gender

markers is lower than that observed in book. Reviews like "*Processador rápido e ótimo para jogar com meus amigos*" (Fast processor and great for playing with my friends) exemplify this balance between neutral terms and personal references.

Meanwhile, toys, food, and fashion hover around an F1 score of 0.61. Here, short phrases or evaluation formulas (e.g., "*Gostei muito*" (I liked it a lot), "*Chegou certinho*" (It arrived correctly)) predominate, offering weaker signals to the classifier. In fashion, for example, generic compliments about the design replace details that could be associated with a specific gender.

The worst results appear in the pets (0.56 ± 0.03) and baby (0.59 ± 0.02) categories. In these domains, a large portion of the reviews describes the product's utility for the pet or the baby, employing impersonal constructions: "*Excelente para alimentar gatos de rua*" (Excellent for feeding stray cats) or "*Muito bom para a pele do meu filho*" (Very good for my son's skin). The absence of autobiographical markers dilutes the linguistic cues. Furthermore, the variability of the vocabulary — from affective terms to objective descriptions — increases the standard deviation and reduces the model's reliability.

In summary, Figure 6 indicates that mBERT distinguishes gender more effectively in texts with extensive personal impressions and less thematic neutrality (books, auto), while it underperforms in categories with functional or impersonal statements (baby, pets). These results suggest that future improvements could come from additional pretraining on domestic corpora or from incorporating pragmatic signals, such as emojis or verb tense, into the input vectors.

4.4 Qualitative Error Analysis

In addition to the quantitative comparison between the models, a qualitative analysis was conducted on the examples where mBERT, the model with the best average performance, recurrently failed. The reviews that presented incorrect predictions in all five folds were extracted, with special attention to the categories with the lowest average performance, such as baby and pets. Table 2 presents a representative set of these cases.

Initially, it is noted that these comments have an extremely concise structure, generic vocabulary, and the absence of explicit linguistic gender markers, such as pronominal or adjectival inflections. This configuration reduces the presence of useful cues for supervised inference. For example, the pets category has an average of 16 words per comment, while the baby category has an average of 13 words per comment. Therefore, the cases in which the classifier failed are mostly very short texts, which make classification difficult for the models.

Table 2: Comments with incorrect predictions.

Category	Gender	Comment
baby	F	*Gostei muito* (I really liked it)
baby	F	*EXCELENTE QUALIDADE* (EXCELLENT QUALITY)
baby	M	*Produto de excelente qualidade* (Product of excellent quality)
baby	M	*GOSTEI DO PRODUTO* (I LIKED THE PRODUCT)
baby	F	*Excelente qualidade* (Excellent quality)
pets	M	*Gostei muito* (I really liked it)
pets	F	*Produto muito bom* (Very good product)
pets	M	*Ótimo produto* (Great product)
pets	F	*Chegou antes do prazo* (It arrived ahead of schedule)
pets	M	*Produto de qualidade* (Quality product)

Additionally, generic and high-frequency expressions, such as "*gostei muito*" (I liked it a lot) and "*produto excelente*" (excellent product), tend to occur in comments from different authors, regardless of gender. This ambiguity can confuse the model during training, weakening the statistical association between the text and the label.

These results reinforce the importance of considering, in future work, strategies that incorporate additional contextual signals, such as the overall discursive style or metadata associated with the author, as well as mechanisms that penalize decisions based on ambiguous linguistic patterns.

5 Limitations and Ethical Considerations

Despite the promising results, this work has limitations that require a thorough discussion, especially concerning the ethical considerations of using gender inference models.

First, it is crucial to clarify that the gender inference performed in this study corresponds to a statistical approximation based on observable patterns in the text data. It does not reflect the complexity, fluidity, and self-identification of gender (Keyes,

2018). Gender is a social and personal construct that transcends binary classifications or characteristics inferred by an algorithm. Ignoring this fundamental distinction can lead to the oversimplification of identities and the reproduction of stereotyped gender notions. Therefore, any application of the proposed models must be carried out with caution, acknowledging their limited nature.

Second, the application of gender inference models in sensitive contexts or automated systems must be accompanied by serious ethical concerns. The risk of algorithmic bias is significant, as the model could perpetuate or even amplify existing societal prejudices (Verma, 2019). In areas such as recruitment, targeted advertising, or security systems, the unregulated use of this technology can result in discrimination and injustice. Therefore, the use of these models must be accompanied by a rigorous ethical impact assessment, ensuring that gender inference is not the sole criterion for making important decisions about individuals.

6 Conclusions and Future Work

This paper presented a systematic evaluation of ten popular models for gender identification in Brazilian Portuguese, based exclusively on the text of Amazon reviews. The corpus contains 4,000 instances balanced between female and male, distributed across ten categories, and has been made available to the community, filling a data gap in Brazilian Portuguese (PT-BR) for author profiling.

Multilingual mBERT, applied in zero-shot mode, achieved the best F1-macro score (0.634), outperforming ChatGPT-4o and Logistic Regression by less than one percentage point. This result demonstrates that, with robust representations and a balanced corpus, moderately costly models can rival state-of-the-art LLMs. The analysis by gender showed a consistently superior performance for female-authored texts (F1 score of 0.654 versus 0.614), attributed to a greater presence of morphological markers and affective vocabulary. At the domain level, four categories (book, automobile, laptop, and game) reached an F1 score ≥ 0.65, while two categories (baby and pets) fell below 0.60, suggesting a direct influence of linguistic style and the degree of text personality.

In terms of practical contribution, the results indicate that a medium-sized transformer, used in zero-shot mode, offers the best cost-benefit ratio, as it does not require fine-tuning, reduces computational demand, and still outperforms or matches proprietary LLMs in Portuguese. Such a configuration favors real-world monitoring applications on e-commerce platforms, where latency and cost are critical factors.

As future work, we plan to: (a) expand the corpus with labels obtained via the crowdsourcing of anonymous profiles, (b) evaluate domain-specific monolingual BERT models, (c) perform and evaluate fine-tuning of BERT models, and (d) apply prompt engineering to LLMs to investigate pragmatic cues. These directions may increase the coverage and fairness of gender profiling systems in the Portuguese language.

7 Acknowledgements

The authors acknowledge the support provided by the Universidade do Estado do Amazonas (UEA) through the Academic Productivity Grant (GPA) (Administrative Ordinance No. 1177/2025-GR/UEA). This work was also supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Basem H Ahmed and Ayman S Ghabayen. 2022. Review rating prediction framework using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, 13(7):3423–3432.

José Almeida Neto and Tiago de Melo. 2023. Exploring supervised learning models for multi-label text classification in brazilian restaurant reviews. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 126–140.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Bassem Bsir and Mounir Zrigui. 2018. Bidirectional lstm for author gender identification. In *International Conference on Computational Collective Intelligence*, pages 393–402. Springer.

Alberto F Cabrera and 1 others. 1994. Logistic regression analysis in higher education: An applied perspective. *Higher education: Handbook of theory and research*, 10:225–256.

Corinna Cortes and Vladimir Vapnik. 1995. Support vector machine. *Machine learning*, 20(3):273–297.

Washington Cunha, Leonardo Rocha, and Marcos André Gonçalves. 2025. A thorough benchmark of automatic text classification: From traditional approaches to large language models. *arXiv preprint arXiv:2504.01930*.

Miguel de Oliveira and Tiago de Melo. 2021. An empirical study of text features for identifying subjective sentences in portuguese. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 374–388. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Rafael Dias and Ivandré Paraboni. 2020. Cross-domain author gender classification in brazilian portuguese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1227–1234.

José Antonio García-Díaz, Ghassan Beydoun, and Rafel Valencia-García. 2024. Evaluating transformers and linguistic features integration for author profiling tasks in spanish. *Data & Knowledge Engineering*, 151:102307.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Sanja Hanić, Marina Bagić Babac, Gordan Gledec, and Marko Horvat. 2024. Comparing machine learning models for sentiment analysis and rating prediction of vegan and vegetarian restaurant reviews. *Computers*, 13(10):248.

Supanat Jintawatsakoon and Ekkapob Poonsawat. 2023. Gender classification of social network text using natural language processing and machine learning approaches. In *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 415–420. IEEE.

Katikapalli Subramanyam Kalyan. 2024. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22.

Gustavo Lopes and Helena Oliveira. 2023. A cascade approach for gender prediction from brazilian portuguese tweets. In *RecSys*.

Paula Luegi, Márcio Leitão, Daniela Avila-Varela, Jéssica Gomes, and Armanda Costa. 2024. Reflexive pronoun resolution in portuguese: testing similarity-based interference. *Frontiers in Language Sciences*, 3:1473948.

Abdul Majeed and Sungchang Lee. 2021. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access*, 9:8512–8545.

Oluwakemi Onikoyi and Amina Adamu. 2023. Improving gender classification in low-resource hausa with profile metadata. In *AfriNLP*.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Francisco Rangel, Anastasia Giachanou, Bilal Hisham Hasan Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CEUR workshop proceedings*, volume 2696, pages 1–18. Sun SITE Central Europe.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Bertimbau: Pretrained bert models for brazilian portuguese. In *ICAI*.

Chanchal Suman, Rohit Shyamkant Chaudhary, Sriparna Saha, and Pushpak Bhattacharyya. 2022. An attention based multi-modal gender identification system for social media users. *Multimedia Tools and Applications*, 81(19):27033–27055.

Pradeep Vashisth and Kevin Meehan. 2020. Gender classification using twitter text data. In *2020 31st Irish Signals and Systems Conference (ISSC)*, pages 1–6.

Claudimar Pereira da Veiga, Cássia Rita Pereira da Veiga, Júlia de Souza Silva Michel, Leandro Ferreira Di Iorio, and Zhaohui Su. 2024. E-commerce in brazil: An in-depth analysis of digital growth and strategic approaches for online retail. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(2):1559–1579.

Shikha Verma. 2019. Weapons of math destruction: how big data increases inequality and threatens democracy. *Vikalpa*, 44(2):97–98.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.