Gendered Stylistic Variation in Brazilian Portuguese Google Play Reviews: A Large-Scale Study

Tiago de Melo
Amazon State University (UEA)
Manaus, Brazil
tmelo@uea.edu.br

Abstract

We study gender-associated stylistic variation in Brazilian Portuguese Google Play reviews. Using IBGE name frequencies, we infer binary gender from first names in 76.7M reviews (96 apps, 20112025), obtaining 22.25M high-confidence labels. Women-associated reviews show markedly higher paralinguistic expressivity (about 60% higher emoji density and more lengthening/punctuation), while lexical diversity (MTLD) is nearly identical across groups. Ratings are mostly positive, with men contributing relatively more 1-star reviews and women more 5-star reviews. These findings contribute to a deeper understanding of digital sociolinguistic behavior within the Brazilian context. We discuss limitations of name-based gender inference and future demographic extensions.

1 Introduction

The Google Play Store serves as a central platform where millions of users express their opinions and experiences through reviews of apps. These comments are a valuable source of feedback for developers, providing signals about bugs, feature requests, and the perceived user experience (Pagano and Maalej, 2013). However, the way feedback is communicated is not uniform and may vary according to sociodemographic factors.

Previous work in the international literature (Noei et al., 2018) suggests that gender is associated with participation and writing behavior in app stores, revealing differences in sentiment, posting frequency and topics of interest. However, most studies focus on English-language corpora and there remains a scarcity of large-scale analyzes targeted at Brazilian Portuguese. Given the lexical richness and particularities of computer-mediated communication in Brazilian digital culture (Vieira et al., 2022), investigating how men and women use written language to evaluate apps is important for a more refined understanding of local user behavior (Guedes et al., 2016; Noei and Lyons, 2022).

This paper presents a large-scale study of gender differences in Brazilian Portuguese Google Play reviews. We collected and organized a dataset with 76,695,564 reviews from 96 popular apps spanning 2011–2025. Because Google Play does not provide explicit gender metadata, we implemented a statistical inference procedure based on users' first names, using frequency data from the 2010 Brazilian Census (IBGE) as reference. This strategy yields a representative sample of 22,254,455 reviews with high-confidence gender labels (*male* or *female*) for subsequent linguistic analyzes.

Our investigation focuses on three complementary dimensions: (i) stylometric analysis, emphasizing expressivity markers and text length; (ii) lexical diversity, measured through the *Measure of Textual Lexical Diversity* (MTLD); and (iii) star-rating distributions. Consequently, we address the following research questions:

RQ1: How do writing styles differ between reviews associated with men and women, particularly regarding paralinguistic expressivity markers in Brazilian Portuguese?

RQ2: To what extent do genders differ in lexical diversity as measured by MTLD?

RQ3: Are there systematic differences in the distribution of star ratings (1–5) between reviews associated with men and women, especially at the extremes?

The results indicate that while the typical length of the review is similar between genders, female-labeled reviews are markedly more expressive, with a 60% higher emoji density and a higher incidence of vowel elongation and repeated punctuation. We also find that reviews with female marks are slightly more positive overall, concentrating on a higher

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 238–246

proportion of 5-star ratings, whereas reviews with male marks contain a higher proportion of 1-star ratings.

We make four main contributions. First, we assemble and describe a large-scale Brazilian Portuguese Google Play review dataset with transparent provenance and language filtering. Second, we provide a stylometric comparison across more than 22 million gender-labeled reviews, quantifying differences in paralinguistic expression cues. Third, we apply MTLD at scale and show that lexical diversity is virtually equivalent across genders despite robust differences in expressivity. Fourth, we characterize gendered differences in star-rating distributions, emphasizing consistent gaps at the extremes (1 and 5 stars).

2 Related Work

2.1 Name-based gender inference

Inferring gender from first names is a common strategy in large-scale observational studies, especially when gender is not available as structured metadata. Reference services and datasets typically associate names with frequency distributions by sex, enabling probabilistic assignments or threshold-based labeling. Santamaría and Mihaljevi (Santamaría and Mihaljević, 2018) compare multiple name–gender inference services and discuss performance, coverage, and biases arising from cultural and linguistic variation as well as thresholding choices. Complementarily, Mihaljevic et al. (Mihaljević et al., 2019) highlight conceptual and external-validity limitations of such inference and recommend methodological transparency, uncertainty reporting, and caution to avoid undue causal interpretations.

In the Brazilian context, IBGE's *Nomes no Brasil* project provides a name database derived from the 2010 Demographic Census, including occurrence frequencies by gender and geographic/temporal breakdowns. In its official description, IBGE reports 130,348 distinct names observed, counted separately for male and female, and clarifies that only the first name was considered (IBGE, 2016). The resource also documents practical details relevant to inference: orthographic variants are treated as distinct entries (e.g., "Ana" vs. "Anna"), diacritics are not represented, and the sex associated with records reflects what was declared at census time (IBGE, 2016). Recent work in Portuguese NLP has also explored stylometric traits in tasks such as plagiarism detection and authorship attribution, reinforcing the value of style metrics (Uka and Berger, 2024).

Unlike studies centered on academic authorship or scientific participation, this work applies name-based inference at scale to Brazilian Portuguese Google Play reviews, propagating the label inferred from the user's name to each review. Our design favors conservative labeling, explicitly preserving *ambiguous* and *unknown* cases; comparative analyzes focus on the most reliable subsets, aiming to characterize stylistic differences without implying causal interpretations.

2.2 Gender in app reviews

Software engineering and repository-mining research have used app-store reviews as signals of opinion, usability, and perceived quality, sometimes incorporating demographic slices when possible (Dąbrowski et al., 2022). In particular, Noei and Lyons (Noei and Lyons, 2022) analyze gender in Google Play reviews using name-based inference and report differences in participation and rating patterns. Such studies motivate complementary investigations in other languages and settings, as name coverage, nickname practices, and writing conventions can vary substantially across communities.

In contrast to analyzes focused on maintenance and engineering aspects, we concentrate on linguistic–stylometric and expressivity metrics in Brazilian Portuguese. We preserve uncertainty (*ambiguous*/*unknown*) and restrict comparisons to high-confidence gender labels to reduce classification bias and support reproducible interpretations.

3 Methodology

3.1 Data collection

Our dataset was collected from reviews posted on Google Play on a diverse set of apps. We selected 96 applications (apps) by popularity and restricted the corpus to Brazilian Portuguese. Language filtering relied on the Google Play metadata field `lang`, keeping only records tagged as `pt-BR`.

3.2 Gender inference

To enable gender-aware analyzes, we infer the probable gender of each review author from the `userName` field in the metadata. The procedure has three steps: (i) extraction and normalization of the name, (ii) extraction of the first-name, and

(iii) labeling into four categories: *male*, *female*, *ambiguous*, and *unknown*.

First, all `userName` values are normalized (e.g., trimming redundant spaces and standardizing case) to reduce superficial variation. We then extract the first alphabetic token of `userName` as a proxy for the first name. This first name is queried against a public IBGE-derived database[1] (2010 Demographic Census), which provides occurrence frequencies by sex, denoted `freq_f` (female) and `freq_m` (male). We rely on the 2010 Census dataset as it remains the most recent comprehensive public resource for name frequency distributions in Brazil. Although a new census was conducted in 2022, equivalent microdata of name-frequency had not been released at the time of this study.

Based on these frequencies, we define the probability of femaleness as follows:

$$p_{\text{fem}} = \frac{\texttt{freq\_f}}{\texttt{freq\_f} + \texttt{freq\_m}}.$$

Similarly, the probability of men is $p_{\text{male}} = 1 - p_{\text{fem}}$. When the first name is missing from the IBGE database, or when $\texttt{freq\_f} + \texttt{freq\_m} = 0$, p_{fem} cannot be reliably estimated.

Labeling is conservative and threshold-based. If $p_{\text{fem}} \geq 0.95$, the name is labeled as *female*; if $p_{\text{fem}} \leq 0.05$, it is labeled as *male*. For $0.05 < p_{\text{fem}} < 0.95$, the name is considered *ambiguous*, as it occurs substantially in both sexes. We assign *unknown* when gender cannot be inferred reliably, particularly for names absent from the IBGE database (nicknames, idiosyncratic spellings, non-name strings, or foreign names) or when a valid `userName` is not available.

After inference on `userName`, the label is propagated to each review, enabling gender-stratified subsets for linguistic analyzes. Name-based inference is a statistical approximation and does not represent self-declared gender identity; therefore, we keep the *ambiguous* and *unknown* categories for transparency and sensitivity analyzes.

3.3 Linguistic analyses

To characterize stylistic differences between genders, we performed three complementary analyzes. In the stylometric analysis, we compute text-format and expressivity metrics, including review length (number of tokens per review: mean, median, and 90th percentile), and paralinguistic expressivity (emoji density per 100 tokens, proportion of reviews containing at least one emoji, vowel elongation defined as repetition of the same letter three or more times, and repeated punctuation such as "!!!" or "???"), and emphasis markers such as uppercase ratio (excluding sentence-initial capitalization). We also compute the incidence of politeness markers, laughter markers, and a small inventory of informal tokens.

Our stylometric metrics are based on established work in style analysis and author profile (Koppel et al., 2002; Rangel et al., 2017), as well as studies focused on Brazilian Portuguese (Dias and Paraboni, 2020). Table 1 defines and illustrates the metrics. Politeness markers, laughter markers, and informal tokens were defined using an *ad hoc* lexicon that will be released with the dataset.

Second, lexical diversity is measured with the *Measure of Textual Lexical Diversity* (MTLD), using the threshold $\tau = 0.72$ for each gender, following the standard value in the literature (McCarthy and Jarvis, 2010). We compute MTLD by concatenating all reviews within each group (male and female), which reduces corpus-size effects in comparisons and yields a deterministic and reproducible procedure.

Third, we analyze emojis beyond overall density by examining group-specific usage patterns. We computed the relative frequency of each emoji (percentage over the total emojis within a group), repertory overlap between groups using Jaccard similarity over the sets of the 50 most frequent emojis, $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, and a functional categorization based on the taxonomy of Barbieri et al. (Barbieri et al., 2016) (e.g., *Affect/Emotion*, *Approval/Evaluation*, *Reaction/Emphasis*).

For emojis with marked differences between groups, the strength of the association is measured using the *odds ratio* (OR), defined as the ratio between the odds of emoji occurrence in the female group and in the male group. Let a be the number of occurrences of the emoji in the female group, b the number of occurrences of other emojis in the female group, c the number of occurrences of the emoji in the male group, and d the number of occurrences of other emojis in the male group. Then:

$$\text{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

We compute 95% confidence intervals for OR using Woolf's approximation (Woolf, 1955) in log

[1] https://brasil.io/dataset/genero-nomes

Table 1: Definition and computation of stylometric metrics.

Metric	Computation example	Description
Mean tokens	Reviews: "*Gostei muito!*" (3 tokens) and "*Não curti o filme.*" (4 tokens). Mean = (3 + 4) / 2 = 3.5.	Total number of tokens divided by the number of reviews.
Median tokens	Token counts: 2, 3, 3, 10, 15. Median = 3.	Middle value after sorting by token count.
90th percentile tokens (P90)	If 90% of reviews have up to 20 tokens, then P90 = 20.	Token-count threshold below which 90% of reviews fall.
Emojis per 100 tokens	Review: "*Amei este app* 😍 😍" (2 emojis, 3 tokens). Thus, $2/3 \times 100 = 67$.	Total emojis divided by total tokens, multiplied by 100.
Reviews with emoji (%)	3 reviews with emojis out of 10 reviews = 30%.	Percentage of reviews that contain at least one emoji.
! per 100 tokens	–	Number of "!" per 100 tokens.
? per 100 tokens	–	Number of "?" per 100 tokens.
Uppercase ratio	5 uppercase letters / 10 letters = 0.5.	Uppercase letters divided by total letters.
Repeated punctuation (%)	Reviews containing "!!!" or "???".	Percentage of reviews with repeated punctuation.
Elongation (%)	Tokens such as "oooi" and "genteeee".	Percentage of reviews containing elongated letters.
Politeness markers (%)	"*por favor*" and "*obrigada*".	Percentage of reviews containing courtesy expressions.
Laughter markers (%)	"kkk", "haha", "rsrs".	Percentage of reviews containing laughter expressions.
Informal tokens (mean)	"tbm", "vc", "pq", "blz".	Average number of informal tokens per review.

space:

$$\mathrm{CI}_{95\%} = \exp\left(\ln(\mathrm{OR}) \pm 1.96 \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right).$$

Finally, we compare star-rating distributions (1–5) between genders using within-group proportions, mean rating per group, and percentage-point differences, focusing on extremes (1 and 5 stars). All analyzes in this section are restricted to high-confidence gender labels (male and female), excluding *ambiguous*, *unknown*, and records without `userName`, as described in Section 3.2.

4 Results

4.1 Dataset

Our dataset comprises 76,695,564 reviews spanning 2011–2025. For gender-aware analyzes, we apply the inference procedure described in Section 3.2, which assigns a gender label to each `userName` and propagates it to the corresponding reviews. This procedure enables us to characterize (i) the population of user names (number of distinct `userName` values and first-name diversity) and (ii) the distribution of reviews by gender. We follow best practices in corpus-construction regarding provenance tracking and typological organization (Sturzeneker et al., 2022).

Table 2 summarizes the counts. In this work, "UserNames" refers to the number of distinct user names labeled by the procedure, "Unique first names" is the number of distinct first names in each category, and "Reviews" is the total number of reviews associated with each category after label propagation from `userName`. There are also reviews with missing `userName`, for which inference does not apply; we report those separately.

All analyzes in this paper consider only the 22,254,455 reviews for which we can assign high-confidence male or female labels (12,917,778 male and 9,336,677 female), excluding the ambiguous and unknown groups and records without `userName`. Although this subset represents only part of the overall crawl, more than 22 million confidently labeled reviews provide a highly representative sample for large-scale linguistic analyzes.

Table 2: Dataset summary and distribution by inferred gender.

Category	UserNames	UserNames (%)	Unique first names	Reviews	Reviews (%)
male	5,015,453	40.30	30,545	12,917,778	16.84
female	3,769,798	30.29	35,718	9,336,677	12.17
ambiguous	1,766,475	14.19	6,001	3,141,146	4.10
unknown	1,894,009	15.22	712,888	36,325,898	47.36
no `userName`	–	–	–	14,974,064	19.52
Total	12,445,735	100.00	785,152	76,695,564	100.00

4.2 Stylometric analysis

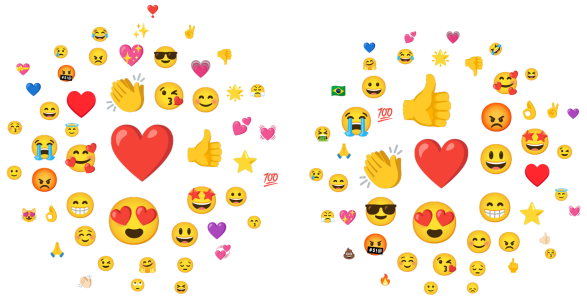
To address RQ1, we report a stylometric analysis of reviews associated with the *male* and *female* groups. Table 3 summarizes the overall differences across metrics. Regarding length, female reviews are slightly longer on average (9.54 vs. 8.89 tokens), while the median is identical (4 tokens) and the 90th percentile is very close (24 vs. 23). This pattern suggests that the length gap is not driven by the typical review (short in both groups), but by a modest and consistent increase in the upper tail, i.e., among longer reviews.

The clearest contrast emerges in the expressivity markers. Female-labeled reviews have a higher emoji density (11.57 vs. 7.17 per 100 tokens) and a higher proportion of reviews with at least one emoji (14.52 vs. 8.96%), with relative gaps above 60%. This aligns with a more expressive style, also visible in the higher incidence of elongation (5.28% vs. 3.04%) and repeated punctuation (2.97% vs. 2.34%), which are commonly associated with emphasis and pragmatic intensification in computer-mediated communication. Male-labeled reviews show a slightly higher uppercase ratio, indicating a marginal preference for all-caps emphasis. The rate of question marks per 100 tokens is very similar across groups, suggesting that question usage is not a discriminative factor in this dataset.

Although we observe small differences in politeness and laughter markers, these effects are low in magnitude and should be interpreted cautiously. Overall, our results indicate that the most robust gender-associated differences concentrate on paralinguistic and intensification cues (emojis, elongation, repeated punctuation), while length-related measures vary only modestly. These findings support the hypothesis that, in Google Play reviews, stylistic variation by gender manifests more strongly in expressivity than in content quantity, motivating finer-grained linguistic analyzes (e.g., lexical and syntactic patterns) in future work.

4.3 Emoji usage by gender

Complementing the stylometric analysis in Section 4.2 (RQ1), we examine the patterns of usage of emoji. Figure 1 shows word clouds for the 50 most frequent emojis in each group.



(a) Female (b) Male

Figure 1: Most frequent emoji clouds for female- and male-labeled reviews. Emoji size is proportional to its frequency within the corresponding group.

We observe substantial overlap in both groups' emoji repertoires, dominated by positive-valence symbols. Nevertheless, quantitative analyzes reveal systematic differences in relative frequencies and composition among the most used emojis. Table 4 lists the top five emojis in each group, with their percentages over the group's total emoji count.

The overlap between the top-50 emojis (by absolute frequency) was measured using Jaccard similarity, yielding $J = 0.82$ (approximately 45 emojis in common). This indicates substantial repertory overlap. However, usage proportions differ significantly. For example, ❤️ is 1.71 times more frequent in the female group (OR = 1.71; 95% CI: [1.70–1.72]), while 👍 is 1.84 times more frequent in the male group (OR = 1.84; 95% CI: [1.83–1.85]).

For a finer analysis, we classify emojis using a functional taxonomy adapted from Barbieri et al. (Barbieri et al., 2016). Among the 20 most frequent emojis in each group, 68% in the female group belong to the Affect/Emotion category (e.g.,

Table 3: Stylometric metrics for reviews labeled as male and female.

Metric	Male	Female	Δ (F–M)	Δ%
N (reviews)	12.917.717	9.336.646	–	–
Mean tokens	8.89	9.54	0.65	7.35
Median tokens	4	4	0	0.00
P90 tokens	23	24	1	4.35
Emojis per 100 tokens	7.17	11.57	4.40	61.35
Reviews with emoji (%)	8.96	14.52	5.55	61.95
! per 100 tokens	3.52	3.97	0.45	12.74
? per 100 tokens	0.204	0.195	-0.009	-4.37
Uppercase ratio	0.0220	0.0190	-0.0030	-13.72
Repeated punctuation (%)	2.34	2.97	0.63	26.90
Elongation (%)	3.04	5.28	2.23	73.28
Politeness markers (%)	2.07	2.20	0.13	6.45
Laughter markers (%)	0.63	0.67	0.04	6.53
Informal tokens (mean)	0.0977	0.1105	0.0128	13.10

Table 4: Top 5 emojis by gender (percentage over total emojis within the group).

Male			**Female**		
Pos.	Emoji	% (M)	Pos.	Emoji	% (F)
1	❤️	7.89	1	❤️	12.76
2	👍	7.85	2	😍	8.60
3	😍	5.53	3	👋	4.51
4	👋	4.95	4	👍	4.43
5	😃	2.87	5	🥰	2.60

❤️, 😍, 🥰, 😘), compared to 40% in the male group. The male group, in turn, uses proportionally more Approval/Evaluation emojis (👍, 👋, ⭐) and Reaction/Emphasis emojis (😡, 😎, 🔥). Table 5 reports the emojis with the largest percentage-point differences between groups.

Table 5: Emojis with the largest percentage-point differences between genders.

Emoji	Category	% Male	% Female	Δ (p.p.)
❤️	Affect/Emotion	7.89	12.76	+4.87
😍	Affect/Emotion	5.53	8.60	+3.07
👍	Approval	7.85	4.43	–3.42
🥰	Affect/Emotion	1.53	2.60	+1.07
😘	Affect/Emotion	1.58	2.56	+0.98
😎	Reaction	2.65	1.50	–1.15
😡	Reaction/Negative	2.67	1.88	–0.79

These findings complement the expressivity patterns in Section 4.2. Differential emoji usage suggests distinct discourse strategies: while female-labeled reviews rely more on affective engagement and emotional intensification, male-labeled reviews tend to emphasize approval, evaluation, or dissatisfaction. This pattern aligns with the classic work on gender in computer-mediated communication (Savicki and Kelley, 2000) and reinforces the importance of paralinguistic resources (emojis, elongation, repeated punctuation) as part of stylistic variation.

Importantly, these are large-scale tendencies, rather than linguistic determinisms. The high repertory overlap (Jaccard = 0.82) indicates that both groups share a broad set of symbols, varying their frequency and preference according to the pragmatic context.

4.4 Lexical diversity (MTLD)

To address RQ2, we evaluated the lexical diversity using MTLD. MTLD estimates the average segment length (in tokens) needed for the type–token ratio (TTR) of successive segments to fall below a threshold τ; higher values indicate greater lexical diversity. We compute MTLD with $\tau = 0.72$ for each gender following standard practice (McCarthy and Jarvis, 2010), concatenating all tokens in each group (*male* and *female*, according to Section 3.2). This reduces corpus-size effects in comparisons, while keeping the procedure deterministic and reproducible.

Overall, MTLD values are very close across groups: $\mathrm{MTLD}_M = 72.98$ and $\mathrm{MTLD}_F = 72.13$ ($\Delta = -0.85$, i.e., -1.17% for females relative to males), despite the large number of tokens analyzed (115.9M for males and 89.8M for females). Practically, this suggests that the average lexical diversity is broadly similar across genders, with

only a marginal advantage for the male group under this metric. Figure 2 shows the yearly stratification (2020–2025), the period that concentrates most gender-inferred reviews, revealing similar temporal trajectories and a sharp decline in 2022 for both groups, followed by partial recovery.

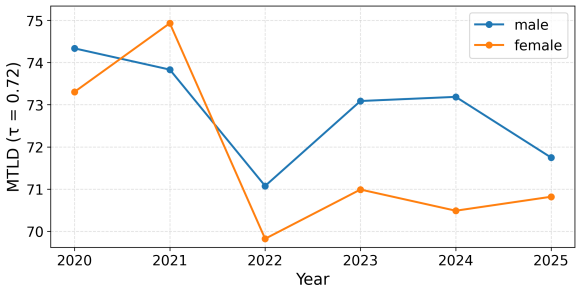


Figure 2: MTLD by year (2020–2025) for male and female groups ($\tau = 0.72$).

The 2022 dip is consistent with a shift in corpus composition (e.g., concentration of reviews in specific apps, topics, or events) and/or increased lexical conventionalization (more recurring vocabulary and formulaic evaluation patterns), which reduces diversity even without changes in typical length. As an observational slice, interpretation must be cautious. The dip may reflect factors external to gender (changes in the set of applications, user profiles, or the dynamics of the platform). Together with stylometric findings, the MTLD results indicate that clearer differences in expressivity markers (e.g., emojis and elongation) do not necessarily imply large discrepancies in lexical diversity.

4.5 Star-rating distributions

To address RQ3, we evaluated gender differences in star-rating distributions. Figure 3 compares the rating distribution (1–5 stars) between reviews labeled with male and female. For each group, we keep only reviews with a valid `score` in $\{1, 2, 3, 4, 5\}$ and a high-confidence gender label (Section 3.2). We then counted the number of reviews at each star level and normalized by the total number of reviews in that group, generating percentages within the group that sum up to 100%. In total, we analyze 12,917,778 reviews in the male group and 9,336,677 in the female group.

Both groups are strongly concentrated on positive ratings, with a predominance of 5 stars (72.07% for males and 74.21% for females). Still, the differences are consistent at the extremes: the male group has a higher fraction of 1-star ratings (13.84% vs. 10.91%, +2.93 p.p.), while the female group has a higher fraction of 5-star ratings (74.21% vs. 72.07%, +2.14 p.p.). These gaps also reflect in the mean score, which is slightly higher for females (4.316 vs. 4.213). In the intermediate range (2–4 stars), discrepancies are small, suggesting that the main distinction between groups occurs primarily between strongly negative (1 star) and strongly positive (5 stars) evaluations.

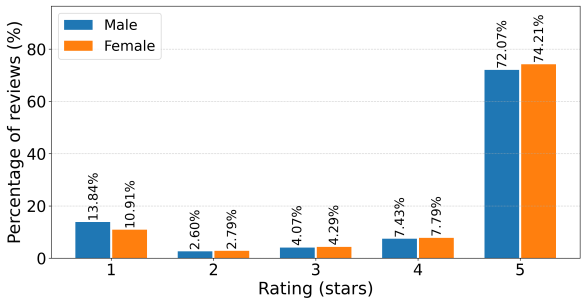


Figure 3: Star rating distribution by gender.

5 Limitations and Ethical Considerations

Interpretation should remain cautious because this is an observational study and gender is inferred from first names. Confounders such as app mix, usage profiles, and temporal effects may contribute to the observed pattern, and results should not be read as causal.

Despite the promising results, this work has limitations that warrant careful discussion, particularly regarding the ethical issues surrounding gender inference. We emphasize that gender inference in this study is a statistical approximation based on observed patterns of name-frequency. It does not capture the complexity, fluidity, or self-identification of gender (Keyes, 2018). Gender is a social and personal construct that goes beyond binary classifications or algorithmically inferred attributes. Ignoring this distinction can oversimplify identities and reinforce gender stereotypes. Therefore, any use of our analysis should be done cautiously, explicitly acknowledging these limitations.

6 Conclusions and Future Work

Our results indicate that the most consistent differences between the reviews labeled with male and female are focused on the expressivity markers. Female-labeled reviews show higher emoji density and incidence and more intensification signals, such as elongation and repeated punctuation. In contrast, global measures of content and vocabulary

are more similar. Typical review length is comparable across groups and lexical diversity measured by MTLD is virtually stable, suggesting that stylistic variation does not necessarily imply large discrepancies in lexical richness. Star ratings are mostly positive in both groups, but differ at the extremes. The male group concentrates more 1-star ratings, whereas the female group has a higher fraction of 5-star ratings and a slightly higher mean rating.

Future work follows three directions. First, improve and audit name-based gender inference by quantifying uncertainty, assessing coverage biases (e.g., names outside the reference list and nicknames), performing sample-based validations to estimate error rates, and conducting sensitivity analyzes that incorporate part of the *ambiguous* and *unknown* subsets. Second, refine the observational analysis by controlling for confounders, for example, via stratification by app/category, matching by activity period, and temporal slicing that separates composition effects (apps and usage profiles) from linguistic differences. Third, broaden the linguistic analyzes with finer-grained measurements of lexical choice and structure (e.g., semantic/affective categories, syntactic patterns, negation, and intensifiers) as well as emoji functional categories and co-occurrence with ratings, deepening the characterization of Brazilian Portuguese stylistic variation while maintaining methodological transparency about limitations.

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

Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 531–535.

Jacek Dąbrowski, Emmanuel Letier, Anna Perini, and Angelo Susi. 2022. Analysing app reviews for software engineering: a systematic literature review. *Empirical Software Engineering*, 27(2):43.

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.

Gustavo Paiva Guedes, Eduardo Bezerra, Lilian Ferrari, and Fellipe Duarte. 2016. Gender differences in the use of portuguese in social networks: Evidence from liwc. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 339–342.

IBGE. 2016. Um brasil de marias e josés: Ibge apresenta banco de nomes com base no censo 2010. Agência de Notícias do IBGE. Acesso em: 2026-01-13.

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.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Helena Mihaljević, Marco Tullney, Lucía Santamaría, and Christian Steinfeldt. 2019. Reflections on gender analyses of bibliographic corpora. *Frontiers in big Data*, 2:29.

Ehsan Noei, Daniel Alencar Da Costa, and Ying Zou. 2018. Winning the app production rally. In *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 283–294.

Ehsan Noei and Kelly Lyons. 2022. A study of gender in user reviews on the google play store. *Empirical Software Engineering*, 27(2):34.

Dennis Pagano and Walid Maalej. 2013. User feedback in the appstore: An empirical study. In *2013 21st IEEE international requirements engineering conference (RE)*, pages 125–134. IEEE.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, 48.

Lucía Santamaría and Helena Mihaljević. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156.

Victor Savicki and Merle Kelley. 2000. Computer mediated communication: Gender and group composition. *CyberPsychology & Behavior*, 3(5):817–826.

Mariana Sturzeneker, Maria Clara Crespo, Maria Lina Rocha, Marcelo Finger, Maria Clara Paixão de Sousa, Vanessa Martins do Monte, and Cristiane Namiuti. 2022. Carolina's methodology: building a large corpus with provenance and typology information. In *DHandNLP@ PROPOR*, pages 53–58.

Adile Uka and Maria Berger. 2024. Could style help plagiarism detection?-a sample-based quantitative study of correlation between style specifics and plagiarism. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 103–108.

Renata Vieira, Ana Paula Banza, Ana Sofia Ribeiro, Cassia Trojahn, Fernanda Olival, Helena Cameron, Herminia Vilar, Ivo Santos, Joaquim Santos, Maria Gonçalves, and 1 others. 2022. Digital humanities and portuguese processing: a research pathway.

Barnet Woolf. 1955. On estimating the relation between blood group and disease. *Annals of human genetics*, 19(4):251–253.