Beyond Systematic Bias: Investigating Gender Differences in Portuguese Text Classification Annotation Patterns

Alexander Feitosa, Érica Carneiro, Gustavo Guedes

CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Brazil
alexander.feitosa@aluno.cefet-rj.br,
ericacqueiroz@gmail.com, gustavo.guedes@cefet-rj.br

Abstract. This study explores how gendered annotation patterns influence sentiment classification in Brazilian Portuguese and whether these patterns are preserved—or amplified—by machine learning models. A corpus of 1,465 diary-style sentences was independently labeled by two gender-balanced annotator groups. Despite a high final agreement, as indicated by Cohen's $\kappa = 0.8177$, statistical analyses revealed divergent annotation behaviors: male annotators exhibited higher internal consistency and lower entropy, while female annotators showed greater variability and a higher proportion of neutral labels. Classifiers trained separately on each group's labels (SVM, LR, NB, RF, DT) reproduced these divergences to varying degrees. Notably, alignment between gendered models dropped to $\kappa = 0.3838$ (Decision Tree) and peaked at $\kappa = 0.6952$ (Logistic Regression), indicating that learning behaviors may differ substantially based on the annotation source. These findings reinforce that annotation is a socially based process. Gendered interpretive divergences can propagate through learning pipelines, shaping model behavior in ways that reflect and potentially amplify gender bias, often going unnoticed without annotation-aware evaluation strategies. Ethical approval was granted under protocol CAAE 82267824.8.0000.5289.

CCS Concepts: • **Computing methodologies** → **Natural language processing**.

Keywords: annotation behavior, gender bias, sentiment classification, supervised learning, fairness in NLP

1. INTRODUCTION

Text classification remains a cornerstone of supervised learning in Natural Language Processing (NLP), enabling structured interpretation of subjective or ambiguous language data. The consistency and accuracy typically associated with supervised models make them a good fit for machine learning applications, even though it can be affected by the involvement of human judgment [Kowsari et al. 2019]. A recent survey reviews over 300 papers on gender bias in NLP, highlighting persistent issues such as binary gender assumptions and a lack of multilingual research focus [Stańczak and Augenstein 2021]. Additionally, [Sun et al. 2019] provided an extensive literature review of gender bias in NLP, categorizing forms of representation bias and mitigation strategies.

However, the systematic investigation of how these biases manifest during the annotation process, when social perspectives become embedded in training data, remains largely unexplored, particularly in sentiment classification tasks. This gap is especially critical in Portuguese-language contexts, where the language's rich system of grammatical gender increases the likelihood of gendered cues influencing annotation. Yet, empirical studies in this area remain scarce.

In this work we focus specifically on sentiment annotation, assigning sentiment labels to a shared corpus of Brazilian Portuguese sentences with one of three polarity labels—negative (-1), neutral (0) or positive (+1). In addition to comparing the final label distributions, we examine levels of agreement, consistency, and labeling entropy within and between the groups to understand how annotation behavior may reflect interpretive differences that influence model learning.

The structure of this article is outlined as follows. Section 2 offers an overview of previous research concerning gender bias in natural language processing. Section 3 details the methodological framework adopted in the present study. Section 4 presents the results of the experimental evaluation. Finally,

Section 5 concludes the article by summarizing the main findings, discussing the study's limitations, and proposing directions for future research.

2. RELATED WORK

The origins of bias in NLP systems have been critically examined by [Blodgett et al. 2020], who argue that biased annotations can be embedded in datasets from their inception. Similarly, [Lim et al. 2024] show that annotation bias can be systematically measured and mitigated using causal mediation analysis, suggesting that the annotation phase warrants as much attention as model selection or evaluation.

[Shah et al. 2020] frame predictive bias as a systematic outcome of interactions between model architectures and latent dataset features, which aligns with our findings on divergences observed across annotator groups. This phenomenon of bias amplification has been empirically demonstrated in NLP tasks such as co-reference and sentiment analysis [Zhao et al. 2017], where models trained on biased data exaggerated demographic associations beyond the original corpus. To mitigate bias in sentiment models, [Davani et al. 2022] proposed the use of post-hoc explanations to expose decision pathways and detect annotation-driven discrepancies. Our study complements this approach by examining bias emergence during training rather than post-prediction.

The gendered dimension of annotation is particularly prominent in Portuguese, where grammatical gender permeates the language. Gendered linguistic patterns influence both how annotators interpret text and how models internalize these interpretations [Silva and Moro 2024]. Any shift in the grammatical system requires collective clarity about the semantic scope of new forms and recognition that their formal representation inevitably imposes categorical boundaries [Schwindt 2020]. [Geva et al. 2019] further shows that annotator demographics such as gender, age, and cultural background can systematically affect labeling outcomes, reinforcing the need for studies in languages other than English.

Beyond the statistical lens, some scholars have called for a broader epistemological shift in how annotations are understood. Rather than treating disagreement as noise to be minimized, [Havens and Hedges 2022] argue for acknowledging uncertainty as an inherent feature of human labeling behavior. These perspectives resonate with broader critiques of dataset construction practices, which emphasize how development choices, label guidelines, and annotator identities all contribute to shaping the social and technical dimensions of learning pipelines [Paullada et al. 2021].

In line with this perspective, recent work has explored the propagation of discrimination in classification tasks by analyzing dataset stratification and label distributions [Minatel et al. 2023]. [Raji et al. 2021] further argue that evaluation benchmarks themselves embed normative decisions, and that understanding bias requires attention to the full data lifecycle — from annotation to deployment. These efforts align with surveys that consolidate known sources of algorithmic bias, including data, modeling assumptions, and feedback loops, as comprehensively outlined by [Mehrabi et al. 2021].

Our study contributes to this evolving discourse by comparing male and female annotator behaviors in Portuguese sentiment annotation and evaluating the extent to which predictive models trained on each group's labels replicate the observed divergences.

3. METHODOLOGY

This section describes the pipeline adopted to: (i) measure agreement between gender-balanced annotators and (ii) evaluate whether text-classification models trained on gender-specific labels reproduce or amplify annotation biases. Figure 1 presents an overview of the steps proposed by our study.

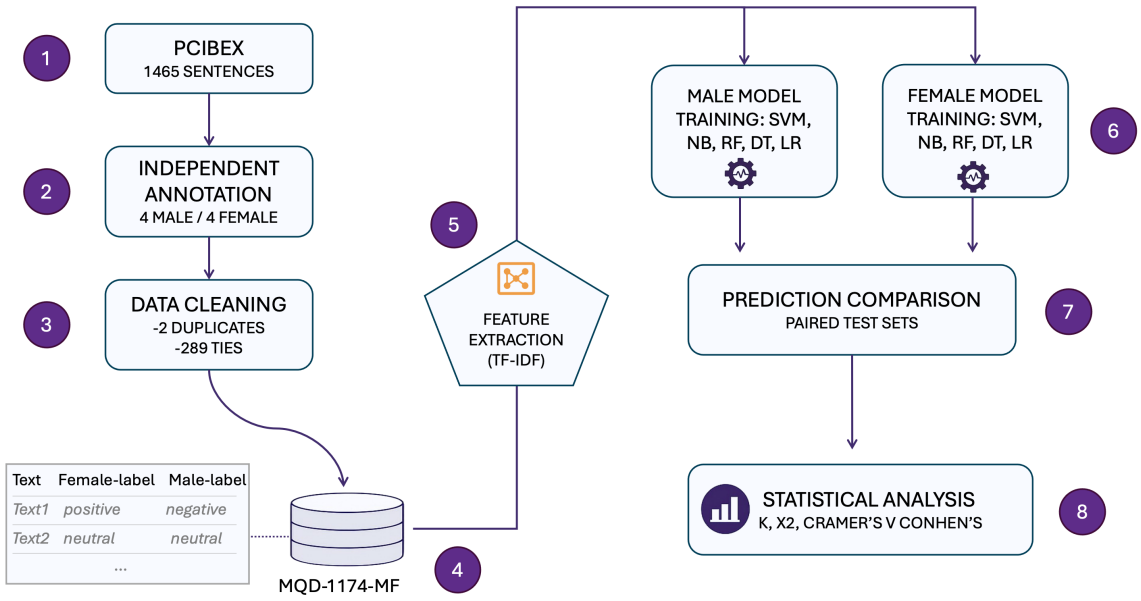

Fig. 1: Gender-balanced annotators and trained models text-classification

3.1 Dataset and Participants

The dataset used in this study, named MQD-1465 [Azevedo et al. 2021], comprises 1,465 Brazilian Portuguese sentences harvested from the website "Meu Querido Diário", an online diary that collected a wide variety of everyday testimonials submitted under anonymity. The dataset retained its original form — including spelling mistakes and expressions that might be seen as uncomfortable — in order to preserve the authenticity and natural tone of the material. This decision was intentional and aligned with the research's exploratory character.

Beyond ethical approval, granted under Plataforma Brasil protocol CAAE 82267824.8.0000.5289, participants provided informed consent by agreeing to the terms outlined in the Informed Consent Form (ICF) prior to their involvement. The experiment involved the deployment of these 1,465 sentences on the PCIbex Farm platform[1] (Fig. 1, Step 1). Eight native speakers (four female and four male) independently labeled every sentence with a three-point polarity scale: *negative* (−1), *neutral* (0) and *positive* (+1) (Fig. 1, Step 2). All participants were adult volunteers identified exclusively by their self-declared gender. Recruitment was conducted through social media channels and broadcast mailing lists, yielding a sample predominantly composed of individuals residing in Rio de Janeiro with completed higher education.

In spite of the risk of some interpretation overlooking we opted to simplify complex annotation scenarios by ensuring a singular target label for model training, discarding instances without a strict majority vote (289 sentences) (Fig. 1, Step 3), which resulted in a shared subset of 1,174 instances labeled independently by male and female annotators (Fig. 1, Step 4). We refer to this final dataset as MQD-1174-MF, as described in Table I. Feature extraction was performed using TF-IDF representations based on Portuguese word unigrams and bigrams, following the removal of stopwords (Fig. 1, Step 5).

[1]https://farm.pcibex.net/

Table I: Labeling of MQD-1174-MF dataset

Gender	Negative	Neutral	Positive
Male	442	253	479
Female	442	307	425

3.2 Model Training

To investigate gender-driven annotation patterns in supervised learning, we trained independent classification models based on the majority labels provided by each annotator group (i.e., male group and female group). Specifically, we generated models trained exclusively on the labels provided by male annotators and compared the resulting classification outcomes with those of models trained on labels from female annotators. This comparative framework allowed us to examine whether subjective annotation tendencies — such as differing interpretations of sentiment — are preserved, attenuated, or amplified during the learning process.

A range of classifiers was employed—Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR), each trained separately on the male- and female-labeled datasets (Fig. 1, Step 6). To ensure fair and paired comparisons, we applied 5-fold cross-validation using identical data splits across gender-specific models (Fig. 1, Step 7). Prediction outputs were then compared using Cohen's Kappa, chi-square tests, and Cramér's V to assess the level of agreement and the strength of association between models trained on different annotation sources (Fig. 1, Step 8).

3.3 Comparative Evaluation

To evaluate the consistency between gender-based human annotations and machine learning model predictions in textual classification, we employed three complementary metrics: agreement rate, Cohen's Kappa , and Cramér's V. Additionally, the chi-square (χ^2) test was used to assess whether the observed agreement significantly deviates from what would be expected under independence, indicating that the associations are not due to chance. The agreement rate offers an intuitive assessment of the consistency between male and female human annotators, as well as the alignment between human-based annotators and model outputs derived from them. In addition, Cohen's Kappa adjusts this agreement for chance, providing a more robust estimation of reliability by accounting for random concordance. Finally, Cramér's V complements this analysis by quantifying the strength of association between categorical outputs, allowing comparisons across different classification patterns. The use of both Cohen's Kappa and Cramér's V allows us to evaluate not only the level of agreement beyond chance but also the strength of association between classifications, offering a more robust and multi-dimensional perspective on human-human and model-model consistency.

3.4 Interpretation Guidelines

Following [Landis and Koch 1977], $\kappa \geq 0.81$ denotes almost perfect agreement, 0.61–0.80 substantial, and 0.41–0.60 moderate. Here, κ refers to Cohen's Kappa coefficient, a widely used statistical measure for inter-rater reliability for categorical items. To assess the strength of association between gender-based annotations, we computed Cramér's V and interpreted its magnitude according to [Alan and Duncan 1997], who define associations below 0.20 as very low, between 0.20 and 0.39 as low, between 0.40 and 0.69 as modest, between 0.70 and 0.89 as high, and from 0.90 to 1.00 as very high. The chi-square (χ^2) test was used to assess whether there is a statistically significant association between the categorical classifications produced by human annotators and machine learning models. A large χ^2 value indicates that the observed classification patterns deviate significantly from what would be expected under independence — that is, if the outputs were unrelated.

4. RESULTS

Table I elucidates a salient finding: both male and female annotator cohorts exhibited impeccable consistency in the enumeration of negative sentences (442 for each group), whereas discrepancies were predominantly observed in the classifications of neutral and positive sentiments. This phenomenon merits further scrutiny in subsequent research endeavors, potentially through focused psycholinguistic investigations into sentiment perception across diverse linguistic contexts in Portuguese.

Table II: Examples of sentences with different classification by gender.

Sentence	**Male Classification**	**Female Classification**
Estamos vivendo em um mundo onde a aparência física vale mais do que a aparência interior.	3 pos and 1 neg	3 neg and 1 neu
Ódio, eu nao tenho por você, nem pelo meu pai mesmo ele sendo ausente.	3 pos and 1 neg	3 neg and 1 neu
Que vontade de voltar no tempo, ter você do meu lado, com seu olhar iluminado.	3 pos and 1 neg	3 neg and 1 pos

Although both groups labeled the same corpus, female annotators showed a slightly more balanced distribution across sentiment classes, with a higher proportion of neutral labels (26.1% vs. 21.6%) and lower polarization. This resulted in a higher entropy score for the female group (1.5673 vs. 1.5355), indicating greater dispersion in label usage. We applied Shannon entropy using the formula below:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

where p_i denotes the relative frequency of each sentiment class (negative, neutral, positive). Rather than inconsistency, a higher entropy may reflect a more cautious or context-sensitive annotation strategy — particularly in ambiguous or emotionally complex sentences, as demonstrated in Table II, which contains some examples of sentences with a great contrast between male and female majority classification. In sentiment analysis, where true distributions are subjective, entropy differences reveal varying interpretive strategies by gender, which can influence the learning pipeline in model behavior.

Table III: Statistical comparison between annotators and models.

Metrics	**Annotators**	**SVM**	**NB**	**RF**	**DT**	**LR**
Cohen's κ	0.8177	0.6466	0.6741	0.5840	0.3838	**0.6952**
Weighted κ	0.8949	0.6710	0.7062	0.6155	0.4810	**0.7369**
χ^2 (stat)	1548.54	750.13	816.11	784.38	353.23	**957.58**
Cramér's V	0.8121	0.5652	0.5896	0.5780	0.3879	**0.6386**

Standard Scikit-Learn configuration with NLTK stopwords. The χ^2 test yielded p-values effectively equal to zero ($p < 10^{-70}$) when applied to both human annotations and model predictions, suggesting strong statistical dependence between classification outcomes

Table III compiles all statistical measures across annotators and models. The results indicate a highly statistically significant association between the classifications provided by male and female annotators ($\chi^2 = 1548.54$, $p < 0.001$; Cramér's $V = 0.8121$; Cohen's $\kappa = 0.8177$). As described in Section 3.4, these values reflect an almost perfect agreement beyond chance and a very high degree of association, indicating that male and female annotators exhibit highly consistent labeling behaviors despite potential underlying differences in interpretive strategy.

When analyzing machine learning models trained on these gender-specific labels, we observed varying degrees of alignment between the model outputs derived from male and female annotations. Logistic Regression (LR), for instance, yielded a Cramér's V of 0.6386 and a Cohen's Kappa of 0.6952. While these values suggest statistically significant alignment between models trained on male and female annotations, they fall short of the thresholds associated with strong agreement. This indicates a moderate level of consistency across gendered training conditions, lower than that observed among human annotators.

In contrast, the Decision Tree (DT) model showed the weakest alignment between its outputs when trained on male and female annotations ($\chi^2 = 353.23$, $p < 0.001$; Cramér's $V = 0.3879$; Cohen's $\kappa = 0.3838$). According to [Alan and Duncan 1997], a Cramér's V of 0.3879 falls within the "low association" range, while the Cohen's Kappa value indicates only fair agreement between the model's predictions under both training conditions. Compared to the high levels of association and agreement observed among human annotators, the performance of DT highlights a notable drop in consistency, reinforcing the idea that not all machine learning models preserve the same interpretative structure derived from human annotation strategies.

We selected Logistic Regression for visual comparison due to its relatively strong, yet imperfect, alignment across gender-specific models. The graphical comparisons presented in Figures 2 and 3 offer additional insight into the distributional effects of gender-specific training. Figure 2 shows near-perfect overlap between male and female human annotations, whereas Figure 3 illustrates a visible shift in label assignments for models trained under opposing gender annotations — particularly for neutral classes.

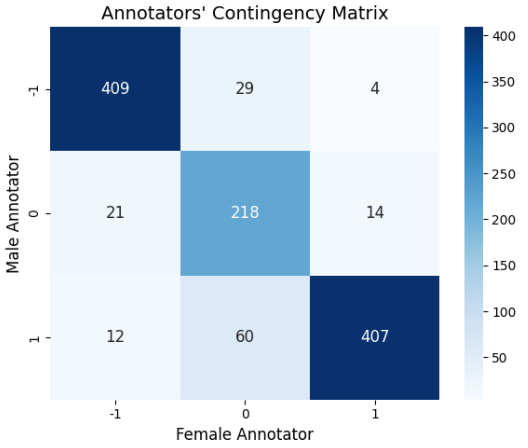

Fig. 2: High agreement in sentiment labels between male and female annotators, showing near-identical distributions.

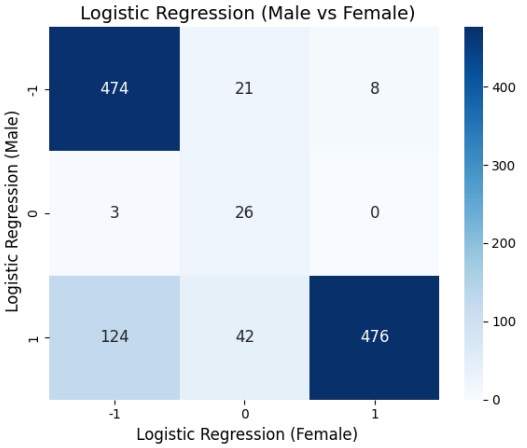

Fig. 3: Label divergence between Logistic Regression models trained on male vs. female annotations, especially in neutral and positive classes.

Altogether, the quantitative and visual results demonstrate that while human annotators were largely aligned in their interpretations, the behavior of classifiers trained on gendered labels diverges considerably — occasionally reinforcing, and in some cases exaggerating, subtle annotation preferences that were less pronounced in the original human data. These findings raise important questions about model robustness and fairness in tasks involving subjective judgment.

5. CONCLUSION AND FUTURE WORK

This study examined how gendered annotation patterns influence sentiment labeling in Brazilian Portuguese texts, and whether such patterns are preserved — or even magnified — when propagated

through supervised classification models. By analyzing 1,465 sentences annotated independently by two gender-balanced groups, we observed high overall agreement in final labels. However, closer inspection revealed notable differences in annotation behavior, including group-specific patterns in consistency, entropy, and class distribution tendencies.

These findings support the view that annotation is not merely a technical step in data preparation, but a social and interpretive process that can shape the behavior of predictive systems. Despite converging on many final labels, the annotation patterns suggest that the two groups (male group and female group) may have employed distinct interpretive tendencies, as evidenced by differences in entropy and label distribution. In practical scenarios, inconsistencies in sentiment evaluation can yield considerable real-world effects.

When classification models were trained separately on the majority labels provided by male and female annotator groups, the resulting predictions not only preserved, but in some cases exaggerated, subtle differences observed in the original human annotations. This observation underscores the need for annotation-aware evaluation strategies, especially in contexts where fairness and representation are central concerns. Rather than treating model disagreement as a post hoc issue, our results suggest that differences in annotation behavior can serve as early indicators of potential disparities in model performance. In that sense, this work offers not only a set of results but also a framework for assessing how annotation behavior influences model behavior—reinforcing the idea that labeling practices are part of the algorithmic pipeline, not just its input.

The range of this research may be constrained by the limited number of annotators involved (eight) and specific nature of the 'Meu Querido Diário' dataset. Considering that the generation of varied datasets is a task that requires significant labor effort [Paullada et al. 2021] and that this dataset was intentionally chosen by its authenticity, we intend to position our study as a foundational case analysis. Beyond increasing the size of the annotator sample to improve the scope of generalization, future research could expand on this foundation in several ways. A qualitative analysis of highly divergent sentence annotations could yield insights into interpretive strategies and ambiguities that statistical measures alone cannot capture. Longitudinal studies might also reveal how annotation behavior changes over time, especially under fatigue or repeated exposure to emotionally charged content. Additionally, extending this analysis to non-binary or more diverse annotator groups could offer a broader view of how identity shapes labeling practices.

In terms of methodology, the importance of advanced validation techniques such as bootstrapping and permutation tests in future work will allow us to assess the stability and significance of observed differences in annotator behavior and model performance under various sampling hypotheses, providing stronger evidence against the possibility of random statistical fluctuation. Integrating these methodologies with statistical indicators such as Krippendorff's Alpha would constitute a remarkable and indispensable advancement for forthcoming research, facilitating an even more comprehensive examination of the reliability of the annotators and the resilience of the findings.

In real-world applications, models trained on datasets containing subtle annotation biases may inadvertently encode and reinforce demographic patterns that are not intrinsic to the task, potentially resulting in unintended disparities in downstream predictions. Recognizing annotation behavior as a critical component of model design can help mitigate potential risks early in the development process. Ultimately, this study reinforces the need to move beyond surface-level agreement and toward a deeper insight into how labeled data is produced, as well as how that production process influences the models that rely on it.

ACKNOWLEDGMENT

This work was supported by the Programa de Pós-graduação em Ciência da Computação (PPCIC) at CEFET/RJ. We thank the anonymous annotators for their contribution of time and effort.

REFERENCES

Alan, B. and Duncan, C. Quantitative data analysis with spss for windows: A guide for social scientists, 1997.

Azevedo, G. d., Pettine, G., Feder, F., Portugal, G., Schocair Mendes, C. O., Castaneda Ribeiro, R., Mauro, R. C., Paschoal Júnior, F., and Guedes, G. Nat: Towards an emotional agent. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, Chaves, Portugal, pp. 1–4, 2021.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, Association for Computational Linguistics, Online, pp. 5454–5476, 2020.

Davani, A., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* vol. 10, pp. 92–110, 2022.

Geva, M., Goldberg, Y., and Berant, J. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pp. 1161–1166, 2019.

Havens, V. and Hedges, M. Uncertainty and inclusivity in gender bias annotation. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, UAE, pp. 25–31, 2022.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. Text classification algorithms: A survey. *Information* 10 (4): 150, 2019.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* 33 (1): 159–174, 1977.

Lim, S. S., Udomcharoenchaikit, C., Limkonchotiwat, P., Chuangsuwanich, E., and Nutanong, S. Identifying and mitigating annotation bias in natural language understanding using causal mediation analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, pp. 11548–11563, 2024.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54 (6): 1–35, 2021.

Minatel, D., da Silva, A. C. M., dos Santos, N. R., Curi, M., Marcacini, R. M., and de Andrade Lopes, A. Data stratification analysis on the propagation of discriminatory effects in binary classification. In *Anais do 11º Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*. Sociedade Brasileira de Computação, Belo Horizonte, MG, pp. 73–80, 2023.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2 (11): 100336, 2021.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. Ai and the everything in the whole wide world benchmark. In *Proceedings of the NeurIPS 2021 Datasets and Benchmarks Track*. NeurIPS, Virtual Conference, pp. 1–10, 2021.

Schwindt, L. C. Predizibilidade da marcação de gênero em substantivos no português brasileiro. *Gênero e língua (gem): formas e usos* vol. 1, pp. 279–294, 2020.

Shah, D. S., Schwartz, H. A., and Hovy, D. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pp. 5248–5264, 2020.

Silva, M. O. and Moro, M. M. Nlp pipeline for gender bias detection in portuguese literature. In *Anais do Seminário Integrado de Software e Hardware (SEMISH)*. SBC, Sociedade Brasileira de Computação, Brasília, Brazil, pp. 1–10, 2024.

Stańczak, K. and Augenstein, I. A survey on gender bias in natural language processing, 2021.

Sun, T., Gaut, A., Tang, S., Huang, Y., Sap, M., Clark, E., Friedman, D., Choi, Y., Smith, N. A., Zettlemoyer, L., et al. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 1630–1640, 2019.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pp. 2979–2989, 2017.