When Annotators Disagree: A Controlled Evaluation of Gender Bias in Sentiment Analysis Using Synthetic Datasets

Érica Carneiro[1], **Alexander Feitosa**[1], **Gustavo Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Brazil

ericacqueiroz@gmail.com, alexander.feitosa@aluno.cefet-rj.br,
gustavo.guedes@cefet-rj.br

Abstract. *This study investigates gender-related annotation bias in sentiment classification. First, we introduce a controlled synthetic dataset generation method that simulates parallel male and female sentiment labels with adjustable inter-annotator agreement. Then, we present the Gender Comparison Methodology, which trains classifiers separately on gender-partitioned labels and evaluates their predictions using shared textual inputs. Agreement is assessed using metrics such as Cohen's Kappa, chi-square test, and Cramér's V. Results show that even moderate disagreement between annotators leads to systematic model divergence, highlighting the importance of annotator identity in shaping classification behavior and informing fairness-aware auditing practices.*

1. Introduction

Language is both a collective and individual construct. While shared linguistic systems enable communication [de Saussure, 1959], individual interpretations are shaped by personal experiences and social identity. In Sentiment Analysis (SA), such variation is often ignored in favor of consensus-based datasets, despite growing evidence that annotator characteristics — including gender — can influence labeling decisions [Kenyon-Dean et al., 2018; Biester et al., 2022; Jiang et al., 2024]. Hence, gender-based differences in perception and communication shape how individuals interpret emotions and express sentiments. In the context of SA, such divergences can influence both the data annotation and the predictions of machine learning systems, which might carry biases from individuals' social perspectives. Meanwhile, the hype surrounding Generative AI often obscures these challenges, portraying Large Language Models (LLMs) as silver bullet solutions. Yet, models trained on crowd-annotated data can reproduce and amplify societal bias [Kiritchenko and Mohammad, 2018; Levshina et al., 2024], especially when disagreement is collapsed into single "ground truth" labels.

Sentiment classification models often rely on datasets annotated under the assumption of objective, homogeneous labeling. However, annotation is not a purely mechanical task — involving deep interpretation shaped by sociocultural factors. Hence, controlling and analyzing such divergences is not only a technical challenge but also an ethical imperative since annotations may carry implicit biases reflecting the annotators' identities, including gender, age, or sociopolitical attitudes [Kenyon-Dean et al., 2018; Biester et al., 2022; Jiang et al., 2024].

Nevertheless, it is also important to note that many studies in the literature do not report the gender composition of annotators, leaving unanswered whether divergen-

ces in labels may have demographic roots. For instance, widely cited datasets often refer to annotators generically as "crowdworkers" without disclosing gender distribution [Kenyon-Dean et al., 2018; Kiritchenko and Mohammad, 2018; Alves et al., 2022]. This omission makes it more difficult to assess how much prejudice might result from gendered interpretation. This leads us to a central hypothesis: what if two sentiment classifiers were trained independently, one using labels solely from male annotators and the other exclusively from female annotators? From this premise, we derive the following research questions: (RI1) Would there be a systematic difference in their predictions?; (RI2) What would this difference tell us about the social aspects of language interpretation?

Some recent studies have begun to explore this question using empirical data. For instance, Feitosa et al. [2025] analyzed gender-based annotation behaviors in a Brazilian Portuguese sentiment dataset and found that male and female annotators differed not only in label distribution but also in internal consistency and entropy. Models trained on these annotations reproduced such divergences to varying degrees, indicating that subjective biases may persist and even intensify during training. While insightful, such studies are constrained by corpus-specific limitations and lack experimental control.

To address this challenge in a controlled and replicable way, we propose a synthetic dataset generation methodology that enables fine-grained control over annotator agreement and label diversity. Building on this, we introduce a gender comparison methodology that simulates male and female annotations independently to assess how classification models respond to divergent labeling sources. This dual contribution aligns with perspectivist principles [Plank, 2022] and advances fairness-oriented evaluation in SA. It should be noted that the synthetic datasets are not intended to replace, replicate, or diminish the nuanced complexity inherent to the human-driven classification process.

This article is organized as follows. Section 2 explains the theoretical background. Section 3 reviews prior research on gender bias in natural language processing. Section 4 details the synthetic dataset generation methodology. Section 5 introduces the gender comparison methodology. Section 6 describes the experimental setup and interpretation guidelines. Section 7 presents the main results and their analysis. Finally, Section 8 summarizes the findings, discusses the limitations, and outlines directions for future work.

2. Theoretical Background

As a subfield of Natural Language Processing (NLP), SA is fundamentally shaped by human perception and is directly influenced by social identity, ideology, and lived experience [Biester et al., 2022]. Rather than being a purely computational task, it inherently reflects the social aspects of human annotators. Although the task of assigning text sentiment polarity (positive, negative, or neutral) may appear objective, annotation decisions are frequently guided by implicit biases and cultural frameworks.

Annotators' ideological attitudes, such as neosexist or authoritarian views, can influence how they label text [Jiang et al., 2024]. These biases are often embedded in benchmark datasets, especially when majority voting masks annotation disagreement [Kenyon-Dean et al., 2018]. As a result, classifiers trained on such data may amplify existing inequalities rather than mitigate them.

Gender, in particular, has been consistently identified as a key source of variation in annotation, influencing not only the perception of emotional tone but also judgments of

offensiveness, sarcasm, and intent [Kiritchenko and Mohammad, 2018; Jiang et al., 2024]. Annotators' ideological attitudes and tendencies towards right-wing authoritarianism or neosexist beliefs affect the way they label content, leading to divergent interpretations of what constitutes sexist or harmful speech [Jiang et al., 2024]. These discrepancies are not just data noise, instead they represent valid, but also conflicting, perspectives frequently lost in the traditional dataset construction processes [Jiang et al., 2024].

This problem is exacerbated in large-scale annotation pipelines where labels are aggregated via majority voting, effectively silencing minority points of view, and giving to the complex and suitable aspects of sentiments simple monolithic "ground truth" values. As shown by Kenyon-Dean et al. [2018], circa 30% of sentiment annotations in crowd-sourced datasets can be classified as "controversial" or subject to annotator disagreement. Yet these instances are often discarded or force-labeled to maintain dataset consistency.

The result is a modeling landscape where even highly accurate systems replicate and amplify the biases embedded in the training data. These models may perform adequately on average but tend to under-perform when applied to content produced in marginalized communities [Kiritchenko and Mohammad, 2018; Levshina et al., 2024].

Given this scenario, rethinking the way data is created, annotated, and interpreted is important for better fairness in machine learning and artificial intelligence model development. The need for perspectivist approaches is underscored by mounting evidence that demographic attributes (such as gender and race) can significantly shape both human annotation and model behavior. Studies have shown that SA systems may assign different emotional intensities to otherwise identical sentences depending solely on the gender or race of the referent [Kiritchenko and Mohammad, 2018; Jiang et al., 2024]. In multilingual contexts, these asymmetries become even more complex, as cross-linguistic variation introduces additional layers of interpretation and bias [Levshina et al., 2024].

Recent research has highlighted an emerging paradigm in the NLP community that advocates for perspectivist data practices [Biester et al., 2022; Plank, 2022]. These approaches challenge the traditional notion of a single correct label, instead promoting the preservation of multiple, coexisting annotations to reflect the inherently subjective nature of interpretation. Such a perspective is particularly relevant to SA, where emotion, tone, and social context intertwine in complex and culturally mediated ways [Biester et al., 2022; Plank, 2022]. Additionally, Assi and Caseli [2024] demonstrate that even advanced generative models like GPT-3.5 Turbo exhibit systematic divergences in output when prompted with gender-marked phrases in both Portuguese and English, reinforcing the persistence of implicit bias across linguistic and cultural contexts.

3. Related Work

Past research has consistently demonstrated the substantial influence of gender on SA. The Equity Evaluation Corpus (EEC) [Kiritchenko and Mohammad, 2018] revealed persistent discrepancies in model predictions based on gendered terms. Jiang et al. [2024] further showed that annotators' ideological positions significantly shape labeling outcomes, particularly in tasks involving gender-based violence. Similarly, user demographics such as gender and age have been found to affect both the expression and interpretation of sentiment in product reviews [Kumar et al., 2020].

Synthetic data generation has gained traction as a tool for fairness-oriented expe-

rimentation. Resources like SentiGEN [Sundarreson and Kumarapathirage, 2024] allow researchers to simulate and control annotation bias across multiple conditions. In parallel, perspectivist NLP frameworks advocate for the retention of multiple interpretations rather than enforcing a single canonical label [Biester et al., 2022; Plank, 2022]. In this spirit, disagreement is revalued as a source of insight rather than noise. Biester et al. [2022] and Kenyon-Dean et al. [2018] caution against discarding "controversial" examples, which may encode valuable variance that contributes to model robustness.

Levshina et al. [2024] challenge the assumption of uniform sentiment polarity across gendered terms, revealing ambivalence rather than pejoration in various linguistic contexts. In parallel, Jiang et al. [2024] and others have shown that ideological and demographic attributes of annotators can significantly influence labeling decisions, reinforcing the need for demographic-aware annotation pipelines.

While these studies provide valuable insights, few have explicitly simulated gendered annotations under controlled conditions for systematic experimental comparison. An empirical contribution in this direction was presented by Feitosa et al. [2025], who analyzed gender-based annotation patterns in a Brazilian Portuguese sentiment analysis corpus of diary-style sentences, labeled as positive, negative, or neutral by male and female annotators. Their findings showed that female annotators tended to produce more variable labels and a higher proportion of neutral classifications, whereas male annotators displayed more consistent and polarized labeling. Models trained on these group-specific annotations reproduced such divergences to varying extents, demonstrating that gendered interpretive patterns can be encoded — and even amplified — during training.

Building on these foundations, the present work introduces a controlled and replicable synthetic framework that allows fine-grained manipulation of annotator agreement and label distribution, thereby enabling a systematic assessment of how gender-specific labeling influences classifier behavior and performance.

4. Synthetic dataset methodology

The synthetic dataset generation procedure is structured around three modular functions, each responsible for a distinct stage of construction. Algorithm 1 outlines this process.

The process illustrates how balanced label pairs are generated with controlled concordance, how synthetic text samples are produced with variable lengths, and how the final dataset is assembled by combining these elements into triplets of text and gendered labels. The algorithm receives three key inputs: the total number of instances N (which must be a multiple of 3 to ensure class balance), a target inter-annotator concordance rate $\gamma \in [0, 1]$, and the parameters $L_{\min}$ and $L_{\max}$, which define the minimum and maximum token length of the generated texts, respectively.

The first function, $\texttt{GenerateBalancedLabelPairs}(\mathcal{C}, n_c, \gamma)$, creates two parallel sequences of sentiment labels — denoted $M = (m_1, \ldots, m_N)$ for male and $F = (f_1, \ldots, f_N)$ for female annotators — based on a fixed class set $\mathcal{C} = \texttt{negative}, \texttt{neutral}, \texttt{positive}$. Each class is represented in equal proportion ($n_c = N/3$ per class), and concordant label pairs are assigned according to the proportion defined by γ. For the remaining discordant instances, label combinations are generated in a way that preserves the overall class balance in both M and F, while ensuring that the two sequences diverge in annotation.

Algorithm 1: Synthetic Sentiment Dataset Generation

Input: N `// Total number of instances (multiple of 3)`
Input: $\gamma \in [0, 1]$ `// Desired concordance rate`
Input: $L_{\min}, L_{\max}$ `// Minimum and maximum text lengths`
Output: Dataset $\mathcal{D} = \{(t_i, m_i, f_i)\}_{i=1}^{N}$

$\mathcal{C} \leftarrow \{\texttt{negative}, \texttt{neutral}, \texttt{positive}\}$; `// Sentiment classes`
$n_c \leftarrow N/3$; `// Instances per class`
$M, F \leftarrow \texttt{GenerateBalancedLabelPairs}(\mathcal{C}, n_c, \gamma)$; `// Balanced and partially concordant`
$T \leftarrow \texttt{GenerateSyntheticTexts}(N, L_{\min}, L_{\max})$; `// Synthetic textual inputs`
$\mathcal{D} \leftarrow \texttt{AssembleDataset}(T, M, F)$; `// Triplets: text, male label, female label`
`Shuffle`$(\mathcal{D})$; `// Randomize order`
return $\mathcal{D}$

The second function, `GenerateSyntheticTexts`$(N, L_{\min}, L_{\max})$, constructs a list $T = (t_1, \ldots, t_N)$ of synthetic text samples. Each t_i is formed by randomly sampling between $L_{\min}$ and $L_{\max}$ placeholder tokens (e.g., `word1`, `word2`, ...). Although devoid of semantic content, these sequences emulate realistic variation in text length and provide a neutral input substrate for isolating the effects of label-based differences on model behavior.

Finally, the `AssembleDataset` function produces the final dataset $\mathcal{D} = (t_i, m_i, f_i)_{i=1}^{N}$ by associating each synthetic text t_i with its corresponding male label m_i and female label f_i. The triplets are then shuffled uniformly to remove any ordering bias, resulting in a ready-to-use dataset suitable for downstream training and evaluation tasks. This procedure offers precise control over annotator agreement, text variability, and label assignment, providing a principled experimental framework for analyzing the interaction between annotation diversity and classifier fairness in SA.

5. Gender comparison methodology

The methodology employed in this study (Fig. 1) aims to systematically assess the divergences in model behavior resulting from gender-specific annotations in SA datasets. The core premise involves evaluating how classifiers trained on identical textual inputs but labeled differently — according to male or female annotators — produce predictions that may reflect or amplify underlying biases.

The process begins with a SA dataset in which each text instance is associated with two independent labels: one assigned by male annotators and the other by female annotators. These labels are categorical (e.g., negative, neutral, or positive).

In Step 1, feature extraction is applied to transform the raw text into numerical representations suitable for machine learning models. Various vectorization techniques may be employed, including term frequency (TF), term frequency-inverse document frequency (TF-IDF), or other bag-of-words-based methods, depending on the experimental setup. Importantly, the feature space remains shared across all subsequent models to ensure comparability.

Step 2 involves the independent training of two distinct classifiers: one using the

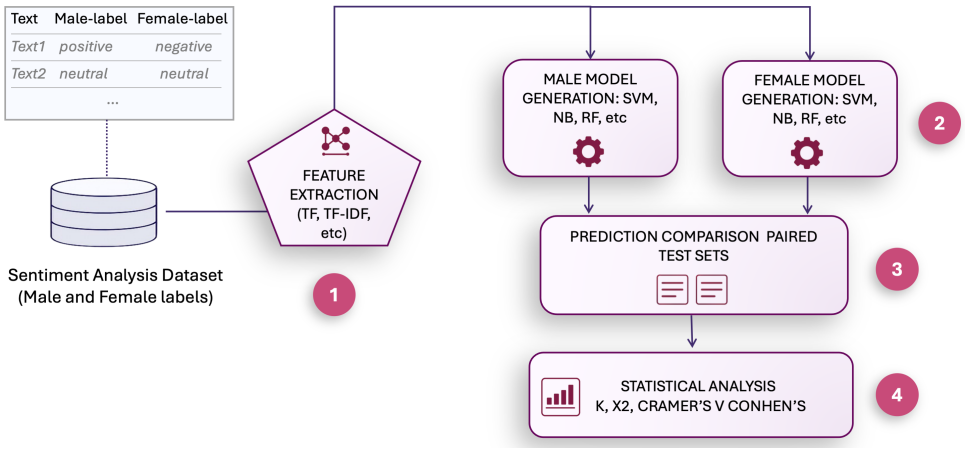

Figura 1. Gender comparison methodology.

male-labeled classes as ground truth and the other using the female-labeled classes. Both models are trained on the same set of input texts and feature representations, differing only in the annotation source used for supervision. This design ensures that any divergence in the models' learned decision boundaries can be attributed solely to the difference in gendered annotations, rather than variations in the input data or architecture.

In Step 3, classifiers are evaluated on paired test sets — identical in content but labeled according to gender. Predictions from the male-based model and the female-based model are compared using the Cohen's Kappa coefficient, which quantifies the agreement between the two classifiers beyond chance. This comparison is conducted across multiple folds in a cross-validation setting to ensure robustness and mitigate sampling variance.

Finally, in Step 4, a set of statistical analyses is conducted to quantify and interpret the divergences between male-based and female-based models. These include the proportion of agreement between annotators, which provides an intuitive measure of consensus, and Cramér's V, which assesses the strength of association between categorical variables — in this case, the predicted class distributions of the male and female models. In addition, Chi-squared tests of independence are performed to test whether the observed differences in model predictions can be statistically attributed to the gender of the label source. Together, these analyses enable a multifaceted assessment of whether and how classification outcomes diverge along gender lines, offering insights into the magnitude and consistency of potential gender bias encoded during supervised learning.

6. Experimental Analysis

This section presents the experimental design, evaluation metrics, and the main findings regarding the impact of gender-specific annotations on sentiment classification. First, we describe the process of synthetic dataset creation, in which we simulate varying degrees of annotation consensus between male and female annotators. These datasets provide a controlled basis for examining classifier behavior under varying inter-annotator agreement. We then describe the computational environment and the statistical tools employed to assess inter-model agreement and annotation divergence. Finally, we report comparative results for four classification models, analyzing disagreement between gender-specific training regimes and interpreting the implications for bias propagation.

6.1. Synthetic Dataset Construction

To systematically explore the relationship between annotation consensus and model agreement, we created 12 synthetic SA datasets following Algorithm 1 in Section 4. Each dataset is labeled using both male and female annotations and is named following the convention `SynSA-XX`, where `XX` denotes the percentage of agreement between male and female annotators (ranging from 50 to 99). For example, `SynSA-80` corresponds to a dataset where 80% of the labels are identical across genders.

Each dataset contains 999 text instances, uniformly constructed to simulate sentiment classification tasks with short documents ranging from 20 to 100 words. Importantly, each instance is associated with two labels — one from a simulated male annotator and another from a simulated female annotator — allowing for controlled training and evaluation of models on gender-partitioned annotation sources. This approach enables investigation of how disagreement between annotators influences model predictions, independent of confounding factors such as content variability or document length.

6.2. Experimental Settings and Interpretation Guidelines

All experiments were conducted using Google Colab, employing Python version 3.11.13 and the Scikit-learn library version 1.6.1. This consistent and reproducible computational environment ensured standardization across all experimental runs, minimizing potential variability due to software dependencies or hardware inconsistencies.

To evaluate the divergence between male and female model predictions and assess the underlying gender-related annotation biases, we employed several statistical measures. The percentage of consensus quantifies the proportion of identical labels assigned by male and female annotators, offering a first indicator of inter-annotator agreement. Cohen's Kappa (k) measures agreement corrected for chance, following the guidelines of Landis and Koch [1977]: $k \geq 0.81$ indicate almost perfect agreement; 0.61–0.80 substantial; 0.41–0.60 moderate; 0.21–0.40 fair; and below 0.20 slight or poor agreement.

In addition, to explore the statistical relationship between the categorical outputs, we performed a chi-square (χ^2) test of independence. Large χ^2 values indicate significant deviation from the null hypothesis of independence — i.e., a stronger association between classifications. The corresponding p-value quantifies the probability of observing such association under the assumption of no relationship, with $p < 0.05$ typically considered statistically significant.

Finally, Cramér's V was computed to assess the strength of association between male and female model predictions. Its values were interpreted based on thresholds proposed by Alan and Duncan (1997): associations below 0.20 were considered very low, between 0.20–0.39 low, 0.40–0.69 moderate, 0.70–0.89 high, and above 0.90 very high.

7. Results and Analysis

The graph presented in Fig. 2 illustrates the Cohen's Kappa values computed between male and female annotators (annotators curve), as well as the inter-model agreement for each of the four classifiers evaluated — Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) — across synthetic datasets with decreasing levels of annotator concordance. These curves allow for a comparative analysis of how different algorithms respond to gender-related annotation discrepancies.

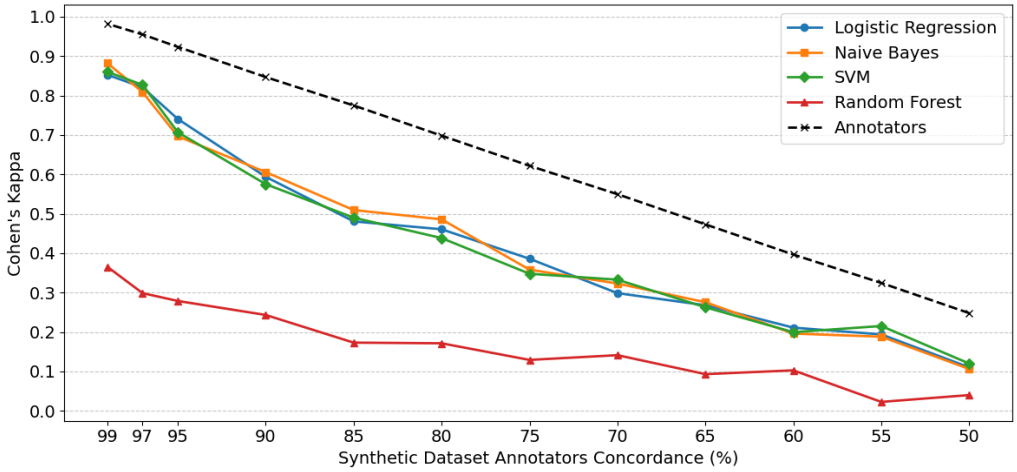

Figura 2. Inter-model agreement (Cohen's Kappa) between four classifiers trained on male and female labels across synthetic datasets and synthetic human annotators agreement.

First, it is evident that the Kappa values between models trained with male and female labels fluctuate in accordance with the original concordance of the human annotators. However, none of the classifiers exhibited a behavior fully aligned with the annotators curve. This discrepancy suggests that classifiers internalize and amplify, in distinct ways, the biases embedded in the original annotations. This phenomenon has been previously identified by Kiritchenko and Mohammad [2018], who demonstrated that SA systems often exhibit gender bias even in simple tasks, attributing different emotional intensities to otherwise identical sentences that differ only in gender references.

Furthermore, the differences observed across models suggest that classifier architecture plays a role in how annotation patterns are internalized and reproduced. LR, NB, and SVM show similar trajectories, with inter-model agreement decreasing steadily as annotator concordance declines. In contrast, RF maintains substantially lower levels of agreement across all datasets. This aligns with findings by Kumar et al. [2020], who argue that demographic attributes such as gender and age shape how sentiments are expressed and interpreted by machine learning models. Thus, it becomes crucial to account for the choice of architecture and the feature extraction methods employed.

A complementary statistical analysis of the results provides further insight into how classifiers (in this case, NB) respond to varying levels of annotation concordance across synthetic datasets, as depicted in Table 1. Beginning with SynSA-99, which exhibits the highest annotator consensus (98.80%), we observe that models trained on male and female labels produce highly concordant outputs, as evidenced by a Cramér's V of 0.982 and a Cohen's Kappa of 0.981. The chi-square statistic, although extremely high (1926.97), confirms that these predictions are significantly dependent, as indicated by the p-value effectively equal to zero. Such strong statistical association suggests a nearly perfect alignment between gender-specific models in scenarios where the data provides little ambiguity — a behavior consistent with high annotator agreement.

As we descend to datasets such as SynSA-90 and SynSA-80, while consensus remains high (89.79% and 79.88%, respectively), Cramér's V declines slightly (0.851 and 0.715), and chi-square values, although still large and significant, decrease proportionally.

Tabela 1. Agreement metrics between male and female labels using NB classifier.

Dataset	Consensus (%)	Cohen's Kappa	Chi²	p-value	Cramér's V
SynSA-99	98.80	0.982	1926.967	0.000	0.982
SynSA-97	97.00	0.955	1823.564	0.000	0.955
SynSA-95	94.89	0.923	1707.568	0.000	0.924
SynSA-90	89.79	0.847	1446.996	0.000	0.851
SynSA-85	84.98	0.775	1227.587	0.000	0.784
SynSA-80	79.88	0.698	1022.289	0.000	0.715
SynSA-75	74.77	0.622	844.656	0.000	0.650
SynSA-70	69.97	0.550	702.965	0.000	0.593
SynSA-65	64.86	0.473	579.080	0.000	0.538
SynSA-60	59.76	0.396	481.834	0.000	0.491
SynSA-55	54.95	0.324	413.877	0.000	0.455
SynSA-50	49.85	0.248	367.889	0.000	0.429

This progressive reduction in statistical association signals the early onset of divergence between male and female model predictions. Despite the underlying textual content being constant, minor differences in the gendered labels begin to impact how models interpret and categorize sentiment. This observation is particularly relevant to fairness-aware NLP, as it reveals how even small deviations in annotation standards can initiate bias amplification during training, as highlighted by Jiang et al. [2024] and Levshina et al. [2024].

Further along the continuum, in mid-range datasets such as SynSA-70 and SynSA-60, we see a marked shift. Here, Cramér's V values drop to 0.593 and 0.491, indicating only modest levels of agreement between gendered models. The chi-square statistics remain statistically significant, but the reduced strength of association suggests that classifiers are learning and reproducing gender-specific annotation patterns in increasingly divergent ways. Importantly, these levels of model disagreement do not necessarily correspond to low-quality annotations; rather, they may reflect the inherent subjectivity present in SA tasks, which becomes more pronounced in settings with moderate annotator consensus.

In the datasets with the lowest agreement among human annotators (SynSA-55 and SynSA-50), the statistical signals confirm gender-driven model divergence. Cramér's V values fall below 0.46, and although chi-square values (413.88 and 367.89) remain highly significant, the patterns captured by each model are clearly driven by gender-specific annotation differences. Notably, in SynSA-50, where consensus is below 50%, the classifier trained on male labels often disagrees sharply with the one trained on female labels, despite using the same feature representation and algorithm. This confirms that supervised models do not merely reflect ambiguity — they transform it into structured, and potentially biased, prediction behavior. Such findings echo concerns raised by Kiritchenko and Mohammad [2018] and Kenyon-Dean et al. [2018] about the limitations of conventional evaluation pipelines in identifying and mitigating representational harms.

As human synthetic annotations become less consistent, classifiers trained on gendered data increasingly diverge, even under identical experimental settings. Thus, consensus, chi-square significance, and Cramér's V collectively offer a triangulated perspective on the fidelity and fairness of gender-based classification behavior. These findings

reinforce the call for model auditing practices that move beyond accuracy metrics, incorporating statistical association measures to assess the robustness and ethical implications of sentiment models in real-world applications.

8. Conclusions and Future Work

This study addressed a critical and often underexplored aspect of supervised machine learning in sentiment analysis: the influence of annotator identity — specifically gender — on the behavior of classification models. While large-scale sentiment datasets are routinely constructed using aggregated human labels, little consideration is given to the diversity or representativeness of those annotators. As highlighted in prior works such as Kiritchenko and Mohammad [2018], annotation bias can significantly shape model behavior, yet standard practices in dataset construction rarely account for this factor explicitly.

Motivated by this gap, we proposed a controlled synthetic dataset generation methodology that simulates sentiment labels from male and female annotators with tunable levels of agreement. We then trained four classification algorithms on gender-partitioned annotations and evaluated their predictions divergence. Our results, measured through inter-model Cohen's Kappa, chi-square tests, p-values, and Cramér's V, reveal that as annotator disagreement increases, so does the divergence in model behavior, suggesting that classifiers do not simply absorb ambiguous labels but structurally encode them into divergent decision boundaries that may amplify latent annotation biases.

These findings resonate with recent concerns in the NLP community regarding the ethical limitations of widely adopted benchmarks. For instance, Kenyon-Dean et al. [2018] argue that the subjective nature of sentiment should not be treated as a flaw to be minimized, but as a meaningful expression of sociolinguistic variation. Our study reinforces this perspective: the divergences between gendered classifiers were not random noise, but consistently structured and correlated with the degree of annotator disagreement.

In line with this perspective, Feitosa et al. [2025] demonstrated that even when final label distributions appear aligned, gender-specific annotation behaviors — such as differences in entropy and label polarization — can lead to divergent model predictions. Their empirical study on Brazilian Portuguese data revealed that classifiers trained on male and female labels behaved differently, particularly in ambiguous or emotionally charged sentences. Our findings reinforce and expand this evidence by showing that, under controlled synthetic conditions, such divergences are not only preserved but can be systematically modulated and statistically quantified.

Moreover, our statistical analysis showed that these divergences were not only present but significant. The chi-square values increased proportionally to annotator consensus, with p-values nearing zero across all datasets, confirming that the probability of these results being due to chance is negligible. These results are in line with observations by Jiang et al. [2024], who demonstrated that even when sentiment labels appear superficially aligned, deeper biases often persist in how models interpret gendered content.

Finally, we advocate for incorporating disagreement-aware auditing protocols into machine learning development cycles. As suggested by Luitel et al. [2025], model disagreement — far from being a defect — can serve as a diagnostic tool to surface and interrogate ethical blind spots. Rather than striving for consensus as the sole marker of data quality, future research should build real-world datasets with balanced demographics.

Referências

Alves, A. A. C., de Souza, L. F. M., Varjolo, L. D., Mauro, R. C., Belloze, K., Paschoal, F., and Guedes, G. (2022). Vita 2.0 – class evaluation system. In *Proceedings of the 17th Iberian Conference on Information Systems and Technologies (CISTI)*, Online. Iberian Conference on Information Systems and Technologies.

Assi, F. and Caseli, H. (2024). Biases in gpt-3.5 turbo model: a case study regarding gender and language. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 294–305, Porto Alegre, RS, Brasil. SBC.

Biester, L. et al. (2022). Analyzing the effects of annotator gender across nlp tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP*.

de Saussure, F. (1959). *Course in General Linguistics*. McGraw-Hill.

Feitosa, A., Carneiro, E., and Guedes, G. (2025). Beyond systematic bias: Investigating gender differences in portuguese text classification annotation patterns. In *Anais do XIII Symposium on Knowledge Discovery, Mining and Learning*. SBC.

Jiang, A. et al. (2024). Re-examining sexism and misogyny classification with annotator attitudes. *arXiv preprint arXiv:2410.03543*.

Kenyon-Dean, K., Ahmed, E., Fujimoto, S., Georges-Filteau, J., Glasz, C., Kaur, B., Lalande, A., Bhanderi, S., Belfer, R., Kanagasabai, N., et al. (2018). Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895.

Kiritchenko, S. and Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Kumar, S. et al. (2020). Exploring impact of age and gender on sentiment analysis using machine learning. *Electronics*, 9(2):374.

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

Levshina, N., Koptjevskaja-Tamm, M., and Östling, R. (2024). Revered and reviled: a sentiment analysis of female and male referents in three languages. *Frontiers in Communication*, 9.

Luitel, S., Liu, Y., and Anwar, M. (2025). Investigating fairness in machine learning-based audio sentiment analysis. *AI and Ethics*, pages 1099–1108.

Plank, B. (2022). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Sundarreson, P. and Kumarapathirage, S. (2024). Sentigen: Synthetic data generator for sentiment analysis. *Journal of Computing Theories and Applications*, 1(4).