RacismoBR: A Manually Annotated Dataset for Racist Discourse Detection in Brazilian Portuguese

João Vítor Vaz and Fabrício Benevenuto and Marcos André Gonçalves
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG, Brasil
{joaovitorvaz, fabricio, mgoncalv}@dcc.ufmg.br

Abstract

Racist discourse on social media appears both through explicit attacks and subtle, context-dependent forms, remaining a challenge for Natural Language Processing. We introduce **RacismoBR**, a culturally grounded dataset for detecting racist discourse in Brazilian Portuguese, manually annotated exclusively by Black researchers to ensure sociolinguistic validity and epistemic representativeness. We conduct a controlled evaluation of binary racism classification in our dataset considering several classification modeling paradigms: classical machine learning, supervised Transformer-based (Small) Language Models, and Large Language models under in-context, few-shot learning. Results show that GPT-4.1 and BERTimbau yield the highest Macro-F1 scores; however, Wilcoxon signed-rank tests reveal no statistically significant differences across models, mostly due to high variability. Across paradigms, classifiers consistently display higher precision for non-racist content and higher recall for racist content. A qualitative analysis highlights persistent difficulties with implicit, euphemized, and context-dependent racism. These findings indicate that culturally grounded annotation plays a more decisive role than architectural sophistication alone in advancing racism detection. **Dataset URL:** https://github.com/joaovitorvaz/RacismoBR-Dataset. **Disclaimer.** This paper contains examples of hateful and offensive language, which are included solely for research and analysis purposes.

1 Introduction

Social networks have become central spaces for information exchange, public debate, and cultural expression. Among these platforms, X (formerly Twitter) plays a prominent role in Brazil, with more than 22 million active users annually (RD Station, 2025). However, the features that enable rapid communication also expand the circulation of discriminatory content, including racism, a structural and enduring phenomenon reproduced in digital environments. These attention-oriented dynamics favor the propagation of racialized insults, insinuations, and other harmful expressions.

Racism, understood as a system of oppression rooted in the hierarchization of phenotypic traits and ethno-racial origins (Almeida, 2019), manifests both explicitly (e.g., direct attacks) and implicitly (e.g., stereotypes, irony, denial). Such discursive complexity poses significant challenges for Natural Language Processing (NLP), as accurate detection requires sensitivity to sociocultural cues that extend beyond surface linguistic patterns.

Despite advances in hate-speech detection (Shen et al., 2025), efforts dedicated specifically to racism in Portuguese remains limited. Existing datasets lack representativeness and often rely on annotation protocols developed in different linguistic and cultural contexts, affecting model sensitivity to nuanced forms of racism. This gap motivates our study, which addresses three research questions: (**RQ1**) how culturally grounded annotation, carried out exclusively by Black researchers, affects model behavior; (**RQ2**) under a strictly controlled and exploratory evaluation setup, how classical machine learning models, fine-tuned Transformer-based Small Language Models (SLMs), and in-context learning (few-shot) Large Language Models (LLMs) behave relative to one another, without the goal of establishing performance rankings or architectural superiority; and (**RQ3**) which systematic error patterns arise, particularly for implicit or context-dependent racism.

Our results show that: (**RQ1**) Rather than providing causal evidence, our results should be understood as observations emerging from a theoretically grounded annotation choice that foregrounds nuanced and context-dependent forms of racist discourse. (**RQ2**) GPT-4.1 and BERTimbau

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 770–779

achieve the highest Macro-F1 scores, but Wilcoxon tests reveal no statistically significant differences across paradigms, mostly due to high variance, also indicating that architectural sophistication alone does not consistently translate into superior effectiveness under the evaluated conditions; and (**RQ3**) although models tend to display higher precision for non-racist content and higher recall for racist content, systematic errors persist in cases of implicit, euphemized, and context-dependent racism, where sociocultural interpretation is crucial.

In sum, this work contributes by: (1) constructing a new epistemically grounded **dataset** of racist discourse in Portuguese, annotated exclusively by Black researchers; (2) performing a **comparative evaluation** of several classical and modern classification models on culturally representative data; and (3) analyzing **recurrent error patterns**, highlighting the challenges of implicit racism for current NLP methods.

2 Related Work

Detecting hate speech and, more specifically, racism remains a central challenge in NLP. Although Transformer-based models have advanced effectiveness, work in Portuguese still faces limitations related to scarce annotated resources, cultural and linguistic specificities, and persistent bias (Fortuna et al., 2019; Leite et al., 2020; Souza et al., 2020). We focus on Brazilian Portuguese, where racialized discourse exhibits distinctive sociolinguistic characteristics.

2.1 Racism and Hate Speech in Portuguese

Early studies on abusive language in Portuguese treated hate speech as a broad category without explicitly distinguishing specific forms of discrimination. Fortuna et al. (2019) introduced a hierarchically annotated dataset for hate speech in Portuguese, enabling analysis of diverse abusive targets under a unified taxonomy, while Leite et al. (2020) presented a multilingual toxic language dataset expanding Portuguese coverage in abusive language detection tasks. These resources established important baselines but primarily adopted general hate speech formulations.

More recent work has shifted toward narrower phenomena, such as sexist hate speech (Salenave Santana et al., 2022), demonstrating how fine-grained annotation schemes are necessary to capture discursive patterns obscured by broader categorizations. Studies on low-resource hate speech detection (Assis et al., 2024) highlight data availability and evaluation design impacts when modeling socially situated phenomena in Portuguese.

Complementary approaches address offensive language, toxicity, and hate speech in digital texts through moderation frameworks (Silva Neto, 2017). Hierarchical resources incorporate racism alongside sexism and xenophobia (Trajano et al., 2022), providing methodological foundations despite not isolating racism analytically.

In Brazilian Portuguese, racism is typically subsumed under broader hate speech or offensive language formulations, while research explicitly targeting racism against Black individuals as a socially and historically situated phenomenon remains comparatively scarce. Social media analyses on Twitter (Silva et al., 2016; Mondal et al., 2017; Reis, 2021) achieved competitive results with linear models, yet limited corpus sizes and sociolinguistic diversity constrain generalization, particularly for implicit or euphemized racism.

Taken together, these works motivate datasets and evaluation protocols explicitly designed to capture racism as a socially and historically situated phenomenon in Brazilian Portuguese.

2.2 Learning Models and Linguistic Challenges

Transformer-based models, especially BERTimbau (Souza et al., 2020), have become strong baselines for Portuguese text classification. Nevertheless, recent evidence suggests that gains in hate speech detection depend not only on architecture but also on dataset representativeness and sociocultural grounding. Subtle and context-dependent racist expressions continue to challenge lexicon-based and shallow models. While hybrid approaches show promise (Putra and Wang, 2024), they remain largely developed for English and rarely adapt to Brazilian sociolinguistic realities, reinforcing the need for culturally grounded evaluation.

2.3 Representativeness, Ethics, and Algorithmic Fairness

Work on bias and ethics in NLP highlights the role of epistemic diversity in annotation and dataset design (Blodgett et al., 2020; Bender et al., 2021). The absence of perspectives from marginalized groups risks reinforcing structural inequalities. Evidence from other domains (Cascalheira and et al., 2024) demonstrates that

culturally informed annotation improves sensitivity to socially embedded phenomena.

To our knowledge, no publicly available dataset for racist discourse in Brazilian Portuguese explicitly centers Black epistemic perspectives or documents annotator positionality. Our contribution is therefore both methodological and ethical: we introduce a dataset annotated exclusively by Black researchers and examine how epistemically grounded annotation affects model behavior and evaluation outcomes.

3 Theoretical Foundations: Racism, Language, and Digital Spaces

Racism operates as a structural system of domination that organizes social, institutional, and discursive relations (Almeida, 2019). In digital environments, these dynamics are intensified by platform algorithms, which mediate the circulation and amplification of racialized content; as (Silva, 2022) argues, such systems not only reproduce but also reshape racial hierarchies through automated decision processes.

From a critical-theoretical perspective, racism encompasses more than overt insults. Fanon's notion of the epidermalization of inferiority (Fanon, 2008) and Hall's discussion of representational regimes (Hall, 1997) highlight how racial hierarchies are internalized, naturalized, and sustained through everyday discourse. These symbolic mechanisms pose specific challenges for NLP, as racist meaning frequently resides in contextual, pragmatic, or metonymic cues difficult for statistical models to capture.

This view aligns with Critical Discourse Analysis, where racialization emerges through lexical and syntactic patterns that reproduce power asymmetries (van Dijk, 2008). In Brazilian Portuguese, these processes intersect with coloniality and the ideology of racial mixture, giving rise to context-dependent forms of implicit or euphemized racism.

Recognizing language as an ideological construct is therefore essential for NLP research. Systems that ignore this dimension tend to reproduce harmful biases (Blodgett et al., 2020; Bender et al., 2021). Building on these insights, this study adopts a socio-technolinguistic perspective, integrating critical theory and computational methods to analyze and detect both explicit and implicit forms of racist discourse. These theoretical insights directly inform the annotation protocol and the interpretation of classification errors presented next.

4 Methodology

The experimental setup described in this section is used consistently across all evaluations reported in Section 5. The presence of offensive language in the dataset is inherent to the task and limited to analytical and methodological purposes.

4.1 Data Collection

Constructing a labeled dataset for racist discourse detection poses significant methodological challenges, particularly due to the non-homogeneous distribution of such messages in digital environments. For this reason, probabilistic sampling techniques such as simple random sampling or stratified sampling (Cochran, 1977) were not appropriate, as they would likely produce datasets with low density of relevant examples. Thus, we adopted an intentional keyword-based sampling strategy, which is better suited for sparse phenomena.

Data collection was conducted between September 2024 and January 2025 using a paid access service to the X API, limited to public content. We obtained the text of posts and associated replies. To ensure the protection of sensitive data, all direct identifiers were removed, preserving only textual content and essential linguistic metadata.

Initial selection was based on the following terms: "mulambo", "racismo", "racista", "angolano/a", "candomblé", "africano/a" and "neguinho/a". These terms were chosen based on studies investigating manifestations of racism in Brazil (Miranda, 2020; Caetano, 2020), capturing recurring forms of racialization observed in the Brazilian context. In total, the collection returned 405 posts, including original publications and replies.

4.2 Expert Annotation

To ensure consistency and cultural sensitivity in labeling, annotation was carried out by eight black researchers organized into four independent pairs. The annotators come from diverse academic backgrounds, regions of Brazil, and phenotypic profiles, ensuring epistemic plurality and reducing interpretive biases.

Annotators were instructed using the following operational definition of racism, adapted from (Almeida, 2019): *Racism is any form of expression, direct or indirect, that conveys the hierarchization*

of phenotypes, the internalization of a negative self-image, or the subordination of black people. This includes discourse that demeans, offends, discriminates, or stigmatizes black individuals on the basis of race, ancestry, or phenotypic traits, whether explicitly or implicitly. Posts were classified into two categories: Class 1: racist; Class 0: non-racist.

Each pair independently labeled their subset. The consensus methodology was inspired by (Cascalheira and et al., 2024), originally applied to datasets related to the experiences of LGBTQIAPN+ minorities. Among the 405 annotated posts, 386 showed complete agreement. The minimum Cohen's Kappa between the annotator pairs was 0.655, indicating good agreement. The 39 posts with disagreement were reviewed by a third independent evaluator, resulting in the reclassification of 16 cases. The final process after re-annotation yielded 113 racist posts and 312 non-racist posts.

Beyond its technical contribution, this annotation design represents a methodological and epistemic advance for studies of racist discourse in Portuguese. By placing Black researchers as the agents who define and apply the labeling criteria, the dataset incorporates sociolinguistic and cultural perspectives largely absent from prior corpora, thereby improving its analytical validity.

4.3 Incorporation of Literature Data

To expand the corpus and increase linguistic diversity, we incorporated 629 posts collected from the X platform and previously labeled as racist in related studies (Fortuna et al., 2019; Leite et al., 2020; Augusto, 2021; Silva Neto, 2017). Because these works adopted broad definitions of hate speech encompassing multiple minority groups, all imported posts were fully re-annotated from scratch using the same strict operational definition of racism against Black individuals and the same annotation protocol applied to the newly collected data. After re-annotation, 316 posts were labeled as racist and 313 as non-racist.

Combining data originally collected between 2017 and 2021 with posts from 2024 and 2025 may introduce temporal variations in language usage. However, the unified re-annotation process under a single operational definition ensures the semantic consistency of the target phenomenon (racist discourse), thereby mitigating potential labeling drifts over time.

These 313 non-racist posts were combined with 312 non-racist posts collected via the X API in this study, resulting in a pool of 625 non-racist instances. For experimental balance, a subset of 429 non-racist posts was selected to match the 429 racist posts used in the experiments. All instances in the final dataset therefore originate exclusively from the X platform.

4.4 Preprocessing

Preprocessing involved the removal of URLs, mentions, emojis, and special characters, as well as lowercasing. For classical models (Naive Bayes, Logistic Regression, Random Forest, and XGBoost), the texts were tokenized and lemmatized. For Transformer-based models (BERTimbau and RoBERTa), we employed the native tokenizer, preserving contextual segmentation. These steps contributed to textual standardization and reduced linguistic noise.

For LLMs, preprocessing was minimal, since these models operate directly on raw text. Only URL and mention removal was applied to avoid unnecessary contextual distraction. The input fed to the LLMs included an instruction format specifying the task, the label space, and the expected output. No additional cleaning or normalization was performed, preserving the discursive and stylistic cues that are often crucial for detecting implicit forms of racism.

4.5 Experimental Protocol and Reproducibility

All models were evaluated under an identical and strictly controlled experimental protocol to ensure direct comparability across paradigms. A fixed random seed (42) was used throughout all experiments. Model evaluation employed Stratified 5-Fold Cross-Validation, preserving class proportions in each fold.

To guarantee full reproducibility, the indices of each fold were generated once and persisted to disk, being reused across all model configurations, including classical machine learning models, Transformer-based classifiers, and LLMs. Class imbalance was addressed through random undersampling of the majority class, resulting in a balanced dataset of 858 posts (429 racist and 429 non-racist) prior to cross-validation.

This protocol ensures that observed effectiveness differences stem from modeling choices rather than data partitioning artifacts or stochastic variation.

4.6 Prompting Strategy for LLMs

LLMs were evaluated under an in-context learning (ICL) setup, without task-specific fine-tuning or gradient updates. We employed a single, fixed prompt template and varied only the number of in-context examples per class, with $k \in \{0, 2, 4\}$, corresponding to 0-shot, 4-shot, and 8-shot prompting in total.

Preliminary comparisons among the 0-, 2-, and 4-example-per-class settings indicated modest and unstable differences in Macro-F1; we therefore adopt the 8-shot configuration (4 examples per class) as a fixed reference condition for comparative evaluation.

In the adopted 8-shot setting, prompts contained eight labeled instances in total (four racist and four non-racist). These examples were class-balanced, fixed per fold, and reused for all test samples within the same fold.

Prompts were deterministic, with outputs restricted to binary labels (0 = non-racist, 1 = racist). Each post was evaluated independently using the same prompt template, with in-context examples drawn from the corresponding training fold and no conversational history.

In the few-shot setting, labeled examples preceded the target input, serving as task demonstrations. The prompt followed a fixed instructional template, structured as:

> **You are a binary classifier of racism.**
>
> **Definition:** Racism is any form of expression, direct or indirect, that conveys the hierarchization of phenotypes, the internalization of a negative self-image, or the subordination of Black people. This includes discourse that demeans, offends, discriminates, or stigmatizes Black individuals based on race, ancestry, or phenotypic traits, whether explicitly or implicitly.
>
> **Instruction:** Respond only with 0 or 1.
>
> **Text:** [Input text]

All experiments were conducted using a Portuguese version of the prompt, semantically equivalent to the template presented above, and used consistently across all folds and models.

4.7 Classification Models

We evaluated a diverse set of classification models, covering different learning paradigms and representational capacities.

Naive Bayes (NB). A probabilistic classifier assuming conditional independence among features, implemented as Multinomial Naive Bayes with Laplace smoothing. Texts were vectorized using TF–IDF with unigrams and bigrams.

Logistic Regression (LR). A linear discriminative model that captures lexical correlations without assuming feature independence. We employed L2 regularization and the `liblinear` solver.

Random Forest (RF). An ensemble model combining multiple decision trees through majority voting, offering robustness to noise and reduced overfitting. TF–IDF representations with unigrams and bigrams were used.

XGBoost. A gradient boosting model based on sequentially optimized decision trees, configured with 200 estimators and moderate depth to balance bias and variance. Texts were represented using TF–IDF features with unigram and bigram n-grams, following the same preprocessing pipeline adopted for the other classical models.

BERTimbau. A Transformer-based model pretrained for Brazilian Portuguese. We fine-tuned the base version on the training partition of each cross-validation fold using native tokenization, truncation to 128 tokens, three training epochs, mixed precision (FP16), and gradient accumulation.

XLM-RoBERTa. A multilingual Transformer pretrained on large-scale cross-lingual corpora. The model was fine-tuned on the training partition of each cross-validation fold under the same configuration adopted for BERTimbau, enabling comparison between monolingual and multilingual pretrained encoders.

GPT-4.1. A large-scale instruction-following language model evaluated in an in-context learning (few-shot) configuration, without task-specific fine-tuning. Each post was processed independently using a structured prompt that explicitly incorporated the same operational definition of racism adopted during human annotation. The model output was restricted to binary labels.

LLaMA 3.1 8B. An open-weight instruction-tuned language model evaluated under the same in-context learning (few-shot) protocol as GPT-4.1. This configuration enables comparison between proprietary and open LLMs under identical conditions.

5 Experimental Results and Analyses

Table 1 shows that, although point estimates of Macro-F1 vary across the evaluated models, these differences remain within largely overlapping confidence intervals. In particular, GPT-4.1 and

BERTimbau achieve the highest mean Macro-F1 scores, but without consistent separation from the remaining models. This numerical proximity already suggests convergent behavior across paradigms when they are evaluated under the same dataset and experimental protocol.

This convergence becomes even more evident in the confusion matrices (Figures 1–4), which reveal highly similar error patterns across classical models, fine-tuned Transformer-based SLMs, and LLMs evaluated under in-context learning. Misclassifications tend to concentrate on the same subsets of instances, especially those involving implicit, euphemized, or context-dependent forms of racism, indicating that models repeatedly struggle with the same discursive phenomena.

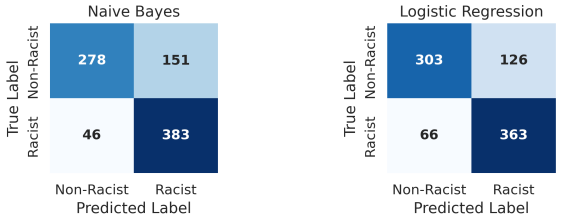

Figure 1: Confusion matrices for Naive Bayes (left) and Logistic Regression (right).

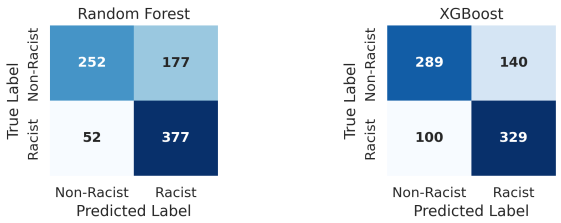

Figure 2: Confusion matrices for Random Forest (left) and XGBoost (right).

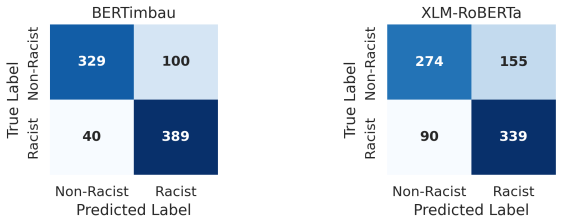

Figure 3: Confusion matrices for BERTimbau (left) and XLM-RoBERTa (right).

This qualitative observation is quantitatively supported by the paired Wilcoxon signed-rank tests on Macro-F1 scores (Table 2), for which no comparison reaches statistical significance ($\alpha = 0.05$). In this setting, the observed point-estimate differences

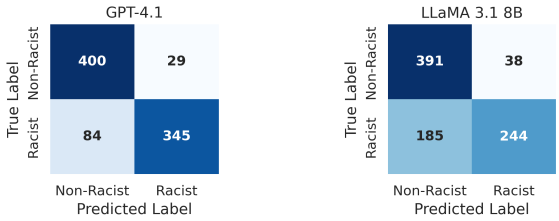

Figure 4: Confusion matrices for GPT-4.1 (left) and LLaMA 3.1 8B (right).

should be interpreted as reflecting fold-level variability under a strictly controlled evaluation design, rather than as evidence of structural superiority of any specific architecture. Importantly, the lack of statistical significance should not be construed as evidence of strict equivalence between models, but rather as delimiting the scope of the inferences that can be drawn from the present dataset and protocol.

A consistent pattern across all evaluated paradigms is the asymmetry between classes: precision is systematically higher for non-racist posts, whereas recall is higher for racist posts. This behavior, observable both in the aggregate metrics reported in Table 1 and in the confusion matrices, reflects discursive properties of racist language in Brazilian Portuguese. Explicitly racialized expressions tend to facilitate the identification of racist content, while neutral, ironic, or metalinguistic uses of racialized terms frequently lead to false positives.

Crucially, these asymmetries and error patterns recur independently of the learning paradigm. Classical machine learning models, fine-tuned Transformer-based SLMs, and LLMs evaluated in a few-shot in-context setting exhibit similar limitations in identifying implicit and context-dependent racism. This convergence suggests that the primary challenges of the task lie less in architectural capacity and more in the sociopragmatic nature of the phenomenon and the characteristics of the data. Taken together, the results reinforce the interpretation that future advances in racism detection for Portuguese are more likely to stem from culturally grounded data collection, epistemically informed annotation schemes, and context-aware evaluation protocols than from the isolated adoption of increasingly complex architectures.

6 Qualitative Error Analysis

To complement the quantitative evaluation, we conducted a qualitative error analysis focused on false

Table 1: Model effectiveness with 95% confidence intervals.

Model	Acc.	Macro-F1	P0	R0	P1	R1
GPT-4.1	0.868±0.025	0.868±0.025	0.827±0.025	0.932±0.031	0.923±0.033	0.804±0.032
BERTimbau	0.826±0.040	0.824±0.040	0.889±0.050	0.770±0.100	0.780±0.060	0.907±0.050
Logistic Regression	0.776±0.014	0.775±0.014	0.821±0.011	0.706±0.028	0.743±0.018	0.846±0.013
Naive Bayes	0.770±0.018	0.767±0.019	0.859±0.018	0.648±0.021	0.718±0.020	0.893±0.016
Random Forest	0.733±0.030	0.727±0.032	0.828±0.022	0.587±0.059	0.682±0.032	0.879±0.016
XGBoost	0.720±0.022	0.719±0.022	0.744±0.026	0.674±0.038	0.702±0.025	0.767±0.032
XLM-RoBERTa	0.715±0.050	0.714±0.050	0.753±0.060	0.640±0.080	0.687±0.060	0.791±0.070
LLaMA 3.1 8B	0.740±0.028	0.732±0.030	0.680±0.024	0.911±0.048	0.864±0.028	0.569±0.014

Table 2: Pairwise Wilcoxon signed-rank tests on Macro-F1 scores.

Model 1	Model 2	p-value
GPT-4.1	BERTimbau	0.0625
GPT-4.1	Logistic Regression	0.0625
GPT-4.1	Naive Bayes	0.1250
GPT-4.1	Random Forest	0.1250
GPT-4.1	XGBoost	0.1875
GPT-4.1	XLM-RoBERTa	0.0625
GPT-4.1	LLaMA 3.1 8B	0.0625
BERTimbau	Logistic Regression	0.0625
BERTimbau	Naive Bayes	0.0625
BERTimbau	Random Forest	0.1250
BERTimbau	XGBoost	0.1875
BERTimbau	XLM-RoBERTa	0.0625
BERTimbau	LLaMA 3.1 8B	0.0625
Logistic Regression	Naive Bayes	0.3125
Logistic Regression	Random Forest	0.1875
Logistic Regression	XGBoost	0.3125
Logistic Regression	XLM-RoBERTa	0.1250
Logistic Regression	LLaMA 3.1 8B	0.1250
Naive Bayes	Random Forest	0.3125
Naive Bayes	XGBoost	0.3125
Naive Bayes	XLM-RoBERTa	0.0625
Naive Bayes	LLaMA 3.1 8B	0.0625
Random Forest	XGBoost	0.3125
Random Forest	XLM-RoBERTa	0.1250
Random Forest	LLaMA 3.1 8B	0.1250
XGBoost	XLM-RoBERTa	0.1875
XGBoost	LLaMA 3.1 8B	0.1875
XLM-RoBERTa	LLaMA 3.1 8B	0.0625

negatives, the most consequential errors in racism detection. These patterns appeared consistently across all models, aligning with their confusion matrices and recall scores for the racist class.

False negatives often involved implicit, euphemized, or context-dependent racism. These posts lacked explicit lexical markers and instead relied on presupposition, evaluative stance, irony, or indirect expressions of racial hierarchy. The examples below illustrate recurrent patterns observed across multiple models and folds, rather than isolated failures. Table 3 presents representative cases, with the original Portuguese in *italics* and English translations alongside. The recurrent sources of error can be grouped into the following categories:

Presuppositional framing and indirect comparison. Racist meaning is conveyed through presupposed statements or analogical constructions that require pragmatic inference rather than explicit lexical cues. Such cases consistently evade both linear classifiers and contextualized models.

Evaluative judgments and aesthetic hierarchies. Statements expressing racial hierarchy through aesthetic or normative evaluation encode discriminatory meaning implicitly, which remains difficult to capture.

Discourses of denial and epistemic invalidation. Utterances that deny the existence of racism reproduce structural inequality while lacking overtly negative lexical markers, leading to systematic misclassification across models.

Mitigated prejudice and face-saving strategies. Disclaimers such as "nada contra" reduce surface-level negativity while preserving discriminatory intent, weakening signals typically exploited by classifiers.

Lexical ambiguity and racialized connotations. Seemingly neutral terms may carry racialized meanings in Brazilian Portuguese that depend on shared cultural knowledge rather than explicit semantics.

Sarcasm and irony. Ironical or sarcastic constructions require detecting a mismatch between literal content and intended meaning, a capability that remains limited even in the best models.

These error patterns highlight structural limitations shared across modeling paradigms, indicating that increased model capacity alone is insufficient to address implicit and context-dependent racism without richer sociocultural grounding.

7 Conclusion

This work introduced **RacismoBR**, a manually labeled dataset for detecting racist discourse against Black people in Brazilian Portuguese, grounded in epistemically informed annotation. Under a unified protocol, we showed that effectiveness differences across classical models, Transformer encoders, and

Table 3: Examples of false negatives: original text and English translation.

Portuguese (original)	English translation
Sabe a diferença entre o negro e o câncer? O câncer evolui.	Do you know the difference between a Black person and cancer? Cancer evolves.
Batom vermelho só é bonito em gente branca.	Red lipstick is only attractive on white people.
Racismo não existe, tira essa parada da cabeça!!	Racism does not exist; stop thinking about that.
Tinha que ser preto (nada contra))))	Had to be Black (nothing against them))))).

LLMs are not statistically significant when culturally representative data are used. Consistent asymmetries emerged across all evaluated models, with higher precision for non-racist content and higher recall for racist content. However, implicit, euphemized, and context-dependent forms of racism remain challenging, with similar error patterns observed across models of different complexities.

Overall, progress in Portuguese racism detection depends more on dataset design, annotation grounded in sociocultural expertise, and context-aware evaluation than on increasingly sophisticated architectures. Future work will expand the dataset, incorporate conversational context, and explore interpretability and multilingual directions to advance equitable NLP.

Limitations

This study operates within the inherent challenges of modeling socially complex phenomena such as racism. The dataset was constructed using a keyword-based sampling strategy, which may shape lexical distribution, but is appropriate for sparse, low-prevalence phenomena and ensured sufficient coverage for robust analysis.

Although conversational context is valuable, our focus on single-turn analysis aligns with the operational reality of most real-time content moderation systems, which often process disjointed streams of comments due to latency constraints, API limitations, or platform-level access restrictions.

The dataset focuses specifically on racism against Black individuals, prioritizing conceptual clarity and annotation consistency. Although this narrows immediate generalization, it enables precise analysis of a well-defined phenomenon without conflating distinct forms of prejudice.

Annotation by Black researchers strengthened epistemic grounding while acknowledging that some interpretive variability is inherent to context-dependent language. Rather than a weakness, this reflects task complexity and reinforces the importance of informed annotation practices. Few-shot examples were fixed within each fold to reduce variance and ensure comparability across models. Exploring prompt sensitivity to example selection is left for future work.

Despite these constraints, the experimental protocol is internally consistent, and the conclusions drawn are valid within the explicitly defined scope of binary racism detection in Brazilian Portuguese under controlled evaluation conditions.

Conclusions regarding LLMs should be interpreted as conditional on the evaluation setup adopted in this study. We evaluated two instruction-tuned LLMs under a fixed in-context learning configuration, using a single prompt template and fixed few-shot examples per fold to ensure comparability and reproducibility. This design does not explore prompt sensitivity or example selection effects, which are known to influence LLM behavior. Consequently, part of the observed variance may reflect properties of the selected in-context examples rather than intrinsic model capabilities.

Ethical Considerations

This study addresses the detection of racist discourse, a task with significant ethical implications due to potential harms from misclassification. Data collection complied with platform policies, used only publicly available text, and removed identifiable information to protect privacy.

Annotation was deliberately conducted by Black researchers as an epistemically grounded choice, recognizing that racism is socially situated and often implicit. Diversity of regional and sociolinguistic backgrounds helped reduce individual bias while strengthening interpretive reliability.

Findings show that distinguishing racist discourse from condemnatory or metalinguistic discussions remains challenging, underscoring risks of harmful false positives in real deployments. These models should therefore support, rather than replace, human oversight.

The dataset will be released under a responsible disclosure protocol, balancing transparency with safeguards against misuse and accompanied by clear documentation. By foregrounding ethical

risks alongside technical outcomes, this work supports socially accountable, context-aware NLP development.

Acknowledgments

This work was 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

Sílvio Almeida. 2019. *Racismo Estrutural*. Pólen Produções, São Paulo.

Gabriel Assis, Annie Amorim, Jonnatahn Carvalho, Daniel de Oliveira, Daniela Q. C. Vianna, and Aline Paes. 2024. Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In *Proc. of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 301–311.

Marcelo Augusto. 2021. Twitter analysis. GitHub repository.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Paulo Henrique Caetano. 2020. *A Palavra-Chave "Racismo" e suas Relações Lexicais no Discurso Brasileiro*. Ph.D. thesis, Universidade Federal de Minas Gerais.

Bruno Cascalheira and et al. 2024. Annotation practices for socially situated language. In *Proc. of the Brazilian Conference on Intelligent Systems (BRACIS)*.

William G. Cochran. 1977. *Sampling Techniques*, 3rd edition. Wiley.

Frantz Fanon. 2008. *Pele Negra, Máscaras Brancas*. EDUFBA, Salvador.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Stuart Hall, editor. 1997. *Representation: Cultural Representations and Signifying Practices*. SAGE Publications, London.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proc. of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) and the 10th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Eloyna Augusta Mesquita Miranda. 2020. As religiães de matriz africana e o racismo religioso no Brasil. Master's thesis, Universidade Federal da Bahia.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 85–94, New York, NY, USA. Association for Computing Machinery.

Cendra Devayana Putra and Hei-Chia Wang. 2024. Advanced bert-cnn for hate speech detection. *Procedia Comput. Sci.*, 234(C):239–246.

RD Station. 2025. As redes sociais mais usadas no Brasil e no mundo em 2025. Accessed: 2025-06-11.

Marcelo Augusto Araújo dos Reis. 2021. *Predição de Comentários em Mídias Sociais sobre Discursos Racistas*. Ph.D. thesis, Universidade de Brasília.

Brenda Salenave Santana, Aline Aver Vanin, and Leandro Krug Wives. 2022. Sexist hate speech: Identifying potential online verbal violence instances. In *Proc. of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 177–187.

Xinyue Shen, Yixin Wu, Yiting Qu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025. Hatebench: benchmarking hate speech detectors on llm-generated content and hate campaigns. In *Proceedings of the 34th USENIX Conference on Security Symposium*, SEC '25, USA. USENIX Association.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 687–690.

Tarcízio Silva. 2022. *Racismo Algorítmico: Inteligência Artificial e Discriminação nas Redes Digitais*. Edições SESC, São Paulo.

Sebastião Rogério da Silva Neto. 2017. Uma abordagem computacional para identificação de indício de preconceito em textos baseada em análise de sentimentos. Master's thesis, Universidade Federal de Alagoas.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Proc. of the International Conference on Artificial Neural Networks (ICANN)*, pages 403–417.

Douglas Trajano, Rafael Bordini, and Renata Vieira. 2022. OLID-BR: Offensive language identification dataset for brazilian portuguese. *Research Square Preprint*. Preprint.

Teun A. van Dijk. 2008. *Discourse and Power*. Palgrave Macmillan.