Detecção de Discursos Racistas em Português na Rede Social X

João Vítor Vaz[1], Fabricio Benevenuto[1], Jussara M. Almeida[1], Marcos André Gonçalves[1], Marisa Vasconcelos[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais

{joaovitorvaz, fabricio, jussara}@dcc.ufmg.br

{mgoncalv, marisavasconcelos}@dcc.ufmg.br

Abstract. *Racism manifests in complex ways on social media, requiring effective approaches for automated detection. This study contributes by building a new dataset of racism annotated by Black researchers, ensuring representativeness in labeling racist posts aimed at the Black population on platform X. We evaluate the performance of traditional machine learning models (Naive Bayes, Logistic Regression, Random Forest and XGBoost) and Transformer-based models, such as BERTimbau, designed for Portuguese. While BERTimbau achieved a reasonably effective F1 score of 0.83, it did not outperform simpler models like Logistic Regression and Naive Bayes. The results highlight challenges in automated detection of online racism, such as the lack of annotated data and linguistic complexities, including irony, sarcasm, and ambiguities. Error analyses reveal that these aspects indeed impact the classifier effectiveness, suggesting the need for more robust approaches to identify racism in Portuguese.*

Resumo. *O racismo se manifesta de maneiras complexas nas mídias sociais, exigindo abordagens eficazes para detecção automatizada. Este estudo contribui construindo um novo conjunto de dados de racismo anotado por pesquisadores negros, garantindo representatividade na rotulagem de postagens racistas direcionadas à população negra na plataforma X. Avaliamos o desempenho de modelos de aprendizado de máquina tradicionais (Naive Bayes, Regressão Logística, Random Forest e XGBoost) e modelos baseados em Transformer, como BERTimbau, voltado para a língua Portuguesa. Embora BERTimbau tenha alcançado uma pontuação F1 de 0,83, indicando razoável eficácia, não superou modelos mais simples, como Regressão Logística e Naive Bayes. Os resultados evidenciam desafios na detecção automatizada de racismo online, como a escassez de dados anotados e as complexidades linguísticas, incluindo ironia, sarcasmo e ambiguidades. Análises de erros revelam que esses fatores de fato impactam a eficácia dos classificadores, sugerindo a necessidade de métodos mais robustos para identificar o racismo em português.*

Aviso de conteúdo: Este artigo contém exemplos de frases racistas. As postagens incluídas exemplificam os desafios encontrados no processo de classificação dos dados.

1. Introdução

As redes sociais desempenham um papel central na disseminação de ideias, opiniões e experiências. Entre elas, o X[1] – anteriormente *Twitter* – se destaca como uma das

[1]https://x.com/

plataformas mais dinâmicas e amplamente utilizadas para interação social. No Brasil, cerca de 22 milhões de pessoas acessam a plataforma anualmente [RD Station 2025], o X se consolidou como um espaço relevante para debates públicos e mobilizações sociais. No entanto, essa estrutura favorece a circulação de informações também possibilita a propagação de discursos de ódio [Rotoli 2023], incluindo manifestações de racismo.

O racismo, entendido como um sistema estrutural que perpetua desigualdades com base em critérios raciais [Almeida 2019], se manifesta de diversas formas no ambiente digital. Comentários, postagens e interações em redes sociais reforçam estereótipos, promovem exclusões e legitimam práticas discriminatórias. No contexto brasileiro, onde as desigualdades raciais são historicamente enraizadas e estruturalmente postas, o estudo desses discursos em plataformas digitais se torna essencial para compreender sua recorrência, impacto e possíveis estratégias de detecção de racismo.

Apesar dos avanços na detecção de discurso de ódio, estudos focados especificamente no racismo ainda são limitados, sobretudo na língua portuguesa. O principal desafio está na escassez de conjunto de dados anotados e na complexidade da linguagem usada nas redes sociais, onde o discurso racista pode ser sutil, irônico ou ambíguo. Além disso, modelos de aprendizado de máquina costumam falhar na captura dessas nuances, comprometendo sua eficácia na moderação de conteúdo.

Nesse contexto, este estudo apresenta duas principais contribuições. Primeiro, propomos a construção de um *conjunto de dados anotado por pesquisadores negros*, adotando uma abordagem com representatividade na rotulação. Ao incorporar vivências diretas do racismo, buscamos maior sensibilidade na identificação de conteúdos sutis e ambíguos, reduzindo vieses e aprimorando a acurácia. Segundo, realizamos uma *análise comparativa* de diferentes modelos de aprendizado de máquina, incluindo abordagens clássicas e baseadas em *Transformers*, discutindo desafios, limitações e caminhos para classificadores mais eficazes no contexto brasileiro. Nosso trabalho contribui para o avanço no entendimento do racismo digital na língua portuguesa, preenchendo lacunas na literatura e fornecendo subsídios para o desenvolvimento de estratégias mais eficazes de mitigação do problema nas redes sociais.

2. Trabalhos Relacionados

Diversos estudos têm sido conduzidos para identificar e classificar postagens que representam ataques às vivências de minorias sociais em diferentes plataformas digitais, visando aprimorar a experiência desses usuários. Esses estudos adotam diferentes abordagens, técnicas e algoritmos para alcançar seus objetivos.

Silva *et al.* [Silva et al. 2018] conduziram um estudo sobre racismo em língua portuguesa no Twitter e encontraram que o *Naive Bayes* foi o algoritmo com melhor desempenho na tarefa. Reis *et al.* [Reis 2021], por outro lado, desenvolveram um modelo de aprendizado de máquina para detectar racismo no X e concluíram que a Regressão Logística apresentou melhor desempenho. Essa divergência pode estar relacionada a diferenças na base de dados utilizadas, no critério de anotação, nas métricas de avaliação aplicadas, ou até mesmo no rigor estatístico adotado na comparação, evidenciando a necessidade de investigações adicionais robustas sobre quais fatores impactam o desempenho dos classificadores nesse contexto.

Criss *et al.* [Criss et al. 2021] analisaram interações racistas no X, considerando

ataques diretos, insultos e microagressões. O estudo destacou a influência do contexto e das redes sociais na interpretação das postagens, exigindo adaptação dos modelos de processamento de linguagem natural. Para melhorar a precisão e evitar reforço de estigmas, propuseram a criação de bases de dados mais representativas, com especialistas e pessoas afetadas, garantindo anotações alinhadas às experiências das minorias.

Putra *et al.* [Putra and Wang 2024] propuseram um modelo para detectar discursos de ódio em mídias sociais, combinando redes neurais convolucionais (CNNs) avançadas com *embeddings* do BERT. Usando os conjuntos de dados *Davidson* e TRAC-1, mostraram que essa abordagem superou métodos tradicionais. O estudo reforça a eficácia do aprendizado profundo na identificação de conteúdos nocivos.

Em [Cascalheira et al. 2024], os autores desenvolveram dois conjuntos de dados, MiSSoM e MiSSoM+, sobre estresse em minorias LGBTQ+ a partir de *subreddits*. Eles incluem rótulos manuais, atributos psicossociológicos e palavras-chave clínicas. A metodologia adotada, pautada em rigor ético, contribui para a criação de bases mais diversificadas. Embora o foco seja distinto, o estudo ilustra como bases bem estruturadas podem auxiliar na pesquisa sobre grupos marginalizados, incluindo vítimas de discriminação racial, reforçando a importância de anotações feitas por pessoas da própria comunidade.

A principal contribuição deste estudo é a construção de um conjunto de dados anotado exclusivamente por pesquisadores negros, oferecendo uma perspectiva mais sensível às nuances do racismo em português. Também realizamos uma análise detalhada dos erros dos modelos, destacando dificuldades com ironia, sarcasmo e ambiguidades linguísticas, pouco exploradas em estudos anteriores. Essa análise orienta a criação de novas bases, o aprimoramento dos modelos e o desenvolvimento de estratégias mais eficazes de moderação automática em português.

Apesar das limitações inerentes à complexidade da linguagem e à dificuldade dos modelos em captar sutilezas discursivas, esta pesquisa abre caminho para abordagens mais justas na detecção automatizada do racismo. Ao combinar a representatividade dos anotadores com análises qualitativas dos erros, fortalecemos a robustez das ferramentas de moderação, contribuindo para mitigar o racismo online e promover ambientes digitais mais inclusivos.

3. Fundamentação Teórica: Racismo e Linguagem Digital

O racismo direcionado a pessoas negras, longe de se restringir a atos isolados, configura-se como um sistema opressor multifacetado. Esse sistema impõe a "epidermização da inferioridade", hierarquizando fenótipos e definindo o indivíduo com base na cor da pele e características físicas [Fanon and da Silveira 2008]. Como consequência, ocorre a internalização de uma autoimagem negativa e a busca por padrões eurocêntricos. Nessa lógica racista, traços mais próximos desses padrões são valorizados, enquanto aqueles considerados *africanizados* são marginalizados. Essa hierarquização reforça as estruturas de opressão, perpetuando a discriminação e a desigualdade.

O racismo estabelece uma hierarquia em que a branquitude é tomada como norma. Isso subordina pessoas negras, restringindo seu acesso a oportunidades e contribuindo para a perpetuação da desigualdade. No plano cultural, esse sistema não somente marginaliza a cultura negra, mas também a relega à condição de inferioridade, ao mesmo tempo

em que impõe valores e padrões eurocêntricos como normativos. O racismo, portanto, atravessa as relações sociais, molda a subjetividade e permeia a experiência cotidiana das pessoas negras, sustentando um ciclo contínuo de opressão e violência simbólica.

No contexto digital, essa hierarquização se manifesta em discursos racistas que podem assumir formas explícitas, como insultos e ataques diretos, ou implícitas, por meio de ironia, estereótipos, negação do racismo e naturalização de desigualdades. A persistência desses discursos nas redes sociais reforça mecanismos de exclusão e contribui para a marginalização de usuários negros.

Tais características impõem desafios à detecção automática do racismo. Modelos de aprendizado de máquina podem enfrentar desafios ao identificar discursos discriminatórios quando o racismo se apresenta de maneira sutil, contextual ou ambígua. Para lidar com essa complexidade, este estudo adota uma abordagem que considera tanto manifestações explícitas quanto implícitas do racismo, garantindo que a anotação dos dados reflita a diversidade de estratégias discursivas usadas para perpetuar a discriminação.

4. Metodologia

Nesta seção, descrevemos o processo metodológico usado na construção da base de dados.

4.1. Coleta de Dados

A construção de uma base de dados rotulada para a detecção de discursos racistas apresenta desafios metodológicos significativos, especialmente na identificação de mensagens relevantes em um ambiente online vasto e diversificado. Devido à especificidade e à distribuição não homogênea do fenômeno estudado, optou-se por uma abordagem intencional em vez de métodos probabilísticos de amostragem.

Métodos probabilísticos, que garantem que cada elemento da população tenha uma chance conhecida e diferente de zero de ser selecionado, como a amostragem aleatória simples ou estratificada [Cochran 1977], não se mostraram adequados para a coleta de discursos racistas, pois a ocorrência dessas mensagens não é distribuída de maneira homogênea na rede. Assim, utilizamos um método baseado em palavras-chave para identificar postagens e comentários potencialmente relevantes.

A coleta foi realizada entre setembro de 2024 e janeiro de 2025, como o apoio de um serviço que acessa a API do X via modelo de acesso pago, limitado a postagens públicas. O serviço retorna os dados com informações detalhadas sobre as postagens e seus respectivos usuários. Para garantir a proteção dos dados sensíveis, identificadores diretos foram removidos, mantendo apenas os atributos textuais das postagens e seus comentários.

A filtragem inicial foi realizada com base nas palavras-chave: “mulambo”, “racismo”, “racista”, “angolano”, “angolana”, “candomblé”, “africano”e “africana”, escolhidas com base em estudos que investigam manifestações de racismo no Brasil [Miranda 2020, Caetano 2020], garantindo que os termos selecionados refletem as particularidades do racismo no contexto brasileiro. Foram coletadas postagens e seus comentários associados, correspondendo a respostas diretas de outros usuários. Essas interações foram incluídas para capturar melhor o contexto e a dinâmica das discussões.

No total foram coletadas 549 postagens, incluindo tanto publicações originais quanto comentários. A seleção final para anotação envolveu uma triagem manual realizada por um pesquisador para remover conteúdos irrelevantes e priorizar textos com potencial relação com o tema, como indícios de linguagem ofensiva, estereotipada ou discriminatória. Essa triagem, no entanto, não constituiu um processo de pré-rotulação. Também foram incluídas postagens com linguagem ambígua ou discussões sobre racismo sem teor ofensivo, buscando proporcionar um contexto mais amplo para a análise.

4.2. Anotação dos Dados por Especialistas

Para garantir consistência na rotulação, os anotadores receberam a seguinte definição de racismo, adaptada para o foco deste estudo em racismo contra pessoas negras: "Racismo é definido como qualquer forma de expressão, direta ou indireta, que manifeste a hierarquização de fenótipos, a internalização de autoimagem negativa ou a subordinação de pessoas negras. Isso inclui discursos que inferiorizam, ofendem, discriminam ou estigmatizam pessoas negras com base em sua raça, ancestralidade ou características fenotípicas, seja de forma explícita (insultos, ataques diretos) ou implícita (ironia, estereótipos, negação do racismo, etc)." Com base nessa definição, as postagens foram classificadas em duas classes: **classe 1 (racista)** e **classe 0 (não racista)**.

A rotulação das postagens foi conduzida por oito pesquisadores negros/as, organizados em quatro duplas. Cada dupla recebeu um subconjunto das postagens, realizando a anotação de forma independente. Os pesquisadores possuem formações diversas, são oriundos de diferentes estados do Brasil e apresentam distintas características fenotípicas, assegurando diversidade e representatividade na análise. A rotulação seguiu o modelo de consenso, com práticas de concordância inter-anotador inspiradas na proposta de [Cascalheira et al. 2024], originalmente aplicada à anotação de dados de minorias LGBTQIAPN+, e adaptadas neste estudo para garantir consistência e sensibilidade na identificação de conteúdos racistas.

Das 425 postagens analisadas nessa etapa, houve concordância total (100%) entre os anotadores em 386 casos. O coeficiente Kappa de Cohen calculado entre pares indicou um bom nível de concordância, com valor mínimo de 0,655. As 39 postagens com discordância foram submetidas à revisão de um terceiro avaliador independente, que seguiu os mesmos critérios. Como resultado, 16 dessas postagens foram reclassificadas como racistas, totalizando 113 postagens classificadas como racistas e 312 como não racistas nessa base construída.

4.3. Incorporação de Dados da Literatura

Para ampliar o conjunto de dados, foram incorporadas 629 postagens rotuladas como racistas em estudos anteriores [Fortuna et al. 2019, Leite et al. 2020, Augusto 2021, Silva Neto et al. 2017]. No entanto, uma análise detalhada revelou que esses estudos adotaram definições mais amplas, englobando discursos de ódio direcionados a diferentes grupos minoritários.

Para garantir a consistência metodológica, todas as 629 postagens foram reavaliadas e re-rotuladas por um dos autores deste estudo, de acordo com a definição restrita de racismo contra pessoas negras adotada neste estudo. Após a reclassificação, 322 postagens foram confirmadas como racistas.

Com isso, o conjunto final de dados rotulados como racistas é composto por 107 postagens da base construída neste estudo e 322 da reanotação da literatura, totalizando 429 postagens racistas. Já a classe de não racistas contém 318 postagens provenientes exclusivamente da base construída pelos especialistas deste estudo. Os dados utilizados neste estudo estão publicamente disponíveis[2], a fim de incentivar a reprodutibilidade e estudos futuros.

4.4. Pré-Processamento dos Dados

Para garantir a qualidade dos dados utilizados na etapa de treinamento e avaliação dos modelos, foi realizado um pré-processamento das postagens coletadas. Foram removidos *URLs*, menções a usuários e caracteres especiais desnecessários. Em seguida, os textos foram convertidos para letras minúsculas, visando padronizar a entrada dos modelos.

O pré-processamento variou de acordo com o tipo de modelo empregado. Para os modelos clássicos de aprendizado de máquina, os textos passaram por lematização e tokenização, garantindo maior uniformidade na estrutura dos textos e facilitando a extração de padrões linguísticos. Já para os modelos baseados em *Transformers* – BERTimbau e RoBERTa – essas etapas não foram aplicadas, uma vez que esses modelos possuem tokenizadores próprios que segmentam e normalizam os textos de forma otimizada para aprendizado contextual.

Essas etapas garantiram que os dados fossem estruturados de forma eficiente, permitindo que os modelos focassem nas características semânticas mais relevantes para a identificação de discursos racistas. A remoção de elementos irrelevantes e a normalização textual aprimoraram a precisão dos modelos, contribuindo para um desempenho mais robusto na classificação das postagens.

4.5. Limitações dos Dados

A construção de bases de dados rotuladas para detecção de discurso racista apresenta desafios inerentes, especialmente relativos ao tamanho da amostra e à representatividade dos dados. Como qualquer estudo que lida com fenômenos sociais complexos, há dificuldades na obtenção de um conjunto diversificado que reflita todas as nuances da realidade.

A base utilizada foi construída a partir de múltiplas fontes, incluindo coletas utilizando a API e dados da literatura, garantindo uma composição heterogênea e representativa nas possibilidades metodológicas adotadas. Ainda assim, reconhece-se que ampliar a quantidade e a diversidade dos exemplos pode contribuir para a evolução dos modelos.

É importante destacar que questões de representatividade de dados são recorrentes em pesquisas sobre aprendizado de máquina aplicado a problemas sociais. Este estudo não apenas mapeia esses desafios, mas também estabelece um ponto de partida para futuras expansões. Assim, longe de ser um fator limitante, essas questões reforçam a necessidade contínua de aprimoramento e validação das abordagens na área.

4.6. Considerações Éticas

Neste estudo, foram adotadas medidas rigorosas para garantir a coleta e anotação ética dos dados. As postagens foram anonimizadas para proteger a privacidade dos usuários.

[2]Disponível em: https://github.com/joaovitorvaz/Base-de-Dados-Racistas-da-Rede-Social-X

A anotação, conduzida por especialistas negros, incorporou perspectivas autênticas sobre o racismo. Buscou-se diversidade entre os anotadores para minimizar vieses e tornar a rotulagem mais representativa. Assim, os modelos foram treinados em um conjunto de dados mais robusto, aprimorando a detecção das diversas formas de discurso racista.

5. Modelos de Classificação

Apresentamos os modelos de classificação utilizados, detalhando suas principais características e implementação.

Naive Bayes. Modelo probabilístico que assume independência condicional entre as características, ou seja, considera que a presença de uma palavra não influencia a probabilidade de ocorrência de outra dentro da mesma classe [Zhang 2004]. Utilizou-se a versão *Multinomial Naive Bayes*, implementada no *scikit-learn* com suavização de *Laplace*. A vetorização dos textos utilizou *CountVectorizer* com unigramas e bigramas para capturar o contexto local das palavras.

Regressão Logística. Modelo linear amplamente utilizado para classificação binária, que equilibra desempenho e interpretabilidade [Gonzalez 2018]. Diferentemente do *Naive Bayes*, não pressupõe independência entre as características, sendo mais apropriado para cenários onde há correlações entre variáveis. Utilizamos a implementação do *scikit-learn*, com o solver `liblinear`, otimizado para dados de alta dimensionalidade e regularização `L2` para mitigar *overfitting*.

Random Forest. Consiste em um conjunto de árvores de decisão, combinando previsões individuais por meio de um processo de votação para melhorar a precisão e reduzir *overfitting* [Breiman 2001]. Neste estudo, utilizou-se a implementação da biblioteca *scikit-learn*, com um número de 100 árvores e critério de divisão `gini`. A vetorização do texto foi realizada com *TF-IDF*, considerando unigramas e bigramas.

XGBoost. Modelo baseado em *Gradient Boosting*, amplamente utilizado para classificação devido ao seu desempenho e eficiência computacional [Chen and Guestrin 2016]. A implementação utilizou a biblioteca *XGBoost*, com número de árvores ajustado para 200 e profundidade máxima de 6. O aprendizado foi otimizado com a função de perda `logloss` e taxa de aprendizado de 0,1.

BERTimbau. Modelo baseado na arquitetura *Transformer* treinado para a língua portuguesa, capaz de capturar contexto bidirecional e relações de longo alcance entre palavras [Souza et al. 2020]. Essa capacidade é essencial para detectar discursos racistas, frequentemente manifestados de forma implícita e sutil. Utilizou-se a versão pré-treinada disponível na biblioteca *Hugging Face*, realizando *fine-tuning* com o conjunto de dados. O pré-processamento envolveu tokenização e truncagem para 128 tokens.

RoBERTa. Assim como o BERTimbau, RoBERTa utiliza a arquitetura *Transformer*. RoBERTa é uma otimização do modelo BERT que permite capturar contexto bidirecional e relações de longo alcance entre as palavras [Liu et al. 2019]. Para este estudo, utilizou-se o modelo *roberta-base*, e foi feito o *fine-tuning* com o nosso conjunto de dados (de treino). O pré-processamento do RoBERTa envolveu tokenização e truncagem para 128 tokens, adaptando os dados para o melhor desempenho deste modelo.

6. Resultados e Avaliação dos Modelos

Avaliamos o desempenho de seis modelos de classificação na detecção de discurso racista, utilizando acurácia, precisão, *recall* e Macro F1-*score* como métricas. A metodologia experimental incluiu divisão estratificada dos dados, validação cruzada de 5 *folds*, otimização de hiperparâmetros com *Grid Search* e aplicação de *Random Undersampling* para lidar com o desbalanceamento de classes. A Tabela 1 apresenta a média das métricas de desempenho dos modelos ao longo das cinco partições (*folds*) de validação cruzada.

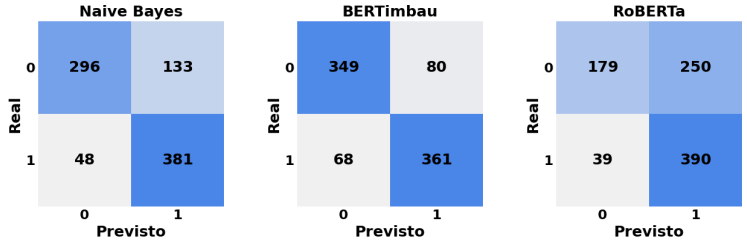
Tabela 1. Desempenho Médio dos Modelos de Classificação

Modelo	Acurácia	Precisão	Recall	F1-score
Regressão Logística	0,7739	0,7491	0,8299	0,7961
XGBoost	0,7821	0,8120	0,7343	0,7711
Random Forest	0,7867	0,7687	0,8231	0,7936
Naive Bayes	0,7891	0,7427	0,8882	0,8083
BERTimbau	0,8275	0,8217	0,8412	0,8299
RoBERTa	0,6632	0,6142	0,9093	0,7286

Para avaliar a significância estatística das diferenças entre os modelos, foi aplicado um teste t pareado nos F1-*scores* dos cinco *folds* de teste para cada par de modelos. A Tabela 2 apresenta os resultados dessa análise, e as Figuras 1 e 2 mostram as matrizes de confusão para cada modelo, detalhando seu desempenho por classe.

Tabela 2. Comparação Par a Par entre Modelos

Modelo 1	Modelo 2	Estatística t	Valor p	Resultado
BERTimbau	RoBERTa	3.5605	0.0236	BERTimbau significativamente melhor
Regressão Logística	*XGBoost*	2.8026	0.0487	Regressão Logística significativamente melhor
Regressão Logística	*Random Forest*	-1.3854	0.2382	Empate
Regressão Logística	*Naive Bayes*	-1.1552	0.3123	Empate
Regressão Logística	BERTimbau	-1.2483	0.2800	Empate
Regressão Logística	RoBERTa	3.5233	0.0244	Regressão Logística significativamente melhor
XGBoost	*Random Forest*	-4.9882	0.0076	*Random Forest* significativamente melhor
XGBoost	*Naive Bayes*	-3.7556	0.0198	*Naive Bayes* significativamente melhor
XGBoost	BERTimbau	-2.3191	0.0812	Empate
XGBoost	RoBERTa	3.8893	0.0177	*XGBoost* significativamente melhor
Random Forest	*Naive Bayes*	-0.0846	0.9366	Empate
Random Forest	BERTimbau	-0.8520	0.4422	Empate
Random Forest	RoBERTa	5.5106	0.0053	*Random Forest* significativamente melhor
Naive Bayes	BERTimbau	-1.0351	0.3591	Empate
Naive Bayes	RoBERTa	5.0415	0.0073	*Naive Bayes* significativamente melhor



a) *Naive Bayes* b) BERTimbau c) *RoBERTa*

Figura 1. Matrizes de confusão - *Naive Bayes*, BERTimbau e RoBERTa (classe 1: racista; classe 0: não-racista).

Os resultados evidenciam alguns aspectos importantes no desempenho dos modelos. Embora o BERTimbau tenha apresentado os maiores valores absolutos de acurácia (0,827) e F1-*score* (0,8299), essas diferenças **não foram** são estatisticamente significativas em relação aos modelos mais simples. Dentro das condições experimentais deste estudo, *não é possível atestar*, que a eficácia do BERTimbau é superior para a detecção de discurso racista. De qualquer forma, o desempenho do modelo é bastante satisfatório, e este desempenho pode ser atribuído ao seu treinamento em um extenso corpus em português, permitindo uma melhor adaptação ao idioma e suas sutilezas. Sua matriz de confusão (Figura 1b) revela um equilíbrio entre verdadeiros positivos (361) e verdadeiros negativos (349), com um número reduzido de falsos positivos (80) e falsos negativos (68).

O modelo RoBERTa apresentou um desempenho que exige análise cuidadosa. Embora tenha demonstrado alta recuperação de instâncias positivas (*recall* de 0,9093), essa sensibilidade veio acompanhada de baixa precisão (0,6142), resultando em muitos falsos positivos—250 postagens não racistas classificadas erroneamente (Figura 1c). Esse padrão pode indicar sensibilidade excessiva a certos traços textuais correlacionados, mas não exclusivos, ao discurso racista. Assim, apesar do *recall* elevado, a baixa precisão compromete sua aplicação em cenários que exigem mínima taxa de falsos positivos.

a) *Random Forest* b) Regressão Logística c) *XGBoost*

Figura 2. Matrizes de confusão - *Random Forest*, Regressão Logística e *XGBoost* (classe 1: racista; classe 0: não-racista).

Entre os modelos tradicionais, o *Naive Bayes* se destacou pelo maior *recall* (0,8882), sendo estatisticamente superior ao *XGBoost* e estatisticamente equivalente aos demais. Esses resultados indicam uma alta sensibilidade do *Naive Bayes* na detecção de discurso racista, embora com menor precisão (0,7427). Isso sugere que o modelo identificou mais instâncias de discurso racista, isto é, maior número de verdadeiros positivos (381), mas gerou um número considerável de falsos positivos (133) (Figura 1a). O *XGBoost*, por sua vez, obteve uma maior precisão (0,8120), minimizando falsos positivos (73), mas com um *recall* inferior (0,7343), o que evidencia o fato de que o modelo deixou passar algumas instâncias, ou seja, falsos negativos (114), como pode ser visto na Figura 2c.

O modelo *Random Forest* apresentou acurácia de 0,7867 e F1-*score* de 0,7936, sendo estatisticamente equivalente aos melhores modelos. Seu elevado *recall* de 0,8231 indica capacidade de detecção de discurso racista. A matriz de confusão (Figura 2a) revela equilíbrio entre verdadeiros positivos (353) e verdadeiros negativos (322), com uma precisão de 0,7687, ligeiramente inferior à do *XGBoost*. Essa característica sugere que o modelo prioriza a identificação da maioria das ocorrências de discurso racista, o que pode ser preferível em cenários onde o custo de não detectar um caso racista é alto.

Por fim, a Regressão Logística apresentou um desempenho competitivo, com

acurácia de 0,7739 e F1-*score* de 0,7961. Seu *recall* de 0,8299 indica boa capacidade de identificar discurso racista, mas sua precisão de 0,7491 sugere um número maior de falsos positivos (121) em comparação a modelos como o *XGBoost* (Figura 2b). Sua matriz de confusão revela um número relativamente baixo de verdadeiros negativos (308), o que sugere que o modelo tende a classificar mais exemplos como discurso racistas, resultando em uma taxa maior de falsos positivos.

7. Análise dos Erros dos Modelos

A análise dos erros nos modelos de classificação de discurso racista revelou desafios significativos na identificação de contextos sutis, ironia, sarcasmo e ambiguidade linguística. Esses aspectos tornam a detecção ainda mais complexa, uma vez que o discurso racista frequentemente se manifesta de forma indireta, utilizando construções linguísticas culturalmente específicas e exigindo um entendimento contextual profundo.

7.1. Erros nos Algoritmos Clássicos

A análise comparativa de modelos de classificação clássicos revelou uma limitação relevante na detecção de discursos racistas implícitos. Embora esses algoritmos demonstrem eficácia na identificação de manifestações explícitas, a capacidade de capturar as nuances sutis e expressões veladas é limitada, resultando em maior incidência de falsos negativos.

A complexidade semântica e contextual da linguagem representa um desafio significativo para esses modelos, especialmente em domínios sensíveis como a detecção de discurso de ódio. Os resultados reforçam essa dificuldade, particularmente na identificação de formas implícitas ou mais sutis de racismo. A Tabela 3 ilustra exemplos de predições incorretas (falsos negativos) obtidas durante os experimentos, em que textos rotulados como racistas foram erroneamente classificados como não racistas.

Tabela 3. Exemplos de falso negativos nos algoritmos clássicos

Texto Original	Rótulo	Predição do Modelo
“Sabe a diferença entre o negro e o câncer? O câncer evolui.”	Racista	Não Racista
“Batom vermelho só é bonito em gente branca.”	Racista	Não Racista
“Por que só tem Globeleza mulata? Isso sim é preconceito!”	Racista	Não Racista
“racismo não existe, tira essa parada da cabeça!!”	Racista	Não Racista
“Tinha que ser preto (nada contra))))”	Racista	Não Racista

A análise dos erros revela padrões recorrentes, mostrando limitações dos modelos:

- **Ironia e Ambiguidade:** Frases como “Sabe a diferença entre o negro e o câncer? O câncer evolui.” utilizam comparações irônicas, cuja intenção discriminatória não é reconhecida pelos modelos.
- **Generalizações e Estereótipos:** Sentenças como “Batom vermelho só é bonito em gente branca.”, apresentam uma generalização ofensiva e são interpretadas de forma literal, ignorando o contexto sociocultural discriminatório.
- **Negação e Linguagem Suavizada:** Expressões como “racismo não existe, tira essa parada da cabeça!!” e “Tinha que ser preto (nada contra))))” demonstram tentativas de minimizar ou negar o racismo, suavizando o discurso discriminatório. Esse tipo de linguagem dificulta a identificação do conteúdo racista subjacente, fazendo com que os modelos confundam tais afirmações com neutralidade ou tentativa de atenuação.

Esses resultados evidenciam a dificuldade dos modelos em compreender as complexidades semânticas e contextuais da linguagem, especialmente relativas a nuances sutis em discursos discriminatórios. Essa limitação é particularmente relevante em tarefas de detecção de discurso racista, na qual a intenção discriminatória nem sempre é explicitada, se ocultando sob ironias, estereótipos ou tentativas de suavização. Para superar esses desafios, é necessário um maior refinamento na captura de contextos mais amplos e uma compreensão mais profunda dos mecanismos subjacentes ao discurso discriminatório.

7.2. Erros nos Algoritmos Baseados em *Transformers*

Modelos baseados em *Transformers* apresentaram resultados promissores na classificação de discurso racista, demonstrando uma notável capacidade de capturar contextos complexos e analisar relações de longo alcance entre as palavras. Contudo, esses modelos ainda enfrentam desafios consideráveis ao lidar com expressões ambíguas, ironias, entre outros.

Entre os modelos analisados, observamos que o RoBERTa apresentou um padrão específico de erro: uma tendência a classificar erroneamente instâncias não racistas como racistas. Essa alta taxa de falsos positivos sugere que o modelo tem dificuldade em diferenciar frases que compartilham certas palavras-chave, mas que nem sempre possuem teor discriminatório. Essa limitação pode estar relacionada ao fato de que o RoBERTa ser pré-treinado predominantemente em inglês, diferentemente do BERTimbau, que foi ajustado para o português com um conjunto de textos mais representativos da língua. Apesar do RoBERTa ser um modelo mais efetivo do que o BERT [Andrade et al. 2023], no qual o BERTimbau é baseado, a adaptação à língua se mostrou mais importante para a tarefa alvo do que a diferença de efetividade dos modelos base pré-treinados.

Além disso, o desempenho desses modelos pode ser afetado por limitações nos dados de treinamento, incluindo a sub-representação de variações linguísticas específicas. Isso compromete a identificação precisa de discursos racistas em certos contextos socioculturais, tornando-os menos eficazes na detecção de manifestações discriminatórias que fogem de padrões mais explícitos. Para mitigar essas limitações, futuros estudos poderiam explorar LLMs treinados com dados em português ou modelos multilíngues, como o XLM-RoBERTa [Conneau et al. 2020], que podem capturar melhor as nuances da língua. A Tabela 4 apresenta exemplos de previsões incorretas (falsos negativos) onde textos rotulados como racistas foram incorretamente classificados como não racistas por modelos baseados em *Transformers*.

Tabela 4. Exemplos de erros de classificação

Texto Original	Rótulo	Predição do Modelo
"Fdx esta merda não devia estar toda em Luanda?"	Racista	Não Racista
"Nossa, que beleza exótica você tem!"	Racista	Não Racista
"Falei morena porque se falasse negra você ficaria constrangida."	Racista	Não Racista
"o meu mundo é bem real, já o seu parece conto de fadas, já que a vida de negro é bem mais fácil."	Racista	Não Racista
"Nada contra negros mas EU PREFIRO aquele branquelo palmito cor de leite pra bejar na boca rsrs"	Racista	Não Racista

A análise desses erros revela desafios específicos que impactam o desempenho dos modelos baseados em *Transformers* na classificação de discurso racista:

- **Dualidade de significados:** Comentários que, à primeira vista, parecem elogios, como "Nossa, que beleza exótica você tem!", podem carregar uma conotação desumanizadora. O termo "exótica" reduz a identidade de uma pessoa a características físicas, um subtexto que o modelo não captura sem uma análise contextual mais profunda.
- **Ambiguidade linguística:** Frases como "Falei morena porque se falasse negra você ficaria constrangida." ilustram como a ambiguidade pode levar a interpretações errôneas, com o modelo falhando em reconhecer a intenção discriminatória.
- **Sarcasmo e ironia:** Comentários como "O meu mundo é bem real, já o seu parece conto de fadas, já que a vida de negro é bem mais fácil." carregam um tom irônico que os modelos podem não reconhecer, levando a classificações imprecisas.
- **Preconceito disfarçado:** Frases como "Nada contra negros, mas..."introduzem uma negação superficial do preconceito antes de afirmar uma visão discriminatória. O modelo pode interpretar essa negação como uma ausência de viés.

Esses desafios reforçam a complexidade do discurso racista, que frequentemente se manifesta por meio de construções linguísticas indiretas e referências específicas. Para uma identificação mais precisa, seria necessário refinar os modelos, incorporando técnicas que aprimorem a compreensão do contexto e das implicações socioculturais do discurso.

8. Conclusão e Trabalhos Futuros

Este estudo investigou a detecção de discursos racistas direcionados à população negra na plataforma X, aliando técnicas de aprendizado de máquina com a construção de uma base de dados anotada por especialistas negros. A integração dessa base com outras da literatura garantiu maior representatividade e permitiu avaliar diferentes modelos classificadores.

Os resultados mostraram que modelos treinados com dados representativos podem ser eficazes na detecção de discursos racistas, permitindo respostas rápidas na remoção de conteúdos ofensivos e promovendo um ambiente digital mais seguro e inclusivo. No entanto, desafios como a distinção entre discursos racistas e não racistas evidenciam as limitações dos modelos, especialmente na captação de nuances linguísticas. A análise de erros revelou padrões que podem guiar futuras pesquisas.

Como direções futuras, propõe-se uma análise mais detalhada dos casos em que mensagens não racistas foram classificadas como racistas, para compreender suas causas e reduzir falsos positivos. Também é importante ampliar o corpus, o que pode beneficiar tanto os modelos tradicionais quanto aqueles baseados em *Transformers*, contribuindo para maior robustez e representatividade. A adoção de LLMs nas análises pode melhorar a interpretação e contexto das mensagens.Além disso, o uso de mensagens contextuais pode ajudar a identificar situações em que o significado depende de interações anteriores.

Este trabalho destaca a importância da detecção de discursos racistas, oferecendo uma análise mais precisa e informada. Ao abordar desafios como a interpretação contextual e a representatividade dos dados, abre caminho para novas abordagens, contribuindo para soluções mais eficazes no combate ao racismo digital e reforçando a necessidade de sensibilidade cultural nas tecnologias voltadas à justiça social.

9. Agradecimentos

Agradecemos o apoio institucional do CNPq, da FAPEMIG e do INCT-TILD-IAR.

Referências

Almeida, S. (2019). *Racismo estrutural*. Pólen Produção Editorial LTDA.

Andrade, C., Belém, F., Cunha, W., França, C., Viegas, F., Rocha, L., and Gonçalves, M. (2023). On the class separability of contextual embeddings representations - or "the classifier does not matter when the (text) representation is so good!". *Inf. Process. Manag.*, 60(4):103336.

Augusto, M. (2021). Twitter analysis. `https://github.com/maugustoo/twitter_analysis`.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Caetano, P. H. (2020). *A palavra-chave racismo e suas relações lexicais: uma análise crítica dos discursos sobre relações raciais brasileiras em corpus de jornal impresso*. Tese (doutorado em ciências humanas), Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

Cascalheira, C., Chapagain, S., Flinn, R., Klooster, D., Laprade, D., Zhao, Y., Lund, E., Gonzalez, A., Corro, K., Wheatley, R., et al. (2024). The LGBTQ+ minority stress on social media (missom) dataset: A labeled dataset for natural language processing and machine learning. In *International AAAI Conference on Web and Social Media*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Cochran, W. (1977). *Sampling Techniques*. Wiley publication in applied statistics. Wiley.

Conneau, A., Khandelwal, K., and Goyal (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Criss, S., Michaels, E., Solomon, K., et al. (2021). Twitter Fingers and Echo Chambers: Exploring Expressions and Experiences of Online Racism Using Twitter. *Journal of Racial and Ethnic Health Disparities*, 8:1322–1331.

Fanon, F. and da Silveira, R. (2008). *Pele negra, máscaras brancas*. Editora da UFBA.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Workshop on abusive language online*.

Gonzalez, L. d. A. (2018). Regressão logística e suas aplicações.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of EMNLP-IJCNLP*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Miranda, E. A. M. (2020). As religiões de matriz africana e o racismo religioso no brasil: os velhos e os novos agentes da perseguição ao candomblé na bahia. Dissertação (mestrado em ciências sociais), Universidade Federal da Bahia, Salvador, Brasil.

Putra, C. D. and Wang, H.-C. (2024). Advanced bert-cnn for hate speech detection. *Procedia Computer Science*, 234:239–246. Seventh Information Systems International Conference (ISICO 2023).

RD Station (2025). As redes sociais mais usadas no brasil e no mundo em 2025: com insights, ferramentas e materiais. Acessado em: 11 de março de 2025.

Reis, M. A. A. d. (2021). Predição de comentários em mídias sociais sobre discursos racistas.

Rotoli, L. U. M. (2023). Manifestações populares no twitter, no período de 2012 a 2021, sobre as políticas para reservas de vagas em universidades brasileiras.

Silva, R., Fernandes, D., and Fernandes, M. (2018). Caracterização de mensagens em língua portuguesa com traços de racismo no twitter. In *Anais da VI Escola Regional de Informática de Goiás*. SBC.

Silva Neto, S. R. d. et al. (2017). Uma abordagem computacional para identificação de indício de preconceito em textos baseada em análise de sentimentos.

Souza, R. A., Almeida, J. M., and Gatti, M. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the Brazilian Conference on Artificial Intelligence*.

Zhang, H. (2004). The optimality of naive bayes. *Aa*, 1(2):3.