UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

JOÃO VÍTOR DOS SANTOS VAZ
Orientador: Profº. Dr. Jadson Castro Gertrudes

# DETECÇÃO DE DISCURSOS RACISTAS NO TWITTER: UMA ABORDAGEM BASEADA EM PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZADO DE MÁQUINA

Ouro Preto, MG
2024

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

JOÃO VÍTOR DOS SANTOS VAZ

# DETECÇÃO DE DISCURSOS RACISTAS NO TWITTER: UMA ABORDAGEM BASEADA EM PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZADO DE MÁQUINA

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

**Orientador:** Profº. Dr. Jadson Castro Gertrudes

Ouro Preto, MG
2024

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

# FOLHA DE APROVAÇÃO

**João Vítor dos Santos Vaz**

**Detecção de discursos racistas no Twitter: uma abordagem baseada em Processamento de Linguagem Natural e Aprendizado de Máquina**

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 15 de Janeiro de 2024.

Membros da banca

Jadson Castro Gertrudes (Orientador) - Doutor - Universidade Federal de Ouro Preto
Amanda da Silva Oliveira (Examinadora) - Bacharel - Universidade Federal de Ouro Preto
Valéria de Carvalho Santos (Examinadora) - Doutora - Universidade Federal de Ouro Preto

Jadson Castro Gertrudes, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 15/01/2024.

Documento assinado eletronicamente por **Jadson Castro Gertrudes**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 18/01/2024, às 10:17, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0653962** e o código CRC **137EAE38**.

*Dedico este trabalho a todos aqueles que alimentam meus sonhos.*

# Agradecimentos

À Universidade Federal de Ouro Preto, sou imensamente grato por ter me proporcionado uma educação de qualidade e por ter sido fundamental em minha formação acadêmica. Agradeço a todos os professores e colaboradores que, com dedicação e comprometimento, contribuíram para o meu crescimento e desenvolvimento ao longo dos anos.

À minha querida vó e minha amada mãe, não existem palavras suficientes para expressar minha gratidão. Vocês sempre acreditaram em meu potencial e me mostraram que é possível realizar sonhos, mesmo diante de adversidades. No amor, o apoio e os ensinamentos que recebi de vocês são inestimáveis. Sou eternamente grato por tudo o que fizeram por mim.

À política de ações afirmativas, por me darem a possibilidade de ingressar na graduação, e aos programas de permanência estudantil, que forneceram suporte financeiro durante minha jornada acadêmica. Esses programas foram essenciais para que pessoas como eu pudessem ter uma trajetória acadêmica menos tortuosa.

Aos amigos que conheci durante minha trajetória na UFOP, cada um de vocês deixou uma marca especial em minha vida. Compartilhamos momentos de aprendizado, superação e diversão, e essas memórias serão sempre valorizadas. Obrigado por fazerem parte da minha jornada e por tornarem os dias na universidade mais significativos e especiais.

Ao meu orientador, Jadson Castro, homem preto do mesmo território que eu, que acreditou na minha pesquisa e me auxiliou do início ao fim com toda atenção e carinho. Saio da graduação com um amigo e futuro colega de pesquisa e profissão.

Ao meu amor, que foi uma fonte constante de apoio e força ao longo dessa pesquisa. Sua presença e encorajamento foram fundamentais para minha motivação e perseverança. Obrigado por estar ao meu lado e acreditar em mim.

Por fim, agradeço a Deus por todas as oportunidades, bênçãos e proteções concedidas ao longo dessa jornada. Sua presença em minha vida é uma fonte de fortaleza e gratidão.

A todos que contribuíram direta ou indiretamente para a realização deste trabalho, meu sincero agradecimento. Cada um de vocês teve um papel importante e especial nessa conquista.

# Resumo

O crescimento notável das redes sociais tem sido acompanhado pelo aumento significativo na disseminação de discursos racistas nesses ambientes. Isso tem suscitado um interesse crescente em estudos relacionados a essa problemática. No entanto, há a existência de uma lacuna no âmbito da pesquisa, particularmente no contexto da língua portuguesa. O presente trabalho visa contribuir nessa área por meio da utilização de técnicas de Aprendizado de Máquina e Processamento de Linguagem Natural. O objetivo central é realizar a junção de bases de dados pertinentes da literatura e aplicar uma série de etapas de pré-processamento de texto. Além disso, há a utilização de duas técnicas de balanceamento de dados: *undersampling* e *oversampling*, bem como a extração de $N$-gramas e a utilização dos algoritmos de Aprendizado de Máquina Supervisionado conhecidos como Regressão Logística, *Support Vector Machine* (SVM) e *Naive Bayes*. Avaliações abrangentes são realizadas, incluindo a validação cruzada $k$-*fold*, utilizando métricas como acurácia e F1-*score*. Os resultados dos testes demonstram que, com base na métrica de acurácia, o melhor resultado é alcançado através do uso do modelo Regressão Logística registrando uma mediana de aproximadamente 93%. Em todas as combinações de $n$-gramas testadas, o modelo de *oversampling* exibe um desempenho superior.

**Palavras-chave**: Aprendizado de Máquina. Processamento de Linguagem Natural. Racismo. *Twitter*.

# Abstract

The remarkable growth of social networks has been accompanied by a significant increase in the spread of racist discourses within these environments. This has sparked a growing interest in studies related to this issue. However, there is a gap in research, particularly within the context of the Portuguese language. The present work aims to fill this gap through the application of Machine Learning techniques and Natural Language Processing. The main goal is to merge relevant databases from the literature and apply a series of text pre-processing steps. In addition, two data balancing techniques are used: undersampling and oversampling, as well as the extraction of $N$-grams and the use of Supervised Machine Learning algorithms known as Logistic Regression, Support Vector Machine (SVM), and Naive Bayes. Extensive evaluations are carried out, including k-fold cross-validation, using metrics such as accuracy and F1-score. The test results show that, based on the accuracy metric, the best result is achieved through the use of the Logistic Regression registering a median of approximately 93%. In all tested n-gram combinations, the oversampling model exhibits superior performance.

**Keywords**: Machine Learning. Natural Language Processing. Racism. *Twitter*.

# Lista de Ilustrações

# Lista de Tabelas

# Lista de Abreviaturas e Siglas

| | |
|---|---|
| DECOM | Departamento de Computação |
| FP | Falsos Positivos |
| FN | Falsos Negativos |
| HTML | Linguagem de marcação de hipertexto |
| PLN | Processamento de Linguagem Natural |
| SVM | *Support Vector Machine* |
| UFOP | Universidade Federal de Ouro Preto |
| VP | Verdadeiros Positivos |
| VN | Verdadeiros Negativos |

# Sumário

# 1 Introdução

Na contemporaneidade, as redes sociais desempenham um papel fundamental na comunicação, pois conectam pessoas e permitem o compartilhamento de ideias, opiniões e experiências. O X - anteriormente chamado de *Twitter*[1] -, em particular, tem se estabelecido como um ambiente amplamente utilizado para interação social. No Brasil, o *Twitter* tem o quarto maior número de usuários ativos, chegando a aproximadamente 24 milhões de pessoas em janeiro de 2023 (Statista, 2023), o que atesta a magnitude e faz valer a eficiência desta rede. No entanto, sua dinamicidade propicia também a rápida disseminação de opiniões e discursos de ódio, tornando-se um ambiente responsável por reforçar problemáticas sociais, como o racismo.

Para uma leitura analítica dos discursos racistas em um ambiente de interação social, é fundamental considerar o conceito de raça, especialmente no contexto do Brasil. Durante o período colonial, a raça foi utilizada como um dispositivo desumanizador para justificar a invasão, sequestro e escravização da população africana (GUIMARÃES, 1999). Desde então, observa-se uma evolução da propagação do racismo, transitando de uma expressão explícita para uma forma mais sutil e implícita. Dentro desse contexto, o *Twitter* desempenha um papel significativo, tanto na amplificação quanto na proliferação desse fenômeno.

A Ciência da Computação, em parceria com os movimentos sociais e indivíduos engajados, pode desempenhar um papel crucial na reparação de desigualdades historicamente construídas. Através do desenvolvimento de soluções tecnológicas, é possível abordar essas problemáticas e promover a implementação eficaz de medidas reparadoras. Nesse sentido, destacam-se os usos do Aprendizado de Máquina e do Processamento de Linguagem Natural como mecanismos essenciais para combater o racismo e contribuir para a desnaturalização das relações étnico-raciais no Brasil. Essas abordagens tecnológicas têm o potencial de identificar e confrontar vieses e discriminações, contribuindo para uma sociedade mais equitativa.

A literatura acadêmica apresenta desfalques referentes a estudos relacionados à identificação do discurso racista na linguagem portuguesa, particularmente no contexto específico do *Twitter*. Compreender e combater efetivamente o racismo nesse ambiente digital requer uma análise aprofundada e uma abordagem direcionada. Dessa forma, a presente pesquisa busca contribuir com o preenchimento dessa lacuna ao explorar e aprimorar abordagens algorítmicas para a predição de comentários racistas direcionados a pessoas negras no *Twitter*. A investigação propõe a combinação de diversos algoritmos com o objetivo de aperfeiçoar a precisão e a eficácia na detecção desses conteúdos na plataforma.

Este estudo contribui para a promoção de um ambiente *on-line* mais seguro, além de fornecer subsídios teóricos e práticos para pesquisadores, profissionais e gestores interessados

[1] https://*twitter*.com/

em compreender e combater o racismo nas redes sociais. Ao fortalecer os estudos sobre o tema, a pesquisa amplia o conhecimento científico acerca do discurso racista no *Twitter*.

## 1.1 Justificativa

A disseminação de discursos racistas nas plataformas digitais, como o *Twitter*, representa um problema social significativo. No entanto, compreender e combater efetivamente esses discursos são desafios complexos devido à imensa quantidade de informações geradas a cada minuto nas redes sociais.

Nesse contexto, surge a necessidade de preencher uma lacuna de pesquisas relacionadas à predição de discursos racistas no *Twitter*, especialmente quando se trata da língua portuguesa. Diante da escassez de estudos aprofundados nesse âmbito, este trabalho propõe-se a desenvolver um estudo direcionado. Ao concentrar-se na língua portuguesa, o estudo busca preencher uma falta significativa de investigações nessa área específica. Através da aplicação de técnicas de PLN e Aprendizado de Máquina, espera-se suprir parte dessa carência e contribuir para o avanço do conhecimento nesse campo de pesquisa.

Este trabalho não apenas beneficia os especialistas em PLN e Aprendizado de Máquina, mas também profissionais de áreas como direitos humanos, políticas públicas e comunicação, que buscam soluções inovadoras para combater a disseminação do discurso de ódio nas redes sociais. Portanto, este estudo justifica-se pela necessidade de preencher a lacuna existente, a falta de pesquisas aprofundadas na língua portuguesa, a oportunidade de desenvolver abordagens eficazes para a predição e prevenção de discursos racistas, bem como o potencial impacto social positivo em diferentes áreas profissionais.

## 1.2 Objetivos

Abaixo estão os objetivos do presente trabalho, divididos em Objetivo Principal e Objetivos Específicos:

### 1.2.1 Objetivo Principal

Com o objetivo de aprimorar a detecção de textos racistas, propõe-se o desenvolvimento de classificadores eficientes, que empregam técnicas avançadas de PLN e Aprendizado de Máquina. Essa abordagem envolve a integração de algumas bases de dados coletadas do *Twitter*, a fim de ampliar a cobertura e a representatividade dos dados utilizados para treinamento e validação do modelo.

### 1.2.2 Objetivos Específicos

1. Coletar bases de dados do *Twitter* em trabalhos já desenvolvidos, que contenham textos relacionados a diferentes contextos, como racismo e outras formas de discriminação.

2. Realizar o pré-processamento dos dados, aplicando técnicas de limpeza, normalização e tratamento de texto, a fim de obter um conjunto de dados pronto para análise e modelagem.

3. Explorar e aplicar técnicas de Aprendizado de Máquina Supervisionado, como algoritmos de classificação, para treinar modelos capazes de identificar textos racistas.

4. Avaliar e comparar a performance dos modelos desenvolvidos usando métricas apropriadas.

5. Documentar os resultados obtidos as principais descobertas sobre a identificação de textos racistas em dados do *Twitter*.

## 1.3 Organização do Trabalho

Este trabalho está estruturado da seguinte maneira:

**Capítulo 1 -** Neste capítulo, foram apresentados o contexto da aplicação, a problemática a ser abordada e solucionada, a justificativa para o desenvolvimento do trabalho e os objetivos a serem alcançados por meio da aplicação.

**Capítulo 2 -** Neste capítulo, é apresentada uma abordagem minuciosa dos conceitos-chave adotados na elaboração deste trabalho, proporcionando uma compreensão aprofundada dos fundamentos necessários para o seu desenvolvimento. Na **seção 2.5**, são realizadas análises críticas e abrangentes de trabalhos anteriores relacionados ao tema de pesquisa proposto.

**Capítulo 3 -** Neste capítulo, é apresentada uma descrição detalhada da arquitetura adotada para a realização do trabalho, juntamente com os resultados obtidos.

**Capítulo 4 -** Neste capítulo, são apresentadas as principais conclusões e reflexões acerca dos resultados do estudo. Além disso, discute-se a possibilidade de continuidade do trabalho, explorando potenciais caminhos para pesquisas futuras ou aprofundamento do tema abordado.

# 2 Fundamentação Teórica

Neste capítulo, há a contextualização do tema abordado no trabalho, por meio da apresentação de uma fundamentação teórica sólida, que engloba conceitos essenciais para a compreensão da pesquisa em questão.

## 2.1 Racismo

O racismo pode ser descrito como uma forma sistemática de discriminação baseada na raça, que se manifesta por meio de práticas conscientes ou inconscientes (ALMEIDA, 2019). Além disso, essa concepção hierarquiza grupos humanos baseado na raça, com uma escala que varia de superior a inferior. Em consonância com aquilo proposto por Almeida, (SOUZA, 2021), ao argumentar acerca da vivência da pessoa negra numa sociedade que elenca a raça branca como superior, afirma que "Ser negro é ser violentado de forma constante, contínua e cruel, sem pausa ou repouso, por uma dupla injunção: a de encarnar o corpo e os ideais de Ego do sujeito branco e a de recusar, negar e anular a presença do corpo negro."

A junção dessas perspectivas demonstra que o discurso racista, quando direcionado às pessoas negras, se utiliza de critérios sociais, culturais ou religiosos para violentar o sujeito negro nas diversas instâncias de sua vivência. Essas referências contribuem para a compreensão do racismo como um fenômeno complexo que afeta tanto as estruturas sociais quanto a experiência subjetiva das pessoas negras. O racismo, portanto, fundamenta-se na busca por justificar desigualdades e discriminações com base na raça, promovendo a supremacia de certos grupos em detrimento de outros.

No contexto brasileiro, a legislação desempenha um papel fundamental no enfrentamento do racismo, sendo um exemplo significativo a Lei 7.716/89, conhecida como Lei do Racismo. Conforme informações obtidas no ACS (2021), essa lei estabelece punições para condutas discriminatórias e preconceituosas com base em raça, cor, etnia, religião ou procedência nacional. Seu principal objetivo é coibir práticas racistas, assegurando a igualdade de direitos e a proteção das vítimas de discriminação racial.

### 2.1.1 Racismo virtual

No contexto brasileiro, o racismo virtual tem se tornado uma preocupação crescente. Dados coletados por Trindade (2020) revelam que os discursos de ódio racistas direcionados a pessoas negras têm aumentado significativamente nos últimos anos. Um mapeamento realizado por Pereira et al. (2016) no *Facebook*[1] e no *Twitter* identificou um total de 32.376 menções de

[1] https://www.facebook.com/

cunho racista, sendo que 97,6% dessas menções eram direcionadas a indivíduos negros. Isso evidencia a frequência e a intensidade dos ataques racistas nas redes sociais contra a população negra. Além disso, de acordo com Boehm (2018), em 2017 foram registrados 63.698 casos de discursos de ódio no ambiente virtual brasileiro, e aproximadamente um terço desses casos eram de cunho racista. Essa estatística reforça a relevância do racismo como uma forma predominante de discurso de ódio nas plataformas online.

Esses dados apontam para a urgência de combater o racismo virtual, que não apenas perpetua a discriminação e a desigualdade racial, mas também causa danos às vítimas, em especial aos sujeitos que se identificam como negros no contexto brasileiro. A detecção de discursos racistas no *Twitter*, por meio de abordagens baseadas em Aprendizado de Máquina e PLN, torna-se fundamental para monitorar e combater esses comportamentos.

## 2.2 Processamento de Linguagem Natural

O PLN é uma área de pesquisa que tem como foco a interação entre computadores e a linguagem humana. Seu objetivo é capacitar os sistemas computacionais a compreender, interpretar, manipular e gerar linguagem humana de maneira natural e eficiente. Nesse sentido, o desenvolvimento de algoritmos e técnicas é fundamental para viabilizar extrações de significados dos textos, respostas a perguntas, tradução de idiomas, resumo de documentos, análise de sentimentos em redes sociais, entre outros. Essas capacidades do PLN têm aplicação prática em diversas áreas, proporcionando benefícios tanto para o avanço da pesquisa científica quanto para a solução de problemas reais.

O PLN engloba uma série de etapas, como normalização, remoção de *stop words*, tokenização, definição de $n$-gramas, entre outras. Nas próximas seções, cada uma dessas técnicas será detalhadamente descrita, proporcionando uma compreensão mais profunda sobre o assunto.

### 2.2.1 Normalização

De acordo com Dias (2021), esse é o processo utilizado para garantir a padronização mínima dos textos, de modo que todos possuam o mesmo formato e sejam tratados de forma consistente e sem ambiguidades. Isso envolve converter todos os textos para letras minúsculas, pois o uso de letras maiúsculas e minúsculas pode resultar em diferentes representações da mesma palavra. Além disso, são removidos URLs, símbolos, números, entre outros elementos irrelevantes para a análise. A remoção de símbolos e números permite que o foco se concentre nas palavras-chave e no significado do texto. Em algumas situações, símbolos, pontuação e números podem adicionar ruído ao texto, especialmente em dados textuais não estruturados, como redes sociais. A remoção desses elementos pode melhorar a qualidade dos dados e a eficácia das análises.

### 2.2.2 Remoção de *Stop words*

Segundo a literatura acadêmica, é amplamente reconhecido que palavras como artigos, conjunções e preposições são comumente denominadas *stop words* ou palavras de parada. Essas palavras, em geral, não possuem relevância significativa para o processamento do texto nem para os mecanismos de aprendizagem. Com o intuito de mitigar o impacto dessas palavras, emprega-se a técnica conhecida como remoção de *stop words*. Essa técnica tem como finalidade primordial a redução da quantidade de termos a serem processados, bem como a eliminação de termos considerados desnecessários durante o processo de análise textual Ferreira (2019).

A Figura 2.1 explicita o que Dias (2021) subentende como *stop words* na língua portuguesa do Brasil.

| StopWords |
| --- |
| a, à, ao, aos, aquela, aquelas, aquele, aqueles, aquilo, as, às, até, com, como, da, das, de, dela, delas, dele, deles, depois, do, dos, e, é, ela, elas, ele, eles, em, entre, era, eram, éramos, essa, essas, esse, esses, esta, está, estamos, estão, estar, estas, estava, estavam, estávamos, este, esteja, estejam, estejamos, estes, esteve, estive, estivemos, estiver, estivera, estiveram, estivéramos, estiverem, estivermos, estivesse, estivessem, estivéssemos, estou, eu, foi, fomos, for, fora, foram, fôramos, forem, formos, fosse, fossem, fôssemos, fui, há, haja, hajam, hajamos, hão, havemos, haver, hei, houve, houvemos, houver, houvera, houverá, houveram, houvéramos, houverão, houverei, houverem, houveremos, houveria, houveriam, houveríamos, houvermos, houvesse, houvessem, houvéssemos, isso, isto, já, lhe, lhes, mais, mas, me, mesmo, meu, meus, minha, minhas, muito, na, não, nas, nem, no, nos, nós, nossa, nossas, nosso, nossos, num, numa, o, os, ou, para, pela, pelas, pelo, pelos, por, qual, quando, que, quem, são, se, seja, sejam, sejamos, sem, ser, será, serão, serei, seremos, seria, seriam, seríamos, seu, seus, só, somos, sou, sua, suas, também, te, tem, tém, temos, tenha, tenham, tenhamos, tenho, terá, terão, terei, teremos, teria, teriam, teríamos, teu, teus, teve, tinha, tinham, tínhamos, tive, tivemos, tiver, tivera, tiveram, tivéramos, tiverem, tivermos, tivesse, tivessem, tivéssemos, tu, tua, tuas, um, uma, você, vocês, vos |

Figura 2.1 – *Stop words* em Português do Brasil

### 2.2.3 Tokenização

De acordo com Anchiêta et al. (2021), a tokenização é o processo de dividir um texto em *tokens*, que podem ser palavras, números, símbolos ou outros elementos. Alguns tokenizadores usam espaços, tabulações e quebras de linhas como separadores, enquanto outros também consideram os sinais de pontuação como separadores. Essa abordagem varia dependendo do tokenizador utilizado.

### 2.2.4 $N$-gramas

$N$-gramas são estruturas de dados amplamente utilizadas em PLN e mineração de textos. Essas estruturas representam sequências de $n$ palavras, letras, sílabas ou fonemas, dependendo da abordagem adotada, com o objetivo de analisar textos e extrair informações linguísticas relevantes (BRODER et al., 1997). A abordagem estatística dos $n$-gramas permite a análise da frequência de palavras e a identificação de padrões linguísticos específicos presentes nos textos. Ao examinar cada par, tripla ou conjunto de palavras que ocorrem juntas, é possível obter dados valiosos sobre

a co-ocorrência de termos dentro do corpus em análise. Isso permite uma compreensão mais profunda das relações entre as palavras e a estrutura geral do texto.

## 2.3 Aprendizado de Máquina

Segundo Paiva, Silva e Moura (2020), Aprendizado de Máquina é uma área que se desenvolve no contexto científico, focada na construção de algoritmos com a capacidade de aprender a partir de experiências, resultando em um aprimoramento progressivo do desempenho por meio da interação com novas informações. Nesse sentido, são desenvolvidos e utilizados modelos com base em dados, visando aprender e otimizar seu desempenho. Esses modelos são treinados utilizando algoritmos e técnicas específicas, com o intuito de extrair padrões e informações dos dados, possibilitando a realização de previsões e tomada de decisões embasadas nesse aprendizado. E, de acordo com Haykin (2009), dentro desse campo científico, é possível identificar três categorias fundamentais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado semi-supervisionado.

Os estudos conduzidos por Carvalho (2021) e a pesquisa realizada por Paiva e Escovedo (2021) abordam as modalidades de aprendizado supervisionado e aprendizado não supervisionado. No primeiro, o sistema possui conhecimento prévio do ambiente e utiliza conjuntos de dados previamente rotulados para aprender os padrões existentes. Dessa forma, quando são apresentados novos dados não rotulados, o algoritmo utiliza o conhecimento adquirido para realizar a classificação. Já no aprendizado não supervisionado, o sistema aprende através da criação de representações internas com base nas características das entradas, explorando os dados em busca de padrões e relacionamentos, sem depender de informações prévias sobre os resultados esperados durante o treinamento. No presente trabalho, avaliaremos algoritmos supervisionados na classificação de textos racistas.

### 2.3.1 Técnicas de Aprendizado de Máquina Supervisionado

Nesta subseção, serão apresentados os algoritmos empregados na elaboração deste trabalho, os quais serão utilizados para atingir os objetivos específicos.

#### 2.3.1.1 Regressão Logística

Regressão Logística é uma técnica que utiliza fundamentos matemáticos para estimar probabilidades ao empregar a função logística (HARRISON, 2019). Além disso, é importante destacar que, conforme elucidado por Gonzalez (2018), Regressão Logística é uma abordagem técnica empregada na modelagem de previsões.

Nesse contexto, um modelo é construído para a variável dependente com base em variáveis independentes. Uma característica central desta técnica reside na sua aplicabilidade a cenários em que a variável dependente apresenta valores categóricos ou binários, já que é um

algoritmo utilizado para prever o valor de um desses fatores com base no outro. Essas previsões frequentemente se manifestam em um conjunto limitado de resultados, tais como "sim"ou "não".

De acordo com Minussi, Damacena e Jr (2002), o modelo de Regressão Logística pode ser descrito conforme a Equação 2.1:

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}} \tag{2.1}$$

Onde $g(x)$ diz respeito à Equação 2.2:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{2.2}$$

Os valores referentes ao segundo membro da Equação 2.2 são calculados por meio do conjunto de dados empregando o método da máxima verossimilhança. Este método tem como objetivo encontrar uma combinação de coeficientes que otimize a probabilidade de observação da amostra (HOSMER; LEMESHOW, 1989).

Com base nas discussões de Géron (2022) e Minussi, Damacena e Jr (2002) sobre a classificação de dados, torna-se evidente que a discriminação acontece em dois grupos de acordo com as especificações descritas a seguir:

- Caso $P(Y = 1) > 0.5$, então $Y = 1$;
- Em divergência ao caso anterior, classifica-se $Y = 0$.

Por fim, é importante ressaltar que esse método encontra aplicações diversificadas, possibilitando a modelagem de relações complexas entre variáveis e as probabilidades associadas à inclusão em uma classe específica.

#### 2.3.1.2 SVM

O modelo de classificação SVM busca criar um hiperplano para distinguir entre duas classes. O hiperplano é um limite de decisão que permite a previsão de rótulos com base em um ou mais vetores de características.

A particularidade do hiperplano no SVM é que ele é orientado de maneira a maximizar a distância até os pontos de dados mais próximos de cada classe. Estes pontos, que estão mais próximos ao hiperplano e, portanto, têm maior influência na sua orientação e posição, são conhecidos como vetores de suporte (HUANG et al., 2018).

Considerando um conjunto de treinamento rotulado: $(x_1, y_1), ..., (x_n, y_n)$, onde $x_i \in R^d$ e $y_i \in \{-1, +1\}$, cada $x_i$ representa um vetor de características e cada $y_i$ representa o rótulo da

classe de um composto de treinamento $i$. Neste contexto, o hiperplano pode ser definido pela seguinte Equação 2.3:

$$wx^T + b = 0 \tag{2.3}$$

Nesta equação, $w$ é o vetor de peso, $x$ é o vetor de características de entrada, e $b$ é o viés. O vetor de peso $w$ e o viés $b$ devem satisfazer as seguintes desigualdades para todos os elementos do conjunto de treinamento:

$$wx_i^T + b \geq +1, \quad \text{se} \quad y_i = 1 \quad ou \quad wx_i^T + b \leq -1, \quad \text{se} \quad y_i = -1 \tag{2.4}$$

O principal objetivo de treinar um modelo SVM é encontrar os valores de $w$ e $b$ que não apenas separam os dados corretamente, mas também maximizam a distância entre o hiperplano e os vetores de suporte mais próximos (HUANG et al., 2018). A maximização desta distância é fundamental para garantir a robustez do modelo.

#### 2.3.1.3 *Naive Bayes*

O modelo *Naive Bayes* é baseado no teorema de *Bayes* e pressupõe a independência condicional entre os recursos ou atributos dos dados (ZHANG, 2004). Essa suposição de independência facilita o cálculo das probabilidades condicionais para cada classe, tornando o modelo eficiente em termos computacionais.

Quando se utiliza o classificador *Naive Bayes* para fazer previsões em novas instâncias, o resultado é aproximado por $P(C = y_i|X)$, isto é, a probabilidade dessa instância $X = x_1, x_2, ..., x_k$ pertencer à classe $y_i$.

A probabilidade de uma instância pertencer à classe $y_i$ dado o conjunto de recursos $X$ é dada pela Equação 2.5:

$$P(C = y_i|X) = \frac{P(C = y_i) \cdot P(X|y_i)}{P(X)} \tag{2.5}$$

Onde:

- $P(C = y_i)$ é a probabilidade a priori da classe $y_i$, ou seja, a probabilidade de encontrar uma instância pertencente a essa classe no conjunto de dados de treinamento;
- $P(X|y_i)$ é a probabilidade condicional dos recursos $X$ dada a classe $y_i$, que representa a probabilidade de observar o conjunto de recursos $X$ nas instâncias da classe $y_i$ no conjunto de treinamento;
- $P(X)$ é a probabilidade marginal do conjunto de recursos $X$, que é utilizada para normalizar a probabilidade e garantir que a soma de todas as probabilidades condicionais seja igual a 1.

O algoritmo usa essas probabilidades para determinar a probabilidade de uma instância pertencer a cada classe possível e, assim, realiza a classificação com base na classe que apresenta a maior probabilidade. Uma das vantagens do Naive Bayes é que ele permite o treinamento em uma única passagem pelo conjunto de dados.

## 2.4 Métricas de Avaliação

No contexto de um modelo de Aprendizado de Máquina, a avaliação do desempenho é um passo crucial após a sua construção. Para esse propósito, é indispensável utilizar métricas de avaliação capazes de mensurar a eficácia do modelo desenvolvido. Cada uma das métricas empregadas nesse processo proporciona informações sobre o desempenho do modelo. A seleção adequada das métricas é de suma importância para realizar uma avaliação precisa e obter uma compreensão abrangente do desempenho do classificador.

### 2.4.1 Acurácia

Acurácia é uma métrica utilizada para avaliar o desempenho de um modelo de classificação e é definida como a razão entre o número de predições corretas e o total de exemplos no conjunto de teste (HARRISON, 2019). A Equação 2.6 apresenta a fórmula da acurácia.

$$A = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.6}$$

Onde:

- VP corresponde a Verdadeiro Positivo;
- VN é Verdadeiro Negativo;
- FP significa Falso Positivo;
- FN representa Falso Negativo.

### 2.4.2 F1-*score*

A medida F1-*score* é uma métrica que leva em consideração a Precisão (P) e a Revocação (R). A Precisão é calculada como a proporção de Verdadeiros Positivos (VP) para a soma de Verdadeiros Positivos e Falsos Positivos (FP), conforme a Equação 2.7. Já a Revocação é a proporção de Verdadeiros Positivos para a soma de Verdadeiros Positivos e Falsos Negativos (FN), conforme a Equação 2.8. O F1-*score* realiza seu cálculo utilizando a média harmônica entre a Precisão e a Revocação Harrison (2019), como descreve a Equação 2.9.

$$P = \frac{VP}{VP + FP} \tag{2.7}$$

$$R = \frac{VP}{VP + FN} \tag{2.8}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{2.9}$$

### 2.4.3 Validação Cruzada $K$-*fold*

Conforme delineado por Castro e Ferrari (2017), a técnica conhecida como validação cruzada $k$-*fold* desempenha um papel crucial na avaliação de modelos. Essa abordagem se baseia na divisão do conjunto de dados em $k$ subconjuntos distintos. Em cada iteração, $k - 1$ subconjuntos são empregados para o treinamento do modelo, enquanto o subconjunto restante é destinado ao teste. Esse ciclo de treinamento e teste é repetido $k$ vezes, abrangendo cada um dos $k$ subconjuntos em diferentes momentos. Ao concluir o processo, a média dos desempenhos obtidos nas bases de treinamento e teste é calculada. Esse valor médio é então adotado como um indicador representativo da qualidade do modelo.

## 2.5 Trabalhos Relacionados

Alguns estudos têm sido realizados com o objetivo de identificar e classificar discursos de ódio e conteúdos racistas em plataformas online, especialmente no Twitter. Cada um desses estudos emprega diferentes abordagens, técnicas e algoritmos para atingir seus objetivos.

No estudo conduzido por Pelle e Moreira (2017), foram criados dois conjuntos de dados a partir de comentários coletados no site de notícias G1[1]. Esses comentários foram extraídos das seções de política e esportes por meio de um *web scraper* utilizado para enviar solicitações ao site e baixar as páginas HTML das notícias, a partir das quais os dados foram extraídos. Os comentários coletados foram classificados em ofensivos e não ofensivos. Anotadores especializados foram empregados para identificar os comentários ofensivos, que incluem, mas não se limitam a, racismo, sexismo, homofobia e xenofobia. Algoritmos de aprendizado de máquina, como *Naive Bayes*, *Support Vector Machine* e *Random Forest*, foram utilizados para classificar os dados. Para validação dos resultados, foram realizadas comparações com estudos anteriores que se concentraram na identificação de discursos de ódio em inglês, obtendo desempenhos semelhantes a esses estudos.

Em um estudo similar, Nascimento (2019) comparou técnicas de Aprendizado de Máquina supervisionado e abordagens de pré-processamento para classificar *tweets* contendo discurso de

[1] https://g1.globo.com/

ódio. Utilizando dois conjuntos de dados, um em inglês e outro em português, o estudo fez uso de três modelos de classificação: SVM, *Naive Bayes* e Regressão Logística. Os resultados indicaram que a combinação de *stemming*, vetorização TF-IDF e o classificador SVM foi a mais eficaz para ambos os conjuntos de dados. Contudo, para o conjunto desbalanceado em português, o *Naive Bayes* mostrou-se mais adequado. O estudo é uma contribuição relevante ao apresentar uma abordagem eficaz para identificação automática de discursos de ódio no *Twitter*, com destaque para a criação de um conjunto de dados em português.

Em paralelo, o estudo de Castro (2019) compara abordagens de pré-processamento e classificação na identificação de discursos de ódio em português. A pesquisa avalia estratégias de pré-processamento e técnicas de classificação aplicadas aos classificadores *Naive Bayes* e SVM. Os resultados indicam que a combinação de seleção de atributos com *undersampling* apresenta o melhor desempenho para a base de dados 'OffComBR2', enquanto a extração de radicais mostra resultados promissores para o conjunto 'OffComBR3'. Ambos os estudos contribuem significativamente para a detecção de discursos de ódio em plataformas públicas.

Já Fortuna et al. (2019) buscaram enfrentar a escassez de dados rotulados em português relacionados ao discurso de ódio. Para isso, criaram um conjunto de dados rotulados hierarquicamente em 81 categorias diferentes, coletados via API do *Twitter*. Os resultados obtidos demonstram que este conjunto é eficiente para treinar modelos de Aprendizado de Máquina na detecção de discurso de ódio em português brasileiro.

Reis (2021), por sua vez, propuseram o desenvolvimento de um modelo de Aprendizado de Máquina para identificar comentários racistas no Twitter, avaliando a eficiência de algoritmos como *Naive Bayes*, SVM e Regressão Logística. A metodologia empregada focou na categorização orientada à categoria, com análise adicional para verificar o contexto das postagens. Apesar dos desafios identificados, como a necessidade de grande volume de dados e a complexidade da linguagem nas redes sociais, a Regressão Logística apresentou a melhor taxa de acurácia.

Por fim, Silva, Fernandes e Fernandes (2018) focaram na detecção e classificação de mensagens racistas em português no *Twitter*. Os autores coletaram *tweets* contendo palavras potencialmente racistas e aplicaram técnicas de pré-processamento para reduzir o ruído nas mensagens. Em seguida, as mensagens foram rotuladas por anotadores humanos quanto à presença ou ausência de características racistas. A Regressão Logística e o *Naive Bayes* foram os algoritmos utilizados para a classificação, sendo que a Regressão Logística apresentou um desempenho superior.

Ao analisar os estudos mencionados, é possível identificar uma série de semelhanças e diferenças em suas abordagens para a identificação e classificação de discursos de ódio ou conteúdos racistas no *Twitter*. Todos os estudos empregam técnicas de aprendizado de máquina e pré-processamento de dados, destacando a importância dessas estratégias na análise de conteúdo online. No entanto, a escolha dos algoritmos específicos e as abordagens para coleta e rotulagem de dados variam significativamente.

Em termos de desempenho, há variações notáveis. Por exemplo, Silva, Fernandes e Fernandes (2018) e Reis (2021) encontraram que a Regressão Logística foi mais eficaz na detecção de características racistas, enquanto Nascimento (2019) descobriram que o *Naive Bayes* era mais adequado para um conjunto de dados desbalanceado em português. Isso sugere que a eficácia dos algoritmos pode depender do contexto específico e do conjunto de dados utilizado. Essas diferenças nos métodos e resultados destacam a complexidade da tarefa de identificar e classificar discursos de ódio e conteúdo racista *online*. Eles também apontam para a necessidade de abordagens inovadoras e adaptativas, capazes de lidar com diferentes contextos e conjuntos de dados, como a proposta neste trabalho.

A Tabela 2.1 é responsável por fornecer uma comparação dos estudos abordados nesta seção.

Tabela 2.1 – Comparação dos estudos mencionados

| **Autor(es)** | **Algoritmos Utilizados** | **Conjunto de Dados** | **Melhor Desempenho** |
|---|---|---|---|
| Pelle e Moreira (2017) | *Naive Bayes*, SVM, *Random Forest* | Dados extraídos do portal G1 | Não especificado |
| Nascimento (2019) | SVM, *Naive Bayes*, Regressão Logística | Dados extraídos do *Twitter* em inglês e português | SVM (conjunto balanceado), *Naive Bayes* (conjunto desbalanceado) |
| Castro (2019) | *Naive Bayes*, SVM | Dados extraídos da WEB | Não especificado |
| (FORTUNA et al., 2019) | LSTM | Dados extraídos do *Twitter* | Não se aplica |
| Reis (2021) | *Naive Bayes*, SVM, Regressão Logística | Dados extraídos do *Twitter* | Regressão Logística |
| Silva, Fernandes e Fernandes (2018) | Regressão Logística, *Naive Bayes* | Dados extraídos do *Twitter* | Regressão Logística |

Fonte: Produzida pelo próprio autor.

Assim como trabalhos anteriores, este estudo emprega algoritmos de aprendizado de máquina supervisionado, especificamente os mais utilizados na literatura - Regressão Logística, SVM e *Naive Bayes*. Além disso, este trabalho contribui para o avanço do campo ao treinar e testar os algoritmos com uma quantidade de dados significativamente maior do que a normalmente utilizada na literatura, especialmente em estudos que consideram a rede social *Twitter*.

# 3 Desenvolvimento

Neste capítulo, serão apresentadas as ferramentas utilizadas no desenvolvimento deste trabalho, bem como serão realizadas discussões sobre os elementos da arquitetura adotada. Além disso, serão apresentados e discutidos os resultados obtidos nesta pesquisa.

## 3.1 Ferramentas de Desenvolvimento

O código desenvolvido neste estudo foi executado no ambiente Colab[1] utilizando a linguagem de programação Python, fazendo uso das principais bibliotecas para análise e processamento de dados a exemplo das bibliotecas pandas, nltk, scikit-*learn*, matplotlib, entre outras. A biblioteca pandas foi empregada para manipulação e análise eficiente de dados tabulares. A biblioteca nltk desempenhou um papel fundamental no processamento de linguagem natural, fornecendo recursos essenciais, como tokenização e remoção de palavras irrelevantes. O scikit-*learn*, uma biblioteca amplamente utilizada para aprendizado de máquina, ofereceu uma variedade de algoritmos e métricas para tarefas como classificação e regressão. Por fim, a biblioteca Matplotlib foi utilizada para a criação de visualizações gráficas, facilitando a interpretação e comunicação dos resultados obtidos. O uso dessas bibliotecas consolidadas na comunidade científica proporcionou recursos eficientes e robustos para o desenvolvimento e análise dos dados.

## 3.2 Método

O método empregado nesta pesquisa é constituído pela arquitetura ilustrada na Figura 3.1. Cada uma das etapas representadas no diagrama será detalhada nas subseções subsequentes.


Figura 3.1 – Arquitetura do método.

Fonte: Produzida pelo próprio autor.

### 3.2.1 Base de Dados

Com o objetivo de melhorar a qualidade dos resultados da pesquisa, optou-se por combinar quatro bases de dados distintas, previamente utilizadas em outros estudos. A união desses dados resultaram em um total de 7.653 *tweets* em português do Brasil extraídos do *Twitter* em diferentes datas, sendo que apenas 629 foram categorizados como racistas.

Uma das bases fundamentais (FORTUNA et al., 2019) contém 5.668 registros em português do Brasil e português de Portugal. Esses *tweets* foram coletados e classificados por anotadores especializados, seguindo um esquema hierárquico de rótulos múltiplos, abrangendo um total de 81 categorias de discurso de ódio. Dentre eles, 94 foram classificados como racistas e 5.574 como não racistas, sendo todos os dados utilizados na construção desta pesquisa. Ademais, foram incorporados à base de dados 138 *tweets* classificados como racistas, selecionados a partir do trabalho de Leite et al. (2020). Esses dados foram categorizados como racistas por um, dois ou três anotadores. É importante destacar que, nessa base de dados, foram identificados seis tipos diferentes de preconceitos, dos quais foram mantidos apenas os dados racistas. Além disso, a base construída por Augusto (2021), disponibilizada no GitHub, contribuiu com 97 *tweets* classificados como racistas e 1.150 *tweets* classificados como não racistas. Por fim, o trabalho de Neto et al. (2017) adicionou 300 *tweets* classificados como racistas e 300 *tweets* classificados como não racistas. A rotulação dos *tweets* foi realizada por dois analistas, e, em caso de divergência, um terceiro foi consultado para chegar a um consenso.

Para fins de elucidação, a Figura 3.3 realiza a ilustração da quantidade de dados coletada de cada autor.

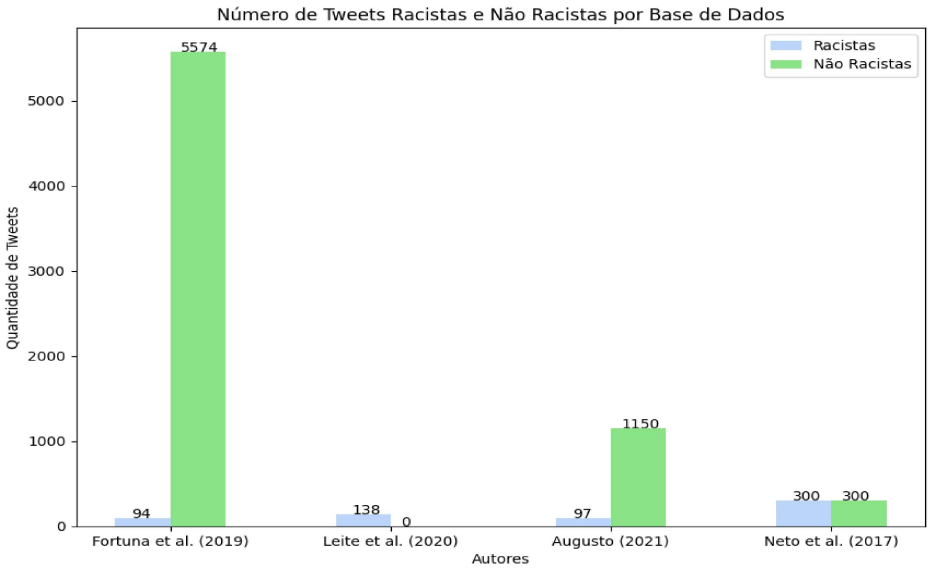


Figura 3.2 – Número de *Tweets* Racistas e Não Racistas por Base de Dados

Fonte: Produzida pelo próprio autor.

[1] https://colab.research.google.com/

A combinação dessas bases de dados proporcionou uma coleção diversificada de *tweets*, abrangendo diferentes contextos e abordagens de anotação. Essa diversidade é um aspecto relevante para obtenção de resultados mais robustos e abrangentes para esta pesquisa.

### 3.2.2 Processamento de Linguagem Natural

Nesta etapa do estudo, os *tweets* presentes na base de dados foram submetidos a um conjunto de procedimentos de PLN. Inicialmente, informações irrelevantes para a análise, como usuários mencionados, sequências de *retweets*, símbolos, números, pontuações e links, foram removidas de cada dado. Em seguida, os textos foram normalizados, convertendo todas as letras para minúsculas. Adicionalmente, partes do texto que não possuíam valor informativo relevante para a análise foram removidas, incluindo *stop words*, que são palavras comuns na língua que geralmente não contribuem para a compreensão do conteúdo. Posteriormente, os dados foram submetidos à tokenização, ou seja, cada frase foi alocada em um vetor, onde cada posição do vetor representa uma palavra presente na frase. Por fim, foi realizada a contagem da frequência de palavras em diferentes sequências, incluindo unigramas, bigramas, trigramas e quadrigramas.

Essas etapas de PLN foram essenciais para preparar os dados textuais e possibilitaram análises posteriores robustas e significativas. Os detalhes teóricos e conceituais dessas técnicas estão abordados no **Capítulo 2**. A Tabela 3.1 ilustra as etapas de processamento de texto descritas acima.

Tabela 3.1 – Etapas de Processamento de Texto

| **Entrada** | **Saída** |
|---|---|
| Tweet original | Ontem a @user1 chegou na escola achando que estava arrasando. Filha, seu cabelo é ruim! #bombril |
| Tweet sem elementos inúteis | Ontem a chegou na escola achando que estava arrasando Filha seu cabelo é ruim |
| Tweet normalizado | ontem a chegou na escola achando que estava arrasando filha seu cabelo é ruim |
| Tweet sem *stop words* | ontem chegou escola achando estava arrasando filha seu cabelo ruim |
| Tokens | ['ontem', 'chegou', 'escola', 'achando', 'estava', 'arrasando', 'filha', 'seu', 'cabelo', 'ruim'] |
| Unigrama | ['ontem', 'chegou', 'escola', 'achando', 'estava', 'arrasando', 'filha', 'seu', 'cabelo', 'ruim'] |
| Bigramas | ['ontem chegou', 'chegou escola', 'escola achando', 'achando estava', 'estava arrasando', 'arrasando filha', 'filha seu', 'seu cabelo', 'cabelo ruim'] |
| Trigramas | ['ontem chegou escola', 'chegou escola achando', 'escola achando estava', 'achando estava arrasando', 'estava arrasando filha', 'arrasando filha seu', 'filha seu cabelo', 'seu cabelo ruim'] |
| Quadrigramas | ['ontem chegou escola achando', 'chegou escola achando estava', 'escola achando estava arrasando', 'achando estava arrasando filha', 'estava arrasando filha seu', 'arrasando filha seu cabelo', 'filha seu cabelo ruim'] |

Fonte: Produzida pelo próprio autor.

### 3.2.3 Divisão dos Conjuntos de Dados

Neste estudo, foi observada uma disparidade na quantidade de dados racistas e não racistas disponíveis. Com o intuito de abordar essa questão, a base de dados foi dividida em duas bases de dados distintas: uma para a classe majoritária e outra para a classe minoritária. Cada base de dados contém exemplos pertencentes às suas respectivas classes.

Para abordar o desequilíbrio entre as classes, duas abordagens foram empregadas neste trabalho: *oversampling* e *undersampling*. As técnicas de *oversampling* e *undersampling* são abordagens para lidar com o desbalanceamento de classes. Conforme Silva (2022), a técnica de *oversampling* envolve a geração de amostras sintéticas para a classe minoritária, como demonstra os exemplos da Tabela 3.2, aumentando seu tamanho até que fique igual ao tamanho da classe majoritária. Por outro lado, a técnica de *undersampling* consiste na redução da classe majoritária, de forma que o número de amostras seja equivalente ao número de amostras da classe minoritária. Ambas têm como objetivo realizar o balanceamento dos dados, isto é, torná-las igualitárias, proporcionando uma distribuição mais equilibrada entre as classes. O resultado dessas operações é uma nova base de dados, que reúne todos os dados das duas classes balanceadas para serem

utilizados nas etapas de treinamento e teste do modelo.

Tabela 3.2 – Exemplos de Amostras Sintéticas

| |
|---|
| Já dizia o ditado, negro é bom morto |
| A diferença entre negro e câncer é que câncer evolui |
| Seu cabelo não é feio porque você penteia |

Fonte: Produzida pelo próprio autor.

### 3.2.4 Treinamento dos Modelos

Foram criados três modelos que representam os algoritmos Regressão Logística, SVM e *Naive Bayes*, adequados para problemas de classificação binária.

Para o modelo de Regressão Logística, foi utilizado o algoritmo *LogisticRegression* com a classe balanceada e um número máximo de iterações definido como 1000. A otimização dos hiperparâmetros foi realizada através de uma busca em grade com validação cruzada de 10 *folds*. Para o treinamento do modelo *Naive Bayes*, foi utilizado o algoritmo *BernoulliNB* com um conjunto de possíveis valores para o parâmetro de regularização *alpha* definido como [0.001, 0.01, 0.1, 1, 10, 100]. O melhor valor para *alpha* foi determinado através de uma busca em grade com validação cruzada de 10 *folds*. No caso do modelo SVM, foi utilizado o algoritmo *SGDClassifier* com o parâmetro de regularização *alpha* otimizado através de uma validação cruzada de busca aleatória. Os possíveis valores para *alpha* foram [ 0.001, 0.01, 0.1, 1, 10, 100].

O treinamento dos modelos foi realizado ao receber os dados de treinamento e seus rótulos. No processo, houve o ajuste dos pesos com base nos dados de treinamento, buscando sempre encontrar a melhor relação entre as características de cada *tweet* e suas classes correspondentes. Os conjuntos de treinamento e teste correspondem, respectivamente, a 70% e 30%.

### 3.2.5 Avaliação dos Modelos

Após realizar a predição dos modelos, foram conduzidas validações cruzadas $k$-*fold* de tamanho 10 para avaliação. Nesse processo, a acurácia foi calculada ao empregar os dados de teste em conjunto com as previsões do modelo, medindo a proporção de acertos obtidos. Adicionalmente, a avaliação abrangente do desempenho do modelo incluiu o cálculo do F1-*score*. A análise do desempenho dos modelos foi conduzida considerando unigramas (1,1), bigramas (2,2), trigramas (3,3) e quadrigramas (4,4), e englobou a aplicação de técnicas de *oversampling* ou *undersampling* para um panorama mais completo da performance.

## 3.3 Resultados

As subseção 3.3.1, subseção 3.3.2 e subseção 3.3.3 exibem os resultados finais da pesquisa, os quais foram obtidos ao utilizar a base de dados construída conforme descrito na subseção 3.2.1.

Para alcançar esses resultados, foram seguidas as etapas de PLN detalhadas na subseção 3.2.2. Além disso, a divisão dos conjuntos de dados, conforme mencionado na subseção 3.2.3, foi empregada para avaliar as métricas de acurácia e F1-*score* do modelo treinado, conforme abordado na subseção 3.2.4.

### 3.3.1 Regressão Logística

A Figura 3.3 ilustra os resultados obtidos utilizando o modelo de classificação Regressão Logística.

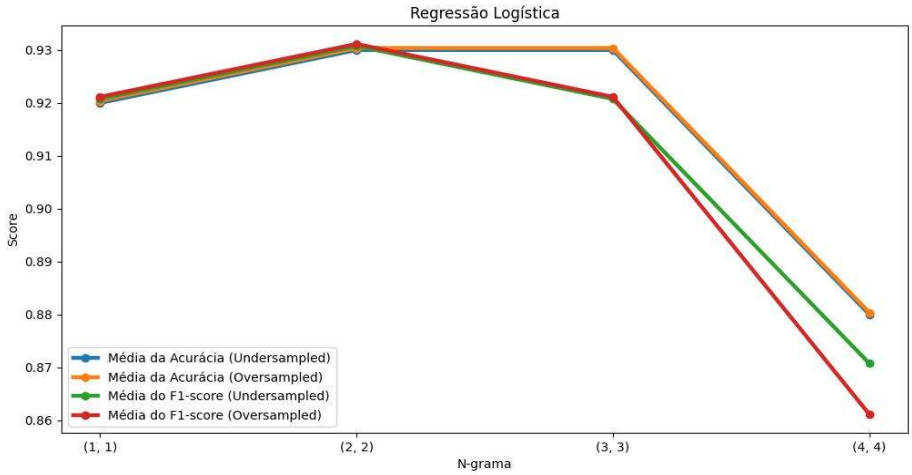


Figura 3.3 – Regressão Logística

Fonte: Produzida pelo próprio autor.

Os resultados apresentados ilustram o desempenho do algoritmo com diferentes configurações de $N$-gramas e técnicas de amostragem.

Ao realizar uma análise detalhada, observa-se que a acurácia média para $N$-gramas (1,1) e (2,2) foi de 92% e 93%, respectivamente, quando aplicado o *undersampling*. A acurácia se manteve em 93% para $N$-grama (3,3), indicando que aumentar o tamanho do $N$-gramas não contribuiu para melhorar a acurácia do modelo. No entanto, ao analisar o $N$-grama (4,4), percebe-se uma queda na acurácia para 88%. Isso sugere que um $N$-grama maior pode estar comprometendo a eficácia do modelo.

Quando utilizada a técnica de balanceamento *oversampling*, os resultados aproximados foram idênticos aos obtidos com o *undersampling*. Isso sugere que a técnica de amostragem escolhida não tem um impacto significativo na acurácia do modelo neste caso.

Analisando o F1-*score* no caso do *oversampling*, foram obtidas as médias de 92%, 93% e 92% para $N$-gramas (1,1), (2,2) e (3,3), respectivamente. Estes resultados aproximados são

idênticos aos obtidos com o *undersampling*. Contudo, ao avaliar o $N$-grama (4,4), o F1-*score* diminuiu para 86%, sendo ligeiramente inferior ao obtido com o *undersampling*, 87%.

Com base nesses resultados, podemos concluir que, para este modelo, a combinação mais eficaz é o uso de um $N$-grama (2,2) com qualquer técnica de amostragem. Esta combinação forneceu a maior acurácia média, 93%, e o maior F1-*score* médio, 93%. No entanto, é importante ressaltar que os resultados foram muito próximos entre si. Portanto, a escolha da melhor combinação pode depender de outros fatores que não foram considerados nesta análise.

### 3.3.2 SVM

A Figura 3.4 ilustra os resultados obtidos utilizando o modelo de classificação SVM.

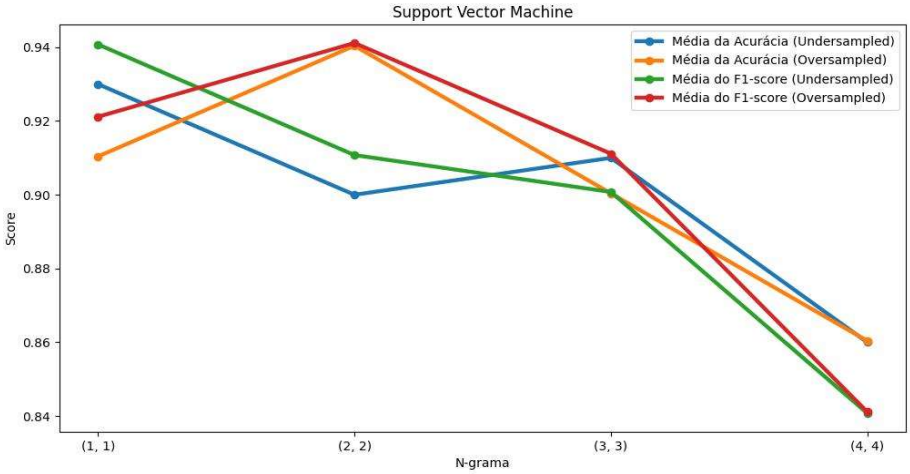


Figura 3.4 – SVM

Fonte: Produzida pelo próprio autor.

O gráfico apresentado na Figura 3.4 fornece uma análise comparativa do desempenho do algoritmo SVM com diferentes configurações de $N$-gramas e duas técnicas distintas de balanceamento de dados: *undersampling* e *oversampling*.

Para a configuração de N-grama (1, 1), o modelo alcançou uma acurácia média de 93% com *undersampling* e 91% com *oversampling*. O F1-*score* médio foi de 94% e 92%, respectivamente. Isso sugere que a técnica de *undersampling* pode ter permitido ao modelo evitar *overfitting*, resultando em um desempenho ligeiramente superior. Na configuração (2, 2), observou-se uma diminuição na acurácia e no O F1-*score* com *undersampling*, atingindo 90% em ambos os casos. No entanto, com *oversampling*, ambas as métricas aumentaram para 94%. Isso pode indicar que a técnica de *oversampling*, ao criar cópias sintéticas das classes minoritárias, pode ter ajudado o modelo a capturar melhor a estrutura dos dados, especialmente quando se trata de blocos de palavras maiores.

Ao aumentar o tamanho do $N$-grama para (3, 3), a acurácia com *undersampling* foi de 91% e o F1-*score* de 90%. Com *oversampling*, ambas as métricas caíram para 90%. Este resultado sugere que o aumento do tamanho do $N$-grama pode ter tornado o modelo mais complexo e menos capaz de generalizar a partir dos dados de treinamento. Finalmente, para a configuração (4, 4), tanto a acurácia quanto o F1-*score* foram de 86% com *undersampling* e se mantiveram estáveis com *oversampling*. Isso indica que, para esta configuração, ambas as técnicas de balanceamento tiveram desempenhos semelhantes, talvez porque a complexidade adicional introduzida pelo aumento do tamanho do $N$-grama tenha limitado a eficácia do balanceamento.

Em suma, esses resultados sugerem que o modelo SVM teve um desempenho melhor com $N$-gramas menores e ao utilizar a técnica de *undersampling*. No entanto, é importante lembrar que esses resultados são específicos para este conjunto de dados e podem não se aplicar a outros contextos.

### 3.3.3 *Naive Bayes*

A Figura 3.5 ilustra o desempenho do algoritmo *Naive Bayes* por meio de um gráfico.

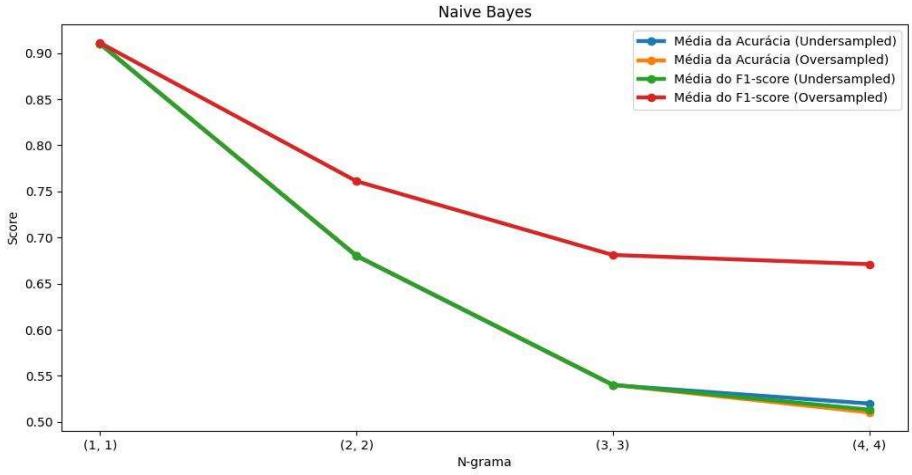


Figura 3.5 – *Naive Bayes*

Fonte: Produzida pelo próprio autor.

Os resultados obtidos apresentam informações sobre N-gramas variando de (1, 1) a (4, 4). Para cada um desses N-gramas, foi calculada a média da acurácia e do F1-*score* para os dados subamostrados (*undersampled*) e superamostrados (*oversampled*).

Para o N-grama (1, 1), a média da acurácia foi de 91% tanto para os dados subamostrados quanto para os superamostrados. O F1-*score* também foi de 91% em ambos os casos. Isso indica que o modelo teve um desempenho muito bom ao considerar apenas uma palavra por vez. Ao aumentar o $N$-grama para (2, 2), houve uma queda significativa na acurácia, passando para 68%. O F1-*score* também diminuiu, mas não tão drasticamente, ficando em 76%. Isso sugere que o

modelo teve mais dificuldade em capturar as relações entre pares de palavras consecutivas. Com o N-grama (3, 3), a acurácia caiu ainda mais, chegando a 54%. O F1-*score* também continuou a cair, atingindo 68%. Esses resultados indicam que o modelo teve ainda mais dificuldades com trios de palavras. Finalmente, para o N-grama (4, 4), a acurácia caiu para 52% e 51% para os dados subamostrados e superamostrados, respectivamente. O F1-*score* se manteve relativamente estável em comparação com o N-grama anterior, com 67%.

Comparando os resultados, é evidente que o aumento do tamanho do $N$-grama resultou em uma diminuição do desempenho do modelo. Isso pode ser devido à complexidade adicional introduzida pelo aumento do número de palavras consideradas juntas. No entanto, é importante notar que, apesar da queda no desempenho, o F1-*score* permaneceu relativamente estável para $N$-gramas maiores. Isso sugere que, embora o modelo possa estar classificando incorretamente um maior número de instâncias, ele ainda é capaz de identificar corretamente muitas das classes positivas.

Devido à falta de recursos, não houve a possibilidade de otimizar os parâmetros do modelo. A falta dessa etapa pode ter contribuído para os resultados menos satisfatórios. O modelo pode não estar adequadamente ajustado aos dados utilizados, resultando em uma performance inferior à sua capacidade real.

Esses resultados indicam que o modelo, com os parâmetros atuais, tem um desempenho melhor ao lidar com $N$-grama (1,1). Os resultados aqui apresentados representam um passo importante na compreensão do comportamento do modelo *Naive Bayes* em relação aos dados utilizados. No entanto, eles também destacam a necessidade de mais pesquisas e experimentação para otimizar o desempenho do modelo.

### 3.3.4 Avaliação Geral Baseada em Acurácia

A Figura 3.6 demonstra uma comparação entre os três algoritmos abordados neste estudo - Regressão Logística, SVM e *Naive Bayes*. Ao comparar as técnicas de *oversampling* e *undersampling*, a primeira se destacou como superior nas análises individuais dos algoritmos. Por isso, esta seção diz respeito à técnica de *oversampling*, com o objetivo de apresentar o algoritmo mais eficaz e a melhor combinação de técnicas para ele utilizando os resultados referentes a acurácia.

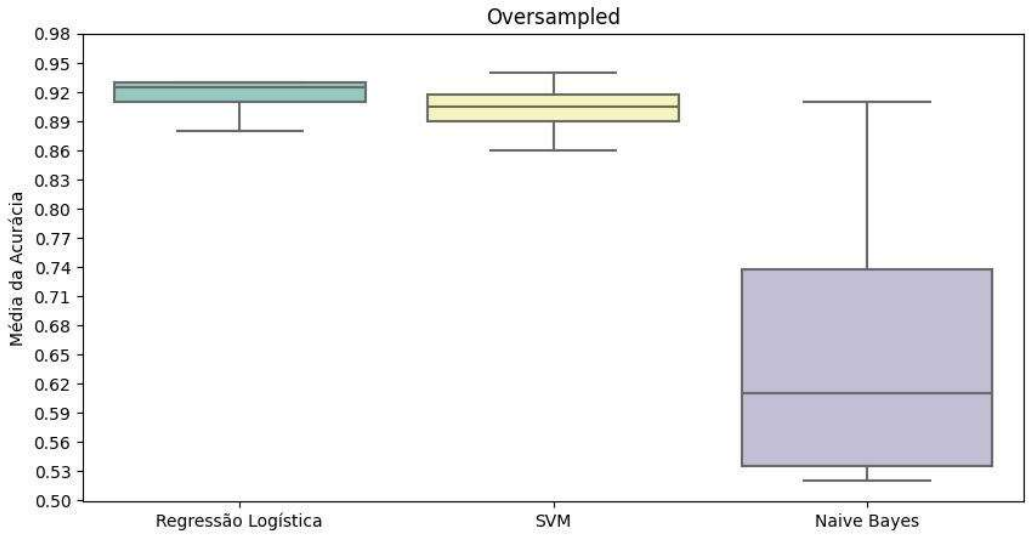


Figura 3.6 – Melhores desempenhos

Fonte: Produzida pelo próprio autor.

Nos algoritmos de Regressão Logística e SVM, o $N$-grama (2,2) provou ser o mais eficiente. Em contraste, para o *Naive Bayes*, o $N$-grama (1,1) se destacou como o ideal. Esses valores correspondem aos pontos de desempenho máximo de cada algoritmo, conforme ilustrado no gráfico referenciado anteriormente.

Com base nesses resultados, conclui-se que o algoritmo Regressão Logística superou os outros algoritmos no conjunto de dados descrito na subseção 3.2.1. Isso indica que, para este conjunto de dados específico - que consiste em *posts* coletados do *Twitter* - o algoritmo Regressão Losgística é o mais adequado para identificar postagens com conteúdo racista. Em uma visão geral, os algoritmos SVM e Regressão Logística apresentaram desempenhos comparáveis, o que reforça a eficácia de ambos na detecção de discursos racistas.

# 4 Considerações Finais

Neste capítulo, serão apresentadas as considerações finais sobre o trabalho, bem como serão apresentadas as direções para trabalhos futuros.

## 4.1 Conclusão

Este estudo investigou a detecção de discursos racistas no *Twitter* utilizando uma abordagem baseada em Aprendizado de Máquina e PLN. A análise dos resultados revelou algumas considerações importantes. Primeiramente, é importante destacar que os resultados obtidos são específicos para o conjunto de dados e as configurações do modelo utilizados, o que ressalta a necessidade de realizar validações adicionais em diferentes conjuntos de dados e tarefas de classificação para generalizar esses achados para outros contextos.

Ao avaliar os modelos Regressão Logística, SVM e *Naive Bayes*, notou-se diferenças significativas em seu desempenho dependendo das configurações de $N$-gramas. O algoritmo de Regressão Logística apresentou um desempenho consistente, enquanto o algoritmo SVM mostrou um desempenho variável dependendo das configurações. Por outro lado, o algoritmo *Naive Bayes* teve uma queda significativa no desempenho com o aumento do tamanho do $N$-grama. Dessa forma, é possível concluir que o tamanho do $N$-grama exerce uma influência significativa nos resultados. A escolha cuidadosa desses fatores pode impactar substancialmente os resultados alcançados, destacando a importância de um ajuste meticuloso desses parâmetros.

Em geral, todos os algoritmos tiveram um desempenho melhor com $N$-gramas menores. A técnica de amostragem não teve um impacto significativo na acurácia dos modelos, exceto no caso do SVM, onde o *undersampling* resultou em um desempenho ligeiramente superior.

Finalmente, a estratégia de combinar várias bases de dados para formar um conjunto de treinamento diversificado provou ser eficaz para lidar com a falta de dados racistas. Essa abordagem produziu resultados superiores aos encontrados na literatura, sugerindo que essa pode ser uma estratégia promissora para melhorar a acurácia dos modelos em estudos futuros.

## 4.2 Trabalhos Futuros

Com o objetivo de avançar na pesquisa e diminuir a lacuna existente na literatura, o autor deste estudo irá expandir este trabalho em várias direções através do seu curso de Mestrado.

Em primeiro momento, planeja-se construir uma base de dados mais robusta e diversificada, composta por *posts* em português. Além disso, pretende-se realizar estudos adicionais sobre a aplicação de PLN neste cenário. O foco será identificar as melhores técnicas de pré-

processamento dos dados que podem melhorar ainda mais a acurácia e outras métricas de avaliação dos modelos de detecção de racismo. Outra direção importante será a exploração de algoritmos de classificação além daqueles tradicionalmente utilizados na literatura.

A combinação dessas estratégias futuras servirá como ferramentas essenciais para mitigar a propagação do racismo no *Twitter*.

# Referências


Statista. **Leading countries based on number of X (formerly Twitter) users as of January 2023**. 2023. Disponível em: <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>. Acesso em 14 de dezembro de 2023.

GUIMARÃES, A. S. A. **Racismo e anti-racismo no Brasil**. [S.l.]: Editora 34, 1999.

ALMEIDA, S. **Racismo estrutural**. [S.l.]: Pólen Produção Editorial LTDA, 2019.

SOUZA, N. S. **Tornar-se negro: ou as vicissitudes da identidade do negro brasileiro em ascensão social**. [S.l.]: Editora Schwarcz-Companhia das Letras, 2021.

ACS. **Direito Fácil - Lei do Racismo**. 2021. Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT). Acesso em 10 de junho de 2023. Disponível em <https://www.tjdft.jus.br/institucional/imprensa/campanhas-e-produtos/direito-facil/edicao-semanal/lei-do-racismo>.

TRINDADE, L. V. P. Mídias sociais e a naturalização de discursos racistas no brasil. **COMUNIDADES, ALGORITMOS E ATIVISMOS DIGITAIS**, p. 26, 2020.

PEREIRA, B.; COSTA, C. T.; CESPEDES, F.; JORGE, S. Dossiê intolerâncias: visíveis e invisíveis no mundo digital. **Associação Brasileira de Comunicação Pública, São Paulo, SP**, 2016.

BOEHM, C. Discursos de ódio e pornografia infantil são principais desafios da internet. **EBC-Empresa Brasil de Comunicação**, v. 6, n. 02, 2018.

DIAS, A. d. S. **Processamento de Linguagem Natural**. E-book. [Digite o Local da Editora]: Editora Saraiva, 2021. Acesso em: 21 jul. 2023. ISBN 9786589881995.

FERREIRA, H. H. Processamento de linguagem natural e classificação de textos em sistemas modulares. 2019.

ANCHIÊTA, R.; NETO, F. A.; MARINHO, J. C.; MOURA, R. Pln: Das técnicas tradicionais aos modelos de deep learning. **Sociedade Brasileira de Computação**, 2021.

BRODER, A. Z.; GLASSMAN, S. C.; MANASSE, M. S.; ZWEIG, G. Syntactic clustering of the web. **Computer networks and ISDN systems**, Elsevier, v. 29, n. 8-13, p. 1157–1166, 1997.

PAIVA, P. D.; SILVA, V. Matias da; MOURA, R. S. Detecção de discurso de ódio utilizando vetores de features aplicados a uma base nova de comentários em português. **Revista de Sistemas e Computação-RSC**, v. 10, n. 1, 2020.

HAYKIN, S. **Neural networks and learning machines, 3/E**. [S.l.]: Pearson Education India, 2009.

CARVALHO, M. V. d. **Aplicações de machine learning na engenharia mecânica: um estudo de caso para diagnóstico da operabilidade de sistemas de abastecimento de água**. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2021.

PAIVA, N.; ESCOVEDO, T. Detecção precoce de alzheimer usando machine learning. 2021.

HARRISON, M. **Machine Learning–Guia de referência rápida: trabalhando com dados estruturados em Python**. [S.l.]: Novatec Editora, 2019.

GONZALEZ, L. d. A. Regressão logística e suas aplicações. Universidade Federal do Maranhão, 2018.

MINUSSI, J. A.; DAMACENA, C.; JR, W. L. N. Um modelo de previsão de solvência utilizando regressão logística. **Revista de Administração Contemporânea**, SciELO Brasil, v. 6, p. 109–128, 2002.

HOSMER, D.; LEMESHOW, S. **Applied logistic regression. New York: John Wiley**. [S.l.]: Sons, 1989.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**. [S.l.]: "O'Reilly Media, Inc.", 2022.

HUANG, S.; CAI, N.; PACHECO, P. P.; NARRANDES, S.; WANG, Y.; XU, W. Applications of support vector machine (svm) learning in cancer genomics. **Cancer genomics & proteomics**, International Institute of Anticancer Research, v. 15, n. 1, p. 41–51, 2018.

ZHANG, H. The optimality of naive bayes. **Aa**, v. 1, n. 2, p. 3, 2004.

CASTRO, L. N. D.; FERRARI, D. G. **Introdução à mineração de dados**. [S.l.]: Saraiva Educação SA, 2017.

PELLE, R. P. D.; MOREIRA, V. P. Offensive comments in the brazilian web: a dataset and baseline results. In: SBC. **Anais do VI Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2017.

NASCIMENTO, R. M. F. d. **Classificação automática de discursos de ódio em textos do Twitter**. Dissertação (B.S. thesis) — Brasil, 2019.

CASTRO, L. d. R. Um estudo empírico sobre técnicas para detecção de discursos de ódio em postagens públicas escritas em português. 2019.

FORTUNA, P.; SILVA, J. R. da; WANNER, L.; NUNES, S. et al. A hierarchically-labeled portuguese hate speech dataset. In: **Proceedings of the third workshop on abusive language online**. [S.l.: s.n.], 2019. p. 94–104.

REIS, M. A. A. d. Predição de comentários em mídias sociais sobre discursos racistas. 2021.

SILVA, R.; FERNANDES, D.; FERNANDES, M. Caracterização de mensagens em língua portuguesa com traços de racismo no twitter. In: **Anais da VI Escola Regional de Informática de Goiás**. Porto Alegre, RS, Brasil: SBC, 2018. p. 205–214. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/erigo/article/view/7140>.

LEITE, J. A.; SILVA, D. F.; BONTCHEVA, K.; SCARTON, C. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In: **Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. [s.n.], 2020. Disponível em: <https://arxiv.org/abs/2010.04543>.

AUGUSTO, M. **Twitter Analysis**. 2021. GitHub repository. Disponível em: <https://github.com/maugustoo/twitter_analysis>.

NETO, S. R. d. S. et al. Uma abordagem computacional para identificação de indício de preconceito em textos baseada em análise de sentimentos. Universidade Federal de Alagoas, 2017.

SILVA, V. d. O. Detecção de fraudes na utilização de cartões usando a técnica de regressão logística: uma aplicação com dados desbalanceados. Universidade Estadual Paulista (Unesp), 2022.