Investigando Racismo Algorítmico na Detecção de Discurso de Ódio no Português do Brasil

Cássia Claudiane Silva da Rosa

Investigando Racismo Algorítmico na Detecção de Discurso de Ódio no Português do Brasil

Cássia Claudiane Silva da Rosa

Orientador: *Profº Drº Renato Porfirio Ishii*

Dissertação entregue à Faculdade de Computação da Universidade Federal de Mato Grosso do Sul - FACOM-UFMS como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

UFMS - Campo Grande
Setembro/2025

"Ser forte diante da opressão não é o mesmo que superá-la. Resistência não deve ser confundida com transformação."
– bell hooks

Agradecimentos

Agradeço a todas as minorias que lutaram antes de mim para que eu pudesse ocupar um lugar dentro da Faculdade de Computação. Agradeço aos meus irmãos, Adrian Rosa e Adriene Rosa, por compartilharem momentos alegres comigo, por me aceitarem e por me apoiarem. Agradeço as amizades que fiz durante estes anos em Mato Grosso do Sul, obrigada por fazerem eu me sentir em casa.

Obrigada Adriana Silva, minha mãe e primeira professora, por me mostrar que a educação tem o poder de transformar vidas, por me ensinar que coisas boas levam tempo para serem alcançadas. Também agradeço pelo apoio emocional e financeiro, durante esses longos anos.

Agradeço aos meus avós, Doralice Gonçalves e Pedro Miranda, pelo cuidado, paciência e trabalho árduo, que foi de muita importância na minha formação. Obrigada por serem as minhas raízes.

Agradeço ao meu companheiro, Gustavo Souza, por me incentivar a não desistir dos meus sonhos e pelas horas dedicadas aos meus cuidados e afetos. Também agradeço aos meus primos, Nailde Dória e Edcarlos Soares, pois sem o acolhimento e apoio deles isto não seria possível.

Ao meu orientador, Renato Ishii, agradeço pela companhia nesta caminhada, pelos conselhos, pelas oportunidades e possibilidades. Agradeço também ao professor Fábio Viduani, por seu companheirismo e dedicação durante minha graduação.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

PLN Processamento de Linguagem Natural

IA *Inteligência Artificial*

AI *Artificial Intelligence*

PL *Projeto de Lei*

BoW *Bag of Words*

TF-IDF *Term Frequency - Inverse Document Frequency*

RSLP *Removedor de sufixos da Língua Portuguesa*

CBoW *Continuou Bag of Words*

MLP *Multilayer Perceptron*

BERT *Bidirecional Encoders Representations from Transformers*

GML *Grandes Modelos de Linguagem*

LLM *Large Language Models*

VP *Verdadeiros Positivos*

VN *Verdadeiros Negativos*

FP *Falsos Positivos*

FN *Falsos Negativos*

UE *União Europeia*

AAE *Inglês Afro-Americano*

SAE *Inglês Americano Padrão*

STM *Modelo de Tópico Estrutural*

TFP *Taxa de Falsos Positivos*

TFN *Taxa de Falsos Negativos*

SMOTE *Técnica de Sobreamostragem Minoritária Sintética*

RL *Regressão Logística*

AUC *Área sob a Curva Característica de Operação do Receptor*

ANOVA *Análise de Variância*

WMW *Willcoxin-Mann-Whitney*

ICCSA *Conferência Internacional sobre Ciência Computacional e suas Aplicações*

ICMLA *Conferência Internacional de Aprendizado de Máquina e Aplicações*

GPU *Unidade de Processamento Gráfico*

TPU *Unidade de Processamento de Tensor*

Abstract

Algorithmic bias gained notoriety in over the years, due to impacts caused by Artificial Intelligence (AI) systems in historically discriminated vulnerable groups [5] [6] [7]. Hate speech detection models are used to identify unwanted contents in social media platforms like Instagram, Threads and X. These platforms are often used as data source during data collection process, in dataset constructions utilized for training AI models to detect hate speech. However, the acquisition of bias can occur in many steps of development cycle of AI models. Therefore, this work contributes to diminish algorithmic racism through application of a methodology that evaluates the presence of racial bias in databases and trained classifiers for hate speech detection in Brazilian Portuguese. The models Logistic Regression (LR), Multilayer Perceptron (MLP), BERTimbau$_{base}$, and Tucano-1b1, were trained with OFFCOMBR-2, HateBR, and TuPy-E databases for hate speech detection. After this, the models were used to predict hate speech on BR-RAPData, a database constructed through data collection of Brazilian RAP lyrics. Our results show that at least one model trained with OFFCOMBR-2, HateBR, or TuPy-E, achieved F1 score over 70%, a substantial result according to the literature [8] [9] [10]. Nonetheless, analyses of the BR-RAPData prediction for hate speech shows that, in some cases, over 50% of the content in this dataset was targeted as hate speech by the trained classifiers.

Keywords — Natural Language Processing; Algorithmic Racism; Hate Speech Detection; Algorithmic Bias; Artificial Intelligence.

Resumo

O tema viés algorítmico tem recebido crescente notoriedade nos últimos anos, devido aos impactos causados por sistemas de Inteligência Artificial (IA) em grupos sociais historicamente discriminados [5] [6] [7]. Os modelos para detecção de discurso de ódio são uma opção viável para identificar conteúdos indesejados em redes sociais, como *Instagram*, *X* e *Threads*. Durante o processo de construção de conjuntos de dados, estas plataformas são usadas como fonte de dados para coletada de publicações, que serão usadas no treinamento de modelos para classificação de discurso de ódio ou linguagem ofensiva. No entanto, em diversos estágios do ciclo de desenvolvimento de um modelo, pode ocorrer a aquisição de discriminação algorítmica. Nesse sentido, este trabalho realiza a investigação de racismo algorítmico em bases de dados e classificadores de discurso de ódio, na língua portuguesa do Brasil. Os modelos de Regressão Logística (RL), *Multilayer Perceptron* (MLP), BERTimbau$_{base}$ e Tucano-1b1 foram treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E, para a classificação de discurso de ódio e, após isso, uma metodologia para investigação de racismo algorítmico foi proposta com base na predição de discurso de ódio sobre a base de dados BR-RAPData, construída neste trabalho, via coleta de músicas de RAP brasileiras. Como resultado, cada modelo treinado com as respectivas bases de dados: OFFCOMBR-2, HateBR ou TuPy-E, obteve valores para métrica F1 acima de 70%, sendo estes, resultados substanciais. No entanto, analises sobre as proporções da classe de discurso de ódio, preditas sobre a base de dados BR-RAPData, mostraram que, em alguns casos, mais de 50% da base de dados de RAP fora rotulada como discurso de ódio pelos classificadores.

Palavras-chave — Processamento de Linguagem Natural; Racismo Algorítmico; Detecção de Discurso de Ódio; Discriminação Algorítmica; Inteligência Artificial.

CAPÍTULO

1

Introdução

Com o grande volume de dados disponíveis em uma velocidade nunca vista, emergiu a "*Era do Big Data*". Neste novo cenário, os sistemas inteligentes passaram a ser ainda mais utilizados, desenfreando uma "moda" nas últimas duas décadas. Cada vez mais inseridos em tomadas de decisões que afetam nossas vidas, os modelos de IA corroboram para a permanência de desigualdades entre os menos favorecidos, enquanto mantém os privilégios dos que estão no poder [11]. Essa prática, não ameaça somente a vida daqueles que estão na margem, mas a sociedade na totalidade. Para Evgeny Morozov [12], o estado democrático é ameaçado por uma corrente política extremamente tecnocrata. Dentre as diversas problemáticas despertadas pelo tecnosolucionismo, preconceitos de gênero e raça são frequentemente expostos em ambientes virtuais de interação humana.

A Discriminação Algorítmica, ou viés codificado, é uma forma de disseminar preconceitos e estereótipos, atingindo grupos vulneráveis mediante sistemas computacionais, contribuindo para o reforço de desigualdades estruturais e sistêmicas [13]. O projeto tecnológico sofre com a carência de representatividade e a exclusão, desde o momento em que é planejado. Como consequência, a vida de pessoas expostas a essas ferramentas são impactadas negativamente. O viés codificado está presente em diversas ferramentas de Inteligência Artificial (IA), como reconhecimento facial [5], predição de reincidência criminal [7], análise curricular [14], tradução de idiomas [15] e detecção de discurso de ódio [16].

Em setembro de 2024, usuários do *Instagram* e *Threads* observaram que, ao pesquisar a palavra "negra", e suas derivações, nos mecanismos de busca dessas mídias sociais, era emitida uma mensagem, alertando sobre a associação deste termo com a venda de drogas[1]. A associação da negritude com elementos pejorativos, advindo

[1]Alma Preta Jornalismo. Instagram e Threads associam palavra 'negra' com venda de drogas; Meta cita problema técnico. Alma Preta, 2024. Disponível em <https://almapreta.com.br/sessao/cotidiano/instagram-e-threads-associam-palavra-

de ferramentas computacionais, não é novidade. Em 2010, a professora universitária Safyia Noble pesquisou "meninas negras" no buscador do *Google*, e obteve resultados relacionados à pornografia [17]. Estas situações são descritas como racismo algorítmico, definido, segundo Tarcízio Silva [18], como uma prática discriminatória perpetuada por meio de sistemas e protocolos de programação dos novos meios digitais, que pode reforçar e até mesmo naturalizar o racismo estrutural.

Em ambientes virtuais, a moderação de conteúdo provê o bem-estar entre usuários por meio de acordos bem definidos, como as diretrizes da comunidade, que especificam quais comportamentos são inadequados dentro daquela rede. Entre esses acordos, está a proibição da veiculação de símbolos e discursos de ódio, uma prática que promove preconceitos contra comunidades vulneráveis e agrava situações de violência dentro e fora da Internet[2]. No entanto, em janeiro de 2025, o diretor-executivo da empresa estadunidense *Meta*, anunciou mudanças significativas sobre a moderação de conteúdo *online* das plataformas de redes sociais, *Instagram*, *Facebook* e *Threads*[3]. Dentre essas mudanças, estão: a substituição de checadores de conteúdos por notas da comunidade; e a exclusão de restrições de conteúdo sobre determinados tópicos, como imigração e gênero. De modo similar, esta mudança de paradigma sobre as regras de ambientes virtuais também ocorreu na rede social *X*, ao ser adquirida pelo bilionário Elon Musk. Ainda mais, em ambos os cenários, esta reviravolta fora marcada pelo desligamento de setores ligados à diversidade e inclusão. Isso mostra que a conduta adotada pelas grandes empresas de tecnologia, ao escolher ignorar problemas estruturais e sistêmicos, reforça as relações de poder existentes na nossa sociedade [19].

Sendo assim, combater o discurso de ódio online é importante para atenuação de seus efeitos sobre as comunidades atingidas. Todavia, analisar o grande contingente de publicações, realizadas todos os dias, manualmente, é uma tarefa fatigante para moderadores de conteúdo online e, por este motivo, as soluções automatizadas são usadas para lidar com o problema, como os modelos de IA. Dessa maneira, pesquisadores integram conceitos da definição de Discurso de Ódio com Processamento de Linguagem Natural (PLN), para construir sistemas capazes de identificar comportamentos abusivos. Essa estratégia, entretanto, pode ter efeitos negativos, como, ao identificar abusos *online*, cometer atos discriminatórios contra aqueles que deveriam ser protegidos por tais ferramentas [20].

Durante o desenvolvimento de instrumentos para detecção de discurso de ódio, existem etapas nas quais os preconceitos e estereótipos são codificados. Na coleta de dados, pesquisadores enfrentam problemas ao lidar com a falta de exemplos de

negra-com-venda-de-drogas-meta-cita-problema-tecnico/>. Acessado em: 25 jul. 2025

[2]TOMAZ, Kleber. Polícia investiga se briga de torcidas em SP foi marcada pela internet. G1 São Paulo, 2012. Disponível em <https://g1.globo.com/sao-paulo/noticia/2012/03/policia-investiga-se-briga-de-torcidas-em-sp-foi-marcada-pela-internet.html>. Acessado em: 25 jul. 2025

[3]HENDRIX, Justin. *Transcript: Mark Zuckerberg Announces Major Changes to Meta's Content Moderation Policies and Operations*. *Tech Policy.Press*, 2025. Disponível em <https://www.techpolicy.press/transcript-mark-zuckerberg-announces-major-changes-to-metas-content-moderation-policies-and-operations/>. Acessado em: 25 jul. 2025

discurso de ódio [21]. Embora esta seja uma manifestação muito presente nas redes sociais, a quantidade de conteúdos não nocivos na Internet é muito maior que a quantidade de conteúdos nocivos. Ainda neste processo, a escolha de palavras-chave para capturar mensagens potencialmente ofensivas, é uma estratégia que censura comunidades que ressignificaram termos pejorativos [22]. Outrossim, os vieses dos anotadores também são inferidos nos dados durante a rotulação [16]. Além disso, a discrepância no conceito de discurso de ódio [23] e a má interpretação do contexto [24], contribuem para que as sentenças sejam anotadas incorretamente [25]. Com isso, a aprendizagem de máquina é comprometida, por meio da utilização de conjuntos de dados que refletem preconceitos sociais [26]. Em adição, abordagens que alcançaram o estado da arte para a tarefa de classificação de discurso de ódio possuem falhas, como o super ajuste dos dados, afetando a veracidade dos resultados obtidos [27]. Outrossim, os Grandes Modelos de Linguages (GML), ou *Large Language Models* (LLM), por serem pré-treinados com textos extraídos de uma variedade de páginas da Internet, são expostos a símbolos e discursos de ódio, podendo reproduzir tais manifestações [28].

Diante de problemáticas apresentadas acima, que envolvem a falta de comprometimento de grandes empresas de tecnologia, desbalanceamento nas bases de dados, viés pessoal adquirido na rotulação e modelos pré-treinados enviesados, comunidades podem ser censuradas devido ao conteúdo que compartilham na Internet. No ano de 2023, a empresa *Meta*, responsável pelas redes sociais *Instagram* e *Facebook*, removeu conteúdos pró-Palestina, em ambas plataformas, sob a premissa de "*implementação errônea e dependência excessiva de ferramentas automatizadas para moderação de conteúdo*"[4]. Enquanto isso, esses mecanismos que censuram os oprimidos, impulsionam propagandas racistas e sexistas dentro de sua rede, evidenciando uma inconsistência na moderação de conteúdo online [18]. Logo, fazendo necessária a investigação de discriminação algorítmica.

1.1 Objetivos

Diante das motivações citadas acima, fez-se necessária a avaliação da presença de viés codificado em bases de dados e modelos de IA, usados para detecção de discurso de ódio automatizada. Sendo assim, este trabalho adotou como principal objetivo a investigação de discriminação algorítmica, especificamente, racismo algorítmico, em modelos e bases de dados para detecção de discurso de ódio *online*, no idioma português do Brasil.

Ao atingir o objetivo principal, os seguintes objetivos específicos também foram concluídos:

1. Identificar quais características dos conjuntos de dados podem estar relaciona-

[4]LUSCOMBE, Richard. *Meta censors pro-Palestinian views on a global scale, report claims*. *The Guardian*, 2023. Disponível em <https://www.theguardian.com/technology/2023/dec/21/meta-facebook-instagram-pro-palestine-censorship-human-rights-watch-report>. Acessado em: 25 jul. 2025.

das ao melhor desempenho dos modelos;

2. Identificar se há correlação entre a qualidade das bases de dados e racismo algorítmico;

3. Construir uma base de dados alinhada racialmente para investigação de racismo algorítmico

Além disso, a partir dos experimentos realizados para concluir os objetivos listados acima, torna-se viável responder às seguintes perguntas:

1. Os modelos selecionados podem classificar corretamente discurso de ódio?

2. Como investigar o racismo algorítmico propagado pela detecção de discurso de ódio automatizada?

Dessa forma, este trabalho contribui com a avaliação de discriminação algorítmica em Inteligência Artificial, sob a hipótese de que há perpetuação de racismo algorítmico em conjuntos de dados no português brasileiro e em modelos de linguagem natural treinados com eles para detecção de discurso de ódio.

Sendo assim, foi definida uma metodologia para investigação de racismo algorítmico nas bases de dados OFFCOMBR-2 [8], HateBR [9] e TuPy-E [10], usadas para treinar modelos de IA para detecção de discurso de ódio. Ainda mais, foi construído, neste trabalho, o conjunto de dados BR-RAPData. Com um roteiro para coleta de dados, desenvolvido na linguagem de programação *python*, com a biblioteca *BeautifulSoup*, a discografia dos artitas brasileiros Racionais MC's, Tasha e Tracie, Negrali, Emicida e Atitude Feminina foram coletadas da plataforma de músicas Vagalume. Seguindo a metodologia de Sap e colaboradores [16], Davidson e colaboradores [29] e Xia e colaboradores [30], esta base de dados fora usada para conduzir os experimentos de verificação de racismo algorítmico, com base na discriminação da variação linguística do português, chamada pretuguês [31], conforme feitos pelos autores citados acima com as bases de dados de Blodgett e colaboradores [32] e Preoțiuc-Pietro e Ungar [33]

Os experimentos realizados incluem a seleção de bases de dados para detecção de discurso de ódio; seleção, treinamento e ajuste de modelos de linguagem natural para classificação; bem como a predição de discurso de ódio sobre o conjunto de dados BR-RAPData, usado como identificador racial; e análises de significância estatística através dos testes de normalidade, qui-quadrado e Kruskal-Wallis.

Como resultado, os modelos Regressão Logística (RL), *Multilayer Perceptron* (MLP), BERTimbau base e Tucano-1b1, treinados com as bases de dados OFFCOMBR-2 [8], HateBR [9] e TuPy-E [10], identificaram diversas sentenças da base de dados BR-RAPData como discurso de ódio, levando à emissão de taxas de falsos positivos (TFP) de até 84%. Além disso, o modelo MLP, treinado com vetores de palavras extraídos com a técnica TF-IDF, das bases de dados OFFCOMBR-2 e TuPy-E, obtiveram métrica F1 igual a 83% e 74%, respectivamente. Enquanto o modelo BERTimbau$_{base}$, treinado com a base de dados HateBR, obteve média F1 igual a 91%. Ainda mais, os modelos treinados atingiram valores para métrica F1 iguais ou acima de 70%, quando

treinados com pelo menos uma das bases de dados investigadas. Logo, respondendo a pergunta 1, os modelos selecionados foram capazes de classificar discurso de ódio corretamente, segundo a literatura [8] [9] [10].

Ainda mais, foram observadas diferentes proporções de discurso de ódio para a base de dados BR-RAPData. Foi constatado que os conjuntos de dados, e os respectivos modelos treinados com eles, reproduziram estereótipos raciais, confirmando a hipótese supracitada. Para isso, fora realizada a investigação de racismo algorítmico, a partir da metodologia de investigação de racismo algorítmico, proposta neste trabalho com base nos trabalhos de Davidson e colaboradores [29] e Sap e colaboradores [16]. Sendo assim, foram avaliadas as amostragens de sentenças classificadas como discurso de ódio, do conjunto de dados BR-RAPData. Com o teste de significância estatística de Kruskal-Wallis, foi constatado que houve variação entre as classificações realizadas por modelos que foram treinados com diferentes bases de dados. No entanto, o teste qui-quadrado revelou que esta variação não permanece ao comparar as proporções emitidas por modelos treinados com a mesma base de dados. Ou seja, modelos treinados com OFFCOMBR-2, HateBR ou TuPy-E classificaram diversas sentenças do conjunto de dados BR-RAPData como discurso de ódio. Entretanto, as proporções de sentenças classificadas como discurso de ódio divergem entre si, conforme variação de base de dados, modelo e técnica de *word embeddings*. Sendo que, técnicas mais avançadas apresentou maiores proporções de discurso de ódio, e, dentre os modelos treinados com a base de dados HateBR, a técnica de MLP, combinada com representação vetorial *Continuous Bag of Words* (CBoW), apresentando a maior taxa de falsos positivos, igual a 84%.

Com os experimentos concluídos, uma discussão fora realizada sobre os resultados, bem como as dificuldades encontradas durante a detecção de discurso de ódio. Pois, essa é uma tarefa que envolve ambiguidade em relação aos conceitos de discurso de ódio e liberdade de expressão, podendo ser incompreensível para uma máquina. Com isso, foi possível perceber que mesmo entre modelos com bom desempenho, houve a permanência de racismo algorítmico. Ainda mais, o desbalanceamento entre classes de discurso de ódio e não discurso de ódio é um problema para detecção de discurso de ódio, que urge por soluções mais significativas, pois, afetam diretamente o desempenho do modelo [26].

As tecnologias e metodologia utilizadas para concluir as tarefas listadas acima estão dispostas nos próximos capítulos deste trabalho. O Capítulo 2 aborda a Fundamentação Teórica, dividida entre as Seções 2.2, 2.3 e 2.4, que abordam Discriminação Algorítmica, Detecção de Discurso de Ódio, e Mitigação e Investigação de Discriminação Algorítmica em Tecnologias para Detecção de Discursos de Ódio, respectivamente. O Capítulo 3 apresenta a metodologia utilizada para investigação de racismo algorítmico, bem como os resultados alcançados. Por fim, o Capítulo 4 reúne a discussão sobre os resultados encontrados, limitações, conclusão e trabalhos futuros.

CAPÍTULO

2

Referencial Teórico

2.1 Considerações Iniciais

Neste capítulo, serão apresentados: na Seção 2.2, conceitos essenciais para compreensão da Discriminação Algorítmica, bem como definições sobre aquisição de viés, racismo algorítmico e ética em IA; na Seção 2.3, serão mostradas técnicas de Processamento de Linguagem Natural e mineração de dados usadas para a Detecção de Discurso de Ódio, desde a etapa de coleta de dados até a escolha e funcionamento dos modelos, além de conceituar o que é considerado discurso de ódio e seus estilos de manifestações; por fim, na Seção 2.4, serão apresentados os trabalhos relacionados à operação de racismo algorítmico em tecnologias para detecção de discurso de ódio, bem como os métodos de avaliação e mitigação de viés utilizados nestes trabalhos.

2.2 Discriminação Algorítmica

Os preconceitos enraizados e ainda fortalecidos fazem com que surjam novas formas de discriminação contra grupos minoritários. Diariamente, pessoas sofrem por conta do seu gênero, da sua orientação sexual, raça, etnia, classe social, religiosidade, entre outras características que as classificam como minorias. Segundo Benjamin [13], a discriminação algorítmica é uma forma de perpetuar os diferentes tipos de discriminação, por meio da adaptação ao decorrente contexto histórico-social, em que as tecnologias desempenham um papel fundamental. Desse modo, as desigualdades são projetadas em ferramentas digitais durante as etapas do seu desenvolvimento. O surgimento da Inteligência Artificial, em meados dos anos 1950, permitiu que pesquisadores da área de Ciência da Computação elaborassem algoritmos para solução de problemas a partir de regras bem definidas, seguindo uma lógica computacional, em que o seu objetivo era a resolução a partir da aprendizagem automática. Estes

sistemas eram capazes realizar decisões baseadas em um conjunto de decisões anteriores, sem que a intervenção humana fosse necessária. Entretanto, o desenvolvimento desta tecnologia fora historicamente pensado por um grupo seleto de pessoas, majoritariamente homens e brancos, e isso permitiu que seus vieses fossem transmitidos para as máquinas [34].

Somada à pouca representatividade, a utilização de ferramentas de IA para tarefas com impacto social foi outro aspecto relevante para que a discriminação algorítmica entrasse em pauta. Além disso, a Era dos Dados (*Big Data*) [35], implicou no compartilhamento e uso desenfreado de informações. Nesse cenário, um conjunto de fatores, como a crescente utilização de mídias sociais, aumentos nas capacidades de armazenamento e processamento de dados, a computação ubíqua, etc., permitiram que os modelos de IA fossem treinados com volumosas bases de dados. Sendo assim, para processar essa abundância de conhecimento, os algoritmos passaram a ter maior complexidade, dificultando a compreensão sobre quais operações eram realizadas até a obtenção de um resultado, assim, permitindo o agravamento do viés codificado.

Frente a isso, pessoas pertencentes a grupos historicamente marginalizados passaram a sofrer preconceitos e violências via sistemas de aprendizado de máquina. A empresa *Amazon* utilizou uma ferramenta de recrutamento que agia discriminadamente contra mulheres, atribuindo notas baixas para os seus currículos, caso tivessem alguma referência ao gênero feminino no documento [14]. Em um estudo feito pela Universidade Federal do Rio Grande do Sul, pesquisadores comprovaram que a tradução automática do *Google Translate* perpetuava estereótipos, centralizando o gênero masculino como padrão para a tradução de áreas profissionais associadas ao gênero masculino, como medicina e engenharia [15].

Pessoas LGBTQIA+ também sofrem com os preconceitos difundidos pela utilização de Inteligência Artificial. A Interface de Programação Aplicada (API) *Perspective*[1], que identifica comentários tóxicos, avaliou os *tweets* de *drag queens* como mais tóxicos que o de supremacistas brancos, devido ao emprego de termos comumente utilizados pela comunidade LGBTQIA+, mostrando pouca eficiência da ferramenta ao lidar com diferentes contextos sociais [36]. Outrossim, a binaridade de gênero empregada por ferramentas de processamento de linguagem natural comete erro de gênero, ao se referenciar a pessoas trans e não-binárias como pessoas cis gênero, além de impulsionar o apagamento dessas comunidades [37].

2.2.1 Aquisição de viés

Conforme ilustrado na Figura 2.1, Madelena Y. Ng e colaboradores [1] argumentam que o ciclo de desenvolvimento de um modelo de IA pode ser compreendido através das seguintes fases: coleta de dados; rotulação de dados; treinamento e avaliação do modelo; implementação, monitoramento e manutenção do modelo. Em algumas dessas fases, há possibilidade de aquisição de discriminações, causando prejuízos para

[1]Perspective. Disponível em <https://www.perspectiveapi.com/>. Acessado em: 25 jul. 2025.

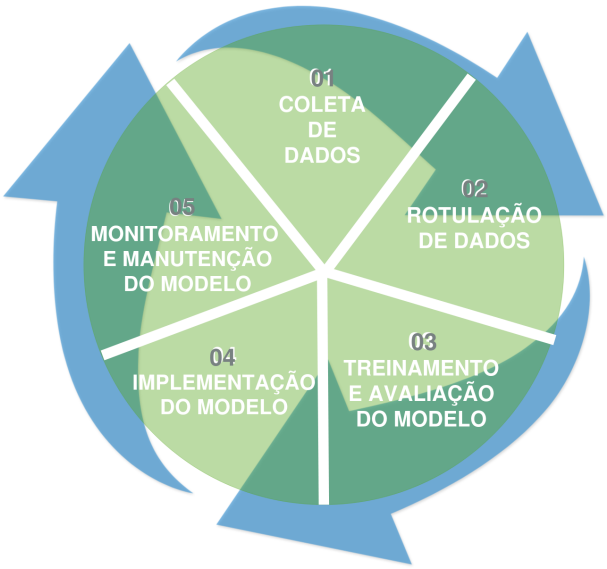


Figura 2.1: Ciclo de desenvolvimento de um modelo de IA, segundo Madalena Y. Ng e colaboradores [1]

a sociedade [38]. As bases de dados utilizadas para o treinamento de modelos são formadas por dados coletados a partir de interações em espaços físicos e virtuais, que transmitem aspectos individuais e coletivos das pessoas na sociedade, incluindo preconceitos, injustiças e desigualdades [39]. Além disso, essas bases de dados podem ser desbalanceadas, contendo mais amostras de uma classe, e, portanto, favorecendo-a [40]. A estratégia adotada para anotação, assim como a quantidade de anotadores, nível de experiência e diversidade entre eles, são fatores que afetam a qualidade da anotação dos dados. Sendo possível que os preconceitos dos anotadores sejam transmitidos para os dados durante esse processo, alterando o desempenho do modelo [24].

Ainda mais, a opacidade no aprendizado de máquina dificulta a compreensão da funcionalidade desses modelos [41], viabilizando que o algoritmo consiga discriminar comunidades, aprendendo por meio de dados demográficos utilizados no seu desenvolvimento [42]. A falta de representatividade nas bases de dados também corrobora para que modelos de IA se tornem enviesados [43]. Em 2018, Buolamwini e Gebru [5] conduziram uma avaliação em ferramentas de reconhecimento facial das empresas *Microsoft*, *Face*++ e *IBM*, e suas respectivas bases de dados utilizadas para treinamento. As pesquisadoras comprovaram que esses sistemas obtiveram um desempenho superior em rostos de homens de pele clara, enquanto foram pouco eficientes em rostos de pessoas de pele escura, sobretudo mulheres. Isso ocorreu porque, a maioria dos rostos usados para compor o conjunto de dados, usados no treinamento dos modelos, eram de pessoas brancas, principalmente homens.

2.2.2 Racismo Algorítmico

Brutalidade policial, desigualdade salarial e maior dificuldade para acessar e permanecer no ensino superior, são exemplos de como o racismo atinge as pessoas negras. Não somente isso, também mostra como ele está inserido em instituições [44],

evidenciando o pacto da branquitude que expõe indivíduos racializados a situações de desvantagem e violência [45]. Além de ser difundido em instituições, o racismo também é propagado via tecnologias digitais. Segundo Tarcízio Silva [18], o racismo algorítmico é uma prática discriminatória perpetuada por meio de sistemas e protocolos de programação dos novos meios digitais, que pode reforçar e até mesmo naturalizar o racismo estrutural.

Com isso, os impactos gerados por essas tecnologias podem ser irreversíveis para a comunidade negra. Um estudo realizado no Brasil, pela Rede de Observatórios de Segurança, demonstrou que 90,5% das pessoas presas através do uso de reconhecimento facial, pela segurança pública, eram negras [46]. Isso mostra como a ferramenta reforça a seletividade e o racismo do sistema penal brasileiro [6]. Nos Estados Unidos, uma pesquisa revelou que o software de predição de reincidência criminal, *COMPAS*, rotulava detentos negros como mais prováveis de cometerem novos crimes do que detentos brancos [7]. Com esses exemplos, é possível notar a disparidade racial presente em sistemas de IA, comprovando que essas ferramentas podem ser utilizadas para prejudicar populações Afrodiaspóricas.

É importante ressaltar que o racismo algorítmico não atinge somente pessoas negras. Mediante experimentos com o GML GPT-3, pesquisadores demonstraram que essa ferramenta apresentou preconceito contra pessoas árabes e muçulmanas, associando-as ao terrorismo [47]. Evidenciando, também, uma discriminação com base na religião. Com as diferentes formas de discriminação perpetuadas, urge o emprego de abordagens mais éticas, que diminuam os efeitos negativos causados por modelos de IA.

2.2.3 Ética em Inteligência Artificial

Para ser possível uma Inteligência Artificial mais confiável, é necessário haver transparência durante todo o processo de desenvolvimento da tecnologia [48]. As pessoas afetadas precisam saber se seus dados estão sendo coletados e para qual finalidade serão utilizados. Também precisam saber se há alguma ferramenta de Inteligência Artificial as monitorando, e quais são os impactos causados por ela em suas vidas. Além de precisarem saber como essas ferramentas funcionam [49]. Assim, a explicabilidade sobre os processos realizados no desenvolvimento de sistemas de IA é fundamental. Isso permite que os envolvidos tenham conhecimento sobre o que é feito durante o ciclo de desenvolvimento dos modelos, permitindo compreender as decisões tomadas pelo algoritmo [42].

Além da transparência, a responsabilização, ou "*accountability*", também é importante. Corporações que atuam na fabricação de *software* precisam ser responsabilizadas pelo desenvolvimento das ferramentas e dos possíveis danos que podem ser causados por sua implementação. Outras partes interessadas, que agem para reforçar relações de poder, mantendo seus privilégios através da utilização de sistemas de IA, também devem ser responsabilizadas pelos seus danos [50].

Perante o cenário de tecnologias que atuam em benefício de uma parcela da po-

pulação, é importante construir algoritmos mais justos e igualitários. Assim, surgem propostas para criar normas que regem a utilização de Inteligência Artificial. A governança em IA é essencial para diminuição de impactos negativos da tecnologia na sociedade. Com ela, é possível administrar e prevenir erros, agindo de maneira ética perante situações de desigualdade, injustiça e discriminação [51]. Com isso, surgem iniciativas públicas e privadas, com intuito de monitorar o uso de Inteligência Artificial e mitigar seus efeitos negativos.

No Brasil, o Centro de Estudos de Segurança e Cidadania criou o projeto O Panóptico[2], responsável por monitorar projetos que utilizam reconhecimento facial no território brasileiro. Segundo os dados desta instituição, existem 209 projetos que implementam técnicas de reconhecimento facial no país, com mais de 71 milhões de pessoas potencialmente vigiadas. Além disso, a campanha Tire Meu Rosto da Sua Mira[3] age para que o uso de reconhecimento facial pela segurança pública no Brasil seja banido. Alegando que esses sistemas agravam o racismo presente no sistema penal brasileiro. A organização *Algorithmic Justice League*[4] trabalha para informar a sociedade sobre os impactos negativos causados pelo uso de IA.

Em 2019, a União Europeia (UE) criou uma abordagem para discutir diretrizes éticas que garantam confiabilidade em tecnologias de Inteligência Artificial. O documento elaborado pela UE aborda questões sobre transparência, responsabilidade, justiça e inclusão social, com intuito de promover uma utilização responsável de IA, respeitando os direitos humanos fundamentais [52]. O Projeto de Lei de Número 2.338 de 2023[5], dispõe sobre o uso da Inteligência Artificial no Brasil, e define que o desenvolvimento, implementação e utilização de sistemas de IA precisam ter como fundamentos: a centralidade da pessoa humana, igualdade, não discriminação, entre outros aspectos [53], demonstrando interesse do Estado Brasileiro em contribuir para implementação de sistemas de IA mais justos e confiáveis.

2.3 Detecção de Discurso de Ódio

2.3.1 Discurso de Ódio

O discurso de ódio é um fenômeno que vitimiza pessoas de diferentes grupos sociais, dentro e fora da Internet. A abundância de publicações que reproduzem atos de violência como racismo [54], misoginia [55] [56], homofobia [57], entre outros [3], são realizadas diariamente por milhares de usuários, e isto exige que haja um filtro eficiente para barrar a veiculação desse tipo de conteúdo. Por exemplo, o *Instagram*, uma

[2]O Panóptico. Disponível em<https://www.opanoptico.com.br/>. Acessado em: 25 jul. 2025

[3]Tire Meu Rosto da Sua Mira. Disponível em <https://tiremeurostodasuamira.org.br/>. Acessado em: 25 jul. 2025.

[4]*Algorithmic Justice League*. Disponível em <https://www.ajl.org/>. Acessado em: 25 jul. 2025.

[5]PACHECO, Rodrigo. PL 2338/2023. Portal da Câmara dos Deputados, 2025. Disponível em <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262>. Acessado em: 25 jul. 2025.

das redes sociais sob o domínio da empresa *Meta*, possui termos de uso[6], definindo acordos entre o serviço oferecido pela plataforma e o usuário; e diretrizes da comunidade[7], estabelecendo quais comportamentos são inapropriados dentro da rede. Os moderadores de conteúdo *online* também desempenham um papel fundamental para manutenção do bem-estar entre usuários dentro das mídias sociais.

No escopo da pesquisa em discurso de ódio, existem definições acerca do que deve ser considerado como discurso de ódio ou não. Segundo Davidson e colaboradores, "*discurso de ódio é uma linguagem usada para expressar ódio contra um grupo-alvo ou que tem a intenção de ser depreciativo, humilhar ou insultar os membros do grupo*" [57, pg.1]. Nesta esfera, também é viável categorizar o discurso de ódio, através da análise do seu conteúdo, em dois sentidos: forma de expressão de discurso de ódio; e tipo de discurso de ódio. Segundo Rosenfeld [58], o discurso de ódio pode ser expresso implicitamente ou explicitamente. As sentenças são classificadas como "discurso de ódio explícito" quando há qualquer manifestação explicitamente odiosa, como em ameaças de morte e insultos racistas. Quando há necessidade de compreensão prévia de determinado assunto, por exemplo, a negação de algum fato histórico, a sentença é considerada na forma de 'discurso de ódio implícito". Além das formas de expressão, existem diferentes tipos de discurso de ódio. Como visto na Tabela 2.1, Fortuna e colaboradores [3] definem nove tipos de discurso de ódio.

Tabela 2.1: Tipos de discurso de ódio, conforme a abordagem adotada por Fortuna e colaboradores em [3].

Tipo	Definição
Sexismo	Discurso de ódio baseado em gênero
Corporal	Discurso de ódio baseado no corpo, como gordofobia
Origem	Discurso de ódio baseado no lugar de origem, como xenofobia
Homofobia	Discurso de ódio baseado na orientação sexual
Racismo	Discurso de ódio baseado na etnia ou raça
Ideológico	Discurso de ódio baseado na ideologia política
Religioso	Discurso de ódio com base na religiosidade
Saúde	Discurso de ódio com base em condições de saúde
Outro estilo de vida	Discurso de ódio baseado em estilos de vida fora do padrão

Além destes, outros tipos de discurso de ódio são alcançáveis a partir da interscci-onalidade entre eles, como a misoginia, que é o discurso de ódio derivado do sexismo, em que o grupo alvo são mulheres [59]. A Tabela 2.2 mostra dois exemplos com diferentes tipos de discurso de ódio expressados de formas distintas. A primeira afirmação evidencia, implicitamente, a raça branca como superior, um exemplo de racismo velado. Enquanto na segunda afirmação, é notório o uso de termos misóginos, bem como o estímulo à violência, categorizando discurso de ódio explícito.

[6]Termos de Uso. *Instagram*. Disponível em <https://help.instagram.com/581066165581870/?helpref=uf_share>. Acessado em: 25 jul. 2025.

[7]*Instagram*. Perguntas frequentes sobre as Diretrizes da Comunidade do Instagram. *Instagram*, 2018. Disponível em <https://about.instagram.com/pt-br/blog/announcements/instagram-community-guidelines-faqs>. Acessado em: 25 jul. 2025.

Tabela 2.2: Exemplo contendo mensagens expressando ódio implicitamente e explicitamente, com dois tipos de discurso de ódio: racismo e misoginia.

Mensagem	Forma de discurso de ódio	Tipo de discurso de ódio
"a revolução branca é a única solução" [60]	Discurso de ódio implícito	Racismo
"Sua loira vagabunda do caralho morte p vc é pouco..." [61]	Discurso de ódio explícito	Sexismo ou misoginia

2.3.2 *Técnicas de Processamento de Linguagem Natural para Detecção de Discursos de Ódio*

A fim de automatizar o processo de identificação de discursos de ódio, são implementadas técnicas de Processamento de Linguagem Natural (PLN); sub-área do campo de Inteligência Artificial, que trata da compreensão da linguagem humana através do aprendizado de máquina. Os modelos de PLN para detecção de discurso de ódio podem ser treinados a partir de conteúdos extraídos da Internet, como publicações de fóruns online ou de redes sociais. Além da coleta de dados, existem outros estágios igualmente importantes para a detecção de discurso de ódio. Por exemplo, técnicas de pré-processamento, representação vetorial de palavras, entre outras discutidas a seguir. Nesta subseção, serão apresentados os estágios do ciclo de desenvolvimento de um modelo de IA, representados na Figura 2.1, para detecção de discurso de ódio ou linguagem ofensiva.

Coleta de dados para detecção de discurso de ódio

As fontes de dados almejadas para coleta de dados são plataformas de mídias sociais, sites de notícias ou fóruns de discussão online. Além da fonte, pesquisadores estabelecem estratégias para coleta de dados, como parâmetro de busca. Alguns autores escolhem termos considerados potencialmente ofensivos [56] [57], que podem ser encontrados em dicionários como "*HateBase*"[8] e *Swear Word List*[9]. Outrossim, minerar dados de fóruns, como o *Stormfront*[10] [54], ou rastrear usuários e páginas que discutam tópicos propícios à disseminação de discurso de ódio, como esportes [62] e política [9] são estratégias comumente adotadas em pesquisas sobre detecção automática de discurso de ódio. Construir uma base de dados com sentenças criadas por especialistas em discurso de ódio, conforme proposto por Röttger e colaboradores [63], ou incrementar a amostragem para balanceamento entre as classes de dados, conforme realizado por Founta e colaboradores [64], são estratégias pouco utilizadas, pois, pode haver pouca variedade nos dados criados. Também é possível usar bases de dados pré-existentes para compor um novo conjunto de dados, como feito por

[8]HateBase. Disponível em <https://hatebase.org/>. Acessado em: 25 jul. 2025.

[9]*Swear Word List*. Disponível em <https://www.noswearing.com/dictionary>. Acessado em: 25 jul. 2025.

[10]Stormfront era um fórum online que propagava ideologias racistas e de supremacia branca.

Badjatiya e colaboradores [65] e Duwairi e colaboradores [66].

Com a fonte de dados e estratégia para coleta dos dados estabelecidas, são necessárias ferramentas para extrair os dados. A API do X[11] permite definir critérios como filtro anti-spam e idioma das publicações, além de ser possível extrair metadados sobre os usuários. Outras alternativas para mineração de texto são as bibliotecas *BeautifulSoup* e *SNScrape*, disponíveis na linguagem de programação *Python*.

É possível notar que, entre as estratégias definidas para coleta de dados, pode haver aquisição de viés, ao adotar termos considerados potencialmente ofensivos como parâmetro de busca. Segundo Davidson e colaboradores [57], optar pela busca de determinados termos pode ocasionar na discriminação de grupos sociais que os utilizam no seu cotidiano.

Rotulação de conjuntos de dados para detecção de discurso de ódio

A rotulação dos dados é outra fase fundamental para a construção de um conjunto de dados para detecção de discurso de ódio. Nesta etapa, autores optam por realizar essa tarefa sozinhos, como feito por Rosa e colaboradores [61] e Waseem e Hovy [56]; delegar a atividade a especialistas [9]; ou atribuí-la ao trabalho colaborativo terceirizado [64]. Existem plataformas como *Appen* e *Amazon Mechanical Turk*, em que é possível atribuir essa tarefa a outras pessoas. Para autores que preferem realizar a anotação em plataformas de código aberto, sozinhos ou com alguma equipe, existem outras opções, como o *Label Studio*.

Após escolher uma plataforma para rotulação, é preciso definir a estratégia que será usada para confirmação de categoria. Outras classes são definidas para além de "discurso de ódio" e "não discurso de ódio", como as subcategorias de discurso de ódio apresentadas na Subseção 2.3.1, além de outras classes como insulto e spam. Para isso, existem regras que estipulam como serão definidos tais rótulos. Trajano e colaboradores [23], utilizam as seguintes estratégias: voto majoritário, em que o rótulo é definido pela maioria dos rotuladores; todos os rótulos, em que há concordância entre todos os anotadores sobre o mesmo rótulo; e pelo menos um voto, em que apenas um único voto decide o rótulo daquela sentença. Também é preferível que os autores escolham pessoas com experiência ou especialistas para fazer a anotação. Todavia, ainda é possível ocorrerem erros durante esse processo, como a classificação incorreta por via da pouca interpretação do contexto ou da definição de discurso de ódio.

Além disso, é recomendado ponderar aspectos sobre as pessoas que estão anotando os dados, como seu país de origem, idade, nível de educação, experiência com rotulação, ideologia política, gênero e raça, visto que essas particularidades podem influenciar no processo de rotulação, inserindo vieses pessoais nos dados durante essa etapa [16]. Dessa forma, é sinalizado que a diversidade entre os anotadores é um fator que influencia a rotulação [9].

[11] X API. *X Developer Platform*. Disponível em <https://developer.x.com/en/docs/x-api>. Acessado em: 25 jul. 2025.

Sendo assim, existem estratégias adotadas por pesquisadores para verificar a qualidade da rotulação dos dados. Abaixo, está descrito o processo de verificação de confiabilidade entre anotadores.

Ao adotar estratégias como voto majoritário na rotulação de sentenças para detecção de discurso de ódio, os rotuladores inferem o rótulo sobre as mesmas sentenças, como feito por Waseem e Hovy [56] e Trajano e colaboradores [23]. Nesse sentido, Kappa de Cohen é uma métrica estatística utilizada para medir a confiabilidade da categorização nominal entre dois anotadores para o mesmo conjunto de dados [67]. Sendo esta, uma medida de confiabilidade considerada "verdadeira", em que os anotadores concordaram, mesmo excluindo as rotulações em que houve concordância por acaso [68].

A Equação 2.1 mostra como calcular o coeficiente de concordância real K, em que p_o é a proporção de concordância observada, definida na equação 2.2, e p_c é proporção de concordância esperada por acaso, definida na equação 2.3. As equações de proporção de concordância entre anotadores foram adaptadas a partir da Kappa de Cohen com pesos, definida por Fleiss e colaboradores [69]. Em que k é o número de categorias nominais dispostas para aquele conjunto, p_{ii} são as rotulações em que houve concordância entre anotadores, proporcional ao número total de rotulações, e $p_i.p_j$ são todas as rotulações de um anotador para uma categoria, proporcionalmente ao número total de rotulações.

Podemos compreender que: se $K = 1$, então houve total concordância entre os anotadores. Se $K = 0$, então a concordância observada é igual à concordância esperada por acaso. É importante ressaltar que se $K = 0$, não significa que não houve concordância entre anotadores, mas que essa concordância ocorreu por acaso. Além disso, Landis e Koch [4] sugerem interpretações para valores de K entre 0 e 1, em que cada intervalo possui um nível de confiabilidade. Veja a Tabela 2.3.

$$K = \frac{p_o - p_c}{1 - p_c} \tag{2.1}$$

$$p_o = \sum_{i=1}^{k} p_{ii} \tag{2.2}$$

$$p_c = \sum_{i=1}^{k} \sum_{j=1}^{k} p_i \cdot p_j \tag{2.3}$$

Outra medida utilizada para verificar a confiabilidade entre anotadores é a Kappa de Fleiss, uma variação de Kappa de Cohen [69]. Nesta versão, a métrica é utilizada para avaliar a concordância entre a rotulação feita por mais de dois rotuladores. Essa estatística é implementada de mesmo modo que a anterior, visto na equação 2.1, mas as fórmulas para calcular as proporções p_o e p_c foram modificadas, conforme mostrado nas equações 2.5 e 2.4, respectivamente.

$$p_o = \frac{1}{N \cdot n \cdot (n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - N \cdot n \right) \tag{2.4}$$

Tabela 2.3: Interpretação sobre os valores de K, seguindo intervalos, de acordo com Landis e Koch [4].

Valor de K	Intensidade da concordância
≤ 0.00	Insuficiente
$0.01 - 0.20$	Pouca
$0.21 - 0.40$	Justa
$0.41 - 0.60$	Moderada
$0.61 - 0.80$	Substancial
$0.81 - 1.00$	Quase perfeita

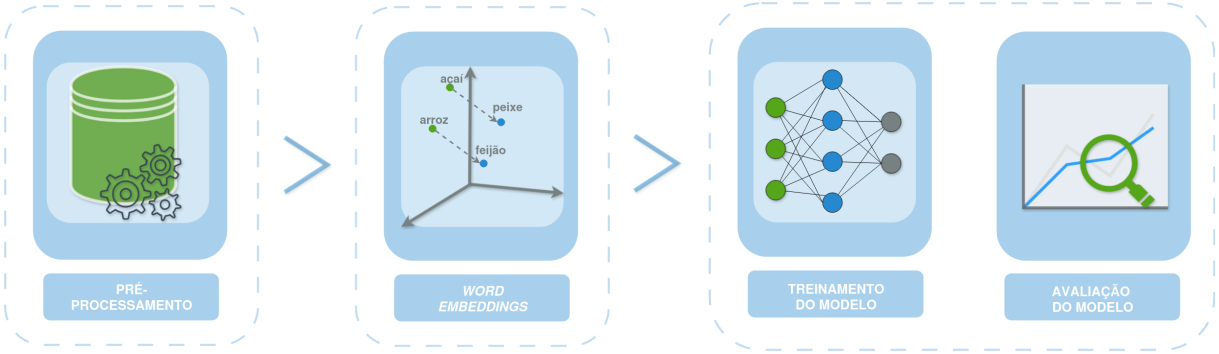


Figura 2.2: Etapas desenvolvidas no treinamento de um modelo para detecção de discurso de ódio

$$p_c = \sum_{i=1}^{k} p_i^2 \tag{2.5}$$

A partir da interpretação do percentual de concordância entre anotadores, é possível ponderar se os responsáveis por essa tarefa possuíam experiência prévia, se entenderam o objetivo da tarefa e o que é discurso de ódio; ou se a definição estipulada agrega diversos tipos de discurso de ódio. De mesmo modo, essas métricas estatísticas são ferramentas úteis para indicar a qualidade da anotação, e consequentemente de uma base de dados.

Treinamento e avaliação de modelos de linguagem natural

A Figura 2.2 ilustra os estágios desempenhados para o treinamento de modelos de IA para detecção de discurso de ódio. Sendo eles: pré-processamento, extração de características com a técnica de *word embeddings*, ou incorporação de palavras, treinamento e avaliação do modelo, com métricas de aprendizado de máquina. Cada fase ilustrada na figura será detalhada a seguir.

O pré-processamento em PLN é realizado a partir de um conjunto de técnicas que deixam o conteúdo das sentenças mais sucinto, descartando alguns caracteres e partes de palavras, a fim de obter ganho durante o processamento. A seguir, são descritas as técnicas de *stemming* e *lemmatization*, usadas para reduzir o tamanho de uma palavra seguindo um conjunto de regras pré-definidas.

- A Estemização ou "*Stemming*" é uma técnica usada para reduzir uma palavra à

sua raiz, eliminando sufixos e prefixos, flexões de gênero, de número, de tempo e de modo dos verbos [70]. A sua execução pode ser feita a partir de abordagens baseadas em regras de linguagem, ou por meio de métodos estatísticos, além de abordagens híbridas [71]. O Removedor de Sufixos da Língua Portuguesa (RSLP) é um *stemmer* desenvolvido a partir da abordagem baseada em regras, capaz de reduzir palavras no plural, palavras no gênero feminino, advérbios, entre outras funcionalidades [72]. O *STEMBR*, desenvolvido especificamente para o português do Brasil, pode realizar reduções de sufixos e prefixos, além de tratar casos especiais, como verbos irregulares [73].

- Similar ao stemming, a Lematização ou "*Lemmatization*", é uma técnica que reduz uma palavra para a sua forma simples, sem conjugações verbais. Diferente da estratégia anterior, a *lemmatization* preserva a classe gramatical das palavras, ajudando a manter o contexto da frase. Esse método também pode ser realizado por meio de abordagens baseadas em regras ou em métodos estatísticos.

Outras etapas são realizadas durante o pré-processamento, sendo elas: as remoções de palavras de parada[12], pontuações, links, emojis, e caracteres especiais; a transformação de letras maiúsculas em letras minúsculas; e a ocultação de nomes de usuários. A remoção de caracteres tem o intuito de diminuir a quantidade de símbolos em um documento, a fim de trazer mais objetividade. Além disso, considerando limitações das capacidades de processamento e armazenamento, é bem quisto que elementos que não agregam importância ao contexto sejam descartados. Em relação à ocultação dos nomes de usuários, estudos sobre detecção de discurso de ódio optam por usar essa abordagem para evitar que os usuários sejam identificados [61] [62].

A Tokenização ou "*Tokenization*" é um processo que consiste em separar os termos de uma sentença, criando representações, chamadas de tokens, que podem identificar partes de palavras, pontuações ou caracteres especiais. Com essa abordagem, uma única palavra pode ser constituída por um ou mais tokens.

O Kit de Ferramentas de Linguagem Natural, "*Natural Language Toolkit*" (NLTK), disponível na linguagem de programação *Python*, pode ser utilizado para realizar as tarefas de pré-processamento supracitadas, entre outras disponíveis em sua biblioteca. Além disso, a ferramenta, por ser de código livre, possui atualizações aprimoradas para a Língua Portuguesa, desenvolvidas por seus usuários [74].

Para ser possível manipular textos durante a etapa de aprendizagem de máquina, eles precisam ser representados por valores numéricos. A Incorporação de Palavras, ou *Word Embeddings*, é uma técnica capaz de codificar tokens para serem representados numericamente. Essa codificação pode ser feita por diferentes métodos, mas todos têm o objetivo de criar representações vetoriais de uma ou mais dimensões, garantindo que seja possível executar operações vetoriais para extrair informações e

[12]Palavras de parada ou "*Stop Words*" são palavras que podem ser retiradas de uma mensagem sem que o contexto seja drasticamente alterado. Por exemplo, "e", "ou" e "um" compõem a lista de palavras de parada do idioma Português.

realizar transformações durante a aprendizagem do modelo. Abaixo, estão descritas diferentes metodologias e técnicas para a criação de incorporação de palavras por meio de um corpus[13] para detecção de discurso de ódio.

- A Sacola de Palavras ou '*Bag of Words*" (*BoW*), é um método capaz de representar todas as palavras de um vocabulário a partir da sua ocorrência. Apesar de simples, essa técnica é usada por modelos estatísticos, como Naïve Bayes, para a tarefa de classificação. Ela consiste em criar um vocabulário, em que serão armazenadas todas as palavras do *corpus*. A partir disso, para cada sentença, é preciso construir um vetor de mesma dimensão do vocabulário e contabilizar a frequência das palavras. Esse método não considera a ordem em que as palavras estão dispostas nas sentenças, nem consegue agregar informações referentes à semântica das palavras. Logo, não é possível medir a correlação entre os termos. Devido a este método contabilizar a frequência de cada palavra do vocabulário, para todas as sentenças, os vetores podem ficar significativamente grandes e preenchidos com vários zeros.

- Outra estratégia de representação é a Frequência de Termos — Frequência Inversa de Documentos, "*Term Frequency - Inverse Document Frequency*" (*TF-IDF*). Através desse método, é possível criar uma representação vetorial de um documento, considerando a relevância de cada palavra com base em todos os documentos do corpus. O primeiro passo é calcular a frequência do termo — TF. Basta contar a frequência que o termo aparece em uma sentença e dividir pela quantidade total de termos da sentença, conforme a Equação 2.6. O segundo passo é calcular a frequência inversa de documentos — IDF. Esse passo envolve computar a quantidade de sentenças, dividida pela quantidade de sentenças que contenham o termo avaliado. Após isso, é necessário calcular o logaritmo do passo anterior, conforme a Equação 2.7. O TF-IDF é obtido por meio da multiplicação entre TF e IDF. A Fórmula 2.8 mostra como calcular o TF-IDF de um termo t em um documento d. Nesse tipo de estratégia, é considerado que, quanto maior for a frequência de um termo, menos importante para o contexto ele é.

$$tf(t,d) = \frac{\textit{frequência de t em d}}{\textit{número de termos em d}} \tag{2.6}$$

$$idf(t) = \log\left(\frac{\textit{número de documentos}}{\textit{número de documentos em que t aparece}}\right) \tag{2.7}$$

$$tf - idf(d,t) = tf(t,d) * idf(t) \tag{2.8}$$

- O algoritmo Word2Vec [76] constrói incorporações de palavras com os pesos obtidos a partir do treinamento de uma rede neural. As representações podem

[13]Para PLN, o *corpus* é um conjunto de dados linguístico sobre determinado tema, que pode ser processado por um computador. Uma coleção de *corpus* é denominado *corpora* [75].

ser distribuídas em um espaço vetorial tal que palavras correlatas obtenham incorporações parecidas. Este modelo possui duas implementações: a Sacola de Palavras Contínua "*Continuous Bag of Words*" (*CBoW*) e o "*Continuous Skip-gram*". No CBoW, as palavras usadas como contexto são submetidas ao modelo para maximizar a probabilidade da palavra alvo ser predita. Cada palavra possui o seu vetor de entrada, codificado com 1, na posição que a palavra aparece no texto, e 0 nas demais posições. Essa representação é denominada "*one-hot-encoded*". Estes vetores de contexto são submetidos a transformações com combinações lineares e funções de ativação. Na camada de projeção, é calculada a média das incorporações. Enquanto na camada de saída, os vetores são submetidos à função *softmax*, responsável por normalizar as probabilidades da palavra alvo. Com isso, o resultado é comparado com a palavra original e os pesos são atualizados, a fim de tornar a predição mais assertiva e minimizar os erros. No *Skip-gram*, o objetivo é inverso ao da Sacola de Palavras Contínua. Cada palavra alvo é submetida ao modelo para que ele consiga maximizar as predições das palavras de contexto, em um determinado intervalo. Por possuir a mesma arquitetura, essa variação também tem as camadas de projeção e saída, em que as probabilidades são geradas e normalizadas, respectivamente. Na Figura 2.3 é possível verificar a arquitetura dos dois modelos.

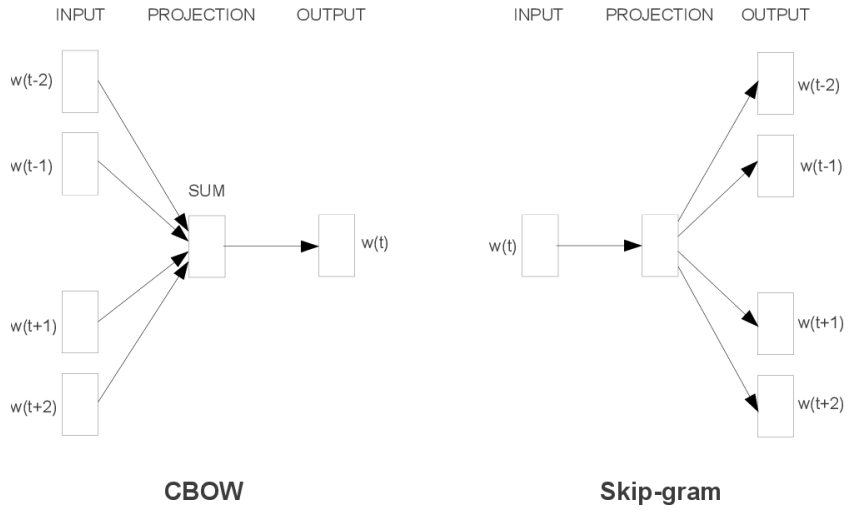


Figura 2.3: Na esquerda, arquitetura do modelo *Continuous Bag of Words*. Na direita, arquitetura do modelo *Continuous Skip-gram*. A entrada para ambos os modelos são representações geradas pelo algoritmo *Bag of Words*. Na saída do modelo CBoW, é emitida a probabilidade da palavra alvo, enquanto na saída do Skip-gram, são emitidas as probabilidades das palavras de contexto, no intervalo pré-estabelecido.

- Em 2017, Vaswani e colaboradores [2] apresentaram uma arquitetura de redes neurais chamada Transformadores, ou '*Transformers*", inicialmente utilizada para tradução de sequências de palavras. Essa arquitetura era capaz de interpretar, com mais precisão, o contexto de uma frase, via mecanismos de auto-atenção implementados em camadas de codificação e decodificação, conforme ilustrado na Figura 2.4. Dessa maneira, surgiu a técnica de incorporação de

palavras contextualizadas, sendo a mais utilizada dentre os modelos que alcançaram o estado da arte, como o BERT [77], em 2019.

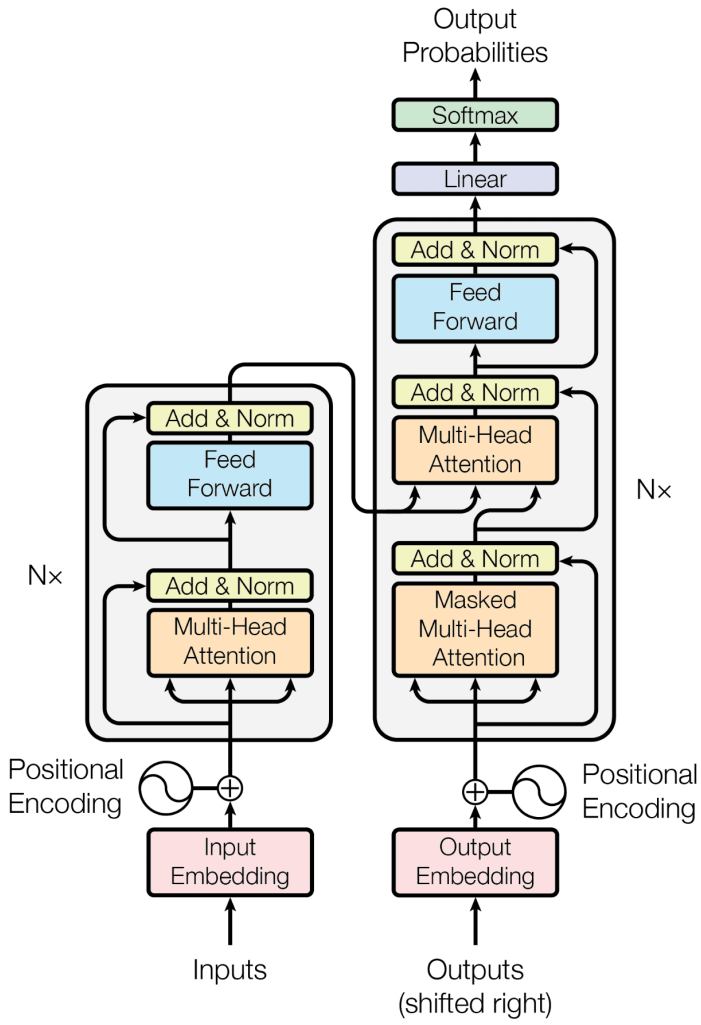


Figura 2.4: Arquitetura de um Transformador proposta por Vaswani e colaboradores [2]. Do lado esquerdo o codificador e do lado direito o decodificador.

Modelos de Aprendizado de Máquina para Classificação de Discursos de Ódio

Desde os mais simples modelos de aprendizado de máquina até os Grandes Modelos de Linguagem (GML) são usados para a detecção de discurso de ódio e linguagem ofensiva através do aprendizado supervisionado[14]. Neste trabalho, alguns modelos foram selecionados para a etapa de treinamento ou ajuste fino. Seguindo as metodologias de Vargas e colaboradores [9] e Rosa e colaboradores [61], os modelos de Regressão Logística (RL), *Multilayer Perceptron* (MLP), BERTimbau$_{base}$ [78] e Tucano-1b1 [79] foram selecionados para realizar a classificação de discurso de ódio.

No algoritmo de Regressão Logística, a função Sigmoid é usada para calcular o rótulo de cada sentença. Para isso, ela utiliza um vetor de recursos computado a

[14]A técnica de aprendizado supervisionado é usada para treinar modelos a partir de um conjunto de dados rotulado.

partir da própria sentença, e parâmetros previamente inicializados, retornando um valor entre 0 e 1, que determina a classe da sentença. Normalmente, o número $0,5$ é usado para delimitar as classes em uma classificação binária. O resultado obtido a partir da função Sigmoid é avaliado pela função de custo, que compara o rótulo predito com o rótulo real. Após isso, é calculado o gradiente, para encontrar o melhor valor para o parâmetro utilizado e minimizar o custo. Quanto menor o custo, mais próximo do rótulo real o rótulo predito está. Durante o treinamento, essas etapas são repetidas por várias iterações, até que o custo mínimo seja alcançado ou até um limite de iterações. Abaixo, na Equação 2.9, é possível ver uma implementação vetorizada da função Sigmoid, em que θ é o vetor de parâmetros e x^i é o vetor de recursos.

$$h(x^i, \theta) = \frac{1}{1 + \epsilon^{-\theta^T x^i}} \tag{2.9}$$

A rede neural de aprendizado profundo do tipo *feed-forward*[15], *Multilayer Perceptron*, funciona a partir da conexão entre neurônios dispostos entre as camadas de entrada, escondidas e de saída. Sendo as camadas escondidas aquelas que ficam entre a camada de entrada e a camada de saída. Em cada camada, os dados de entrada são transformados a partir de funções de ativação, como a sigmoid, descrita anteriormente na Equação 2.9. Segundo Popescu e colaboradores [80], uma rede MLP com apenas uma camada escondida é capaz de gerar regiões de decisão sob a forma de semi-planos. Além disso, ao adicionar mais camadas escondidas, a rede neural gera regiões de decisão convexas, oriundas da intersecção entre os semi-planos gerados pelos neurônios.

A seguir, será explicado o funcionamento dos mecanismos de atenção e autoatenção, provenientes da arquitetura de Transformadores. Sendo esta, a arquitetura utilizada pelos modelos BERTimbau$_{base}$ e Tucano-1b1

Um Transformador é composto por camadas empilhadas de codificadores e decodificadores de tamanho fixo $N = 6$. Em cada codificador, existem duas sub-camadas: a primeira, é uma atenção multi-cabeças; a segunda é uma rede neural do tipo "feed forward", em que os elementos são passados adiante, sem a existência de retro propagação ou laços. Cada uma dessas sub-camadas é seguida de uma função de conexão residual e uma função de normalização. Em cada decodificador existem três sub-camadas: a primeira é uma rede neural do tipo "feed forward"; a segunda, uma atenção multi-cabeças, modificada para que um vetor não dependa da saída dele próprio, recebendo os *embeddings* de saída do codificador, deslocados em uma posição; e a terceira, é uma atenção multi-cabeças, que recebe os *embeddings* de saída do codificador sem alterações.

O mecanismo de atenção deste tipo de arquitetura permite estabelecer dependências entre a entrada e a saída, auxiliando o modelo a criar uma espécie de comparação entre os termos, propiciando a referência entre eles. A Equação 2.10 mostra como é calculada a função de atenção de um Transformador. Seu resultado é obtido através do produto escalar de matrizes Q e K, de dimensão d_k, e V, de dimensão d_v. Em que

[15]Neste tipo de rede neural, não existe retropropagação entre as camadas de neurônios. Ou seja, os dados de entrada seguem à frente até a camada de saída.

o produto escalar de Q e K é multiplicado por um fator $\frac{1}{\sqrt{d_k}}$, e o resultado é aplicado em uma função *softmax*. O resultado final é obtido com o produto escalar entre a matriz V e a saída da função *softmax*. A autoatenção é uma implementação paralela de várias camadas de atenção, que os autores chamam de "multi-head attention", ou "atenção multi-cabeças".

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.10}$$

As incorporações de palavras obtidas através desta arquitetura fazem referências entre si, de modo que a incorporação de um palavra dependa das outras palavras da sentença, permitindo que uma única palavra tenha diferentes representações, dependendo das palavras mais próximas a ela e, por isso, seu contexto é capturado com mais exatidão. Para a detecção de discurso de ódio, esta arquitetura e outras derivadas dela, como o modelo BERT [77], podem ser mais eficientes para identificação de publicações que expressem discurso de ódio, pois, nos casos em que um termo possui o seu significado modificado, dependendo do contexto da frase, o modelo consegue interpretar o seu "novo" significado.

Saleh e colaboradores [81], usaram Representações de Codificadores Bidirecionais de Transformadores (BERT) para detecção de discurso de ódio. Comprovando que o modelo apresenta bons resultados e, dispensa o uso de modelos como Word2Vec para criar incorporações de palavras previamente. O BERT [77] é um grande modelo de linguagem pré-treinado para modelagem de linguagem mascarada e pode ser ajustado para realizar tarefas específicas. Com a implementação de representações de codificadores bidirecionais, advindos da arquitetura de transformadores, o modelo utiliza os mecanismos de autoatenção bidirecionalmente. Isso significa que o modelo consegue processar simultaneamente a entrada, interpretando o contexto da sentença em direções distintas, da esquerda para a direita e da direita para a esquerda. O BERT alcançou o estado da arte em 2019 em tarefas de processamento de linguagem natural e, devido à necessidade de interpretação contextual, ele é um dos modelos mais utilizados para a tarefa de classificação de discurso de ódio. Ainda mais, o BERTimbau [78] é uma variação do modelo BERT, treinado com um corpus na língua portuguesa, disponível com 12 camadas e 110M de parâmetros, na versão BERTimbau$_{base}$ e 24 camadas com 335M de parâmetros, na versão BERTimbau$_{large}$.

O Tucano [79] é um modelo base de código livre pré-treinado com mais de 500B de tokens, que foi treinado a partir da técnica de autoaprendizado. Para seu pré-treinamento, foi utilizada a extensa base de dados GigaVerbo. Este, é um modelo mono lingual para a língua portuguesa, que utiliza apenas o codificador da arquitetura de Transformadores. As versões disponíveis para uso são: Tucano-160m, com 12 camadas, Tucano-630m, com 14 camadas, Tucano-1b1, com 22 camadas, e Tucano 2b4, com 24 camadas.

Considerando que os modelos de grande escala necessitam de muitos recursos ambientais e computacionais para o seu treinamento, uma alternativa interessante seria usar modelos de aprendizado de máquina. Akuma e colaboradores [82] reali-

zaram uma análise comparativa com BoW e TF-IDF para extrair recursos e treinar modelos de aprendizagem de máquina para detecção de discurso de ódio com publicações do *X*. Uma das abordagens com o melhor resultado foi a combinação de BoW e Regressão Logística.

Métricas para Avaliação de Classificadores

Após o treinamento de modelos de aprendizado de máquina, são feitos cálculos com base nas distribuições da predição do modelo sobre o conjunto de teste ou validação. A partir da matriz de confusão, que dispõe os valores preditos na forma de verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN), podem ser computadas uma série de métricas que indicam se houve um bom desempenho ou não do modelo. A seguir, serão descritas as métricas que podem ser utilizadas para avaliação de um classificador binário de discurso de ódio.

- A acurácia é uma métrica que computa o desempenho de um modelo ao classificar corretamente as classes. Conforme a Equação 2.11, a acurácia é calculada dividindo a quantidade total de predições corretas sobre o total de itens da base de dados.

$$acc = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.11}$$

- A métrica de precisão indica a consistência das predições para a classe positiva, realizadas por um modelo em relação ao número total de positivos preditos. A Equação 2.12 mostra como calcular a precisão.

$$precisao = \frac{VP}{VP + FP} \tag{2.12}$$

- A revocação, mostrada na Equação 2.13, é uma métrica usada para avaliar as predições para a classe positiva, realizadas por um modelo em relação ao número total de positivos.

$$revocacao = \frac{VP}{VP + FN} \tag{2.13}$$

- A métrica F1 ou média F1, é calculada a partir da média harmônica entre a precisão e revocação, conforme mostrado na Equação 2.14

$$F1 = 2 * \frac{precisao * revocacao}{precisao + revocacao} \tag{2.14}$$

- A curva característica de operação do receptor (ROC) é usada para avaliar as predições realizadas por um modelo. A métrica AUC, área sob a curva ROC, indica a eficiência do modelo ao distinguir as classes em suas predições.

- As taxas de falsos positivos (TFP) e falsos negativos (TFN) são métricas usadas para avaliar os erros de um modelo em relação à classe de negativos e positivos, respectivamente. Conforme mostrado nas Equações 2.15 e 2.16.

$$TFP = \frac{FP}{VN + FP} \tag{2.15}$$

$$TFN = \frac{FN}{VP + FN} \tag{2.16}$$

2.4 Mitigação e Investigação de Discriminação Algorítmica em Tecnologias para Detecção de Discursos de Ódio

As dificuldades encontradas por pesquisadores, ao categorizar e identificar discursos de ódio, tornam-se mais evidentes diante do uso de plataformas de interação social via Internet. Com a dinâmica frenética das redes sociais, não é incomum surgirem novas formas de discriminação, como o racismo recreativo, que utiliza expressões humorísticas para disseminar preconceitos, sendo veiculado, principalmente, em plataformas de mídias sociais, mediante postagens humorísticas conhecidas como "memes" [83]. Além disso, a linguagem humana sofre com constantes mudanças ao decorrer do tempo, viabilizando a criação ou adaptação de termos para um contexto específico, em que o objetivo seria atacar uma pessoa ou grupo de pessoas [25] [84]. Ainda mais, figuras de linguagem, como ironia e sarcasmo, são frequentemente utilizadas em discursos de ódio, dificultando a classificação automatizada via algoritmos de Inteligência Artificial [85] [86]. Não obstante, alguns termos considerados potencialmente odiosos ou ofensivos, possuem diferentes significados dependendo do contexto em que estão sendo empregados. Por isso, é importante frisar que o uso de uma palavra, isoladamente, não deve ser atrelado automaticamente ao discurso de ódio, pois, isso pode discriminar, através do dialeto utilizado, algum grupo social [87] [16]. Em adição, a complementação entre imagens e frases é um problema ainda maior para a detecção de discurso de ódio, visto que isso exige uma grande compreensão de contexto e, em determinados casos, um conhecimento prévio sobre o que está sendo abordado naquelas imagens e frases, sendo possível identificá-las automaticamente com o uso de técnicas multimodais [88].

Outro fator relevante frente ao discurso de ódio tange à questão de uma definição formal para tal fenômeno [57]. Mesmo que exista um consenso sobre o que é considerado discurso de ódio, muitos trabalhos não assumem diferenças explícitas entre linguagem ofensiva ou tóxica e discurso de ódio [62]. Outros, por exemplo, denotam o discurso de ódio como uma instância de linguagem abusiva, como Trajano e colaboradores [23] e Poletto e colaboradores [89]. Para o aprendizado de máquina, é importante ter uma definição concreta sobre esse assunto, pois os rotuladores irão identificar as mensagens com base nos critérios estipulados sobre o que é considerado

discurso de ódio. Além disso, anotações com pouca concordância podem diminuir as taxas de acerto e, farão, consequentemente, com que os modelos errem mais [9].

Frente a esses problemas, detectar todo e qualquer tipo de discurso de ódio é uma tarefa árdua. E ao ignorar estes desafios, torna factível erros cometidos por algoritmos de aprendizado de máquina. Ainda mais, conforme visto na Seção 2.2, estes erros podem ser danosos para algum grupo específico, qualificando discriminação algorítmica. Além de entender o conceito de discurso de ódio, outra tarefa importante para mitigar o viés codificado é compreender o funcionamento dos sistemas de Inteligência Artificial, e isto abrange conceber quais são as etapas do ciclo de desenvolvimento de modelos para detecção automática de discursos de ódio.

Neste sentido, para entender quais elementos são essenciais para a investigação de racismo algorítmico, foi feito um mapeamento de literatura, com intuito de encontrar trabalhos que utilizem diferentes técnicas e ferramentas para identificar o preconceito de raça em bases de dados para detecção de discurso de ódio, e respectivos modelos treinados com elas.

Através da ferramenta *ResearchRabbit*, foram coletados trabalhos correlatos ao tema de viés na detecção de discurso de ódio que, utilizavam, principalmente, bases de dados na língua inglesa para verificação de discriminação algorítmica. Posteriormente, a coleta fora expandida para encontrar contribuições com bases de dados na língua portuguesa. Ainda mais, os artigos foram filtrados, com a intenção de selecionar apenas os que realizavam identificação de racismo algorítmico em bases de dados para detecção de discurso de ódio ou linguagem ofensiva. A Figura 2.5 mostra um grafo composto por trabalhos sobre investigação de discriminação algorítmica e detecção de discurso de ódio, indicando a crescente correlação entre os temas a partir do ano de 2014.

Como mostrado, a investigação de discriminação algorítmica faz-se presente em pesquisas sobre detecção automática de discurso de ódio. Frente a isso, para investigar a influência dos anotadores na detecção de discurso de ódio, Waseem [24] conduziu um processo de rotulação com anotadores experientes e amadores. O conjunto de dados usado por ele foi composto por uma amostragem dos dados extraídos da base de dados de Waseen e Hovy [56]. Para medir a confiabilidade da rotulação, foi calculado o coeficiente de concordância de Kappa de Cohen entre os anotadores amadores. Ainda que tenha mitigado o viés pessoal, ao fazer uma rotulação mais diversa, o pesquisador salienta não haver diferenças significativas entre a rotulação feita por amadores e a rotulação feita por especialistas.

Mesmo utilizando um conjunto de dados com discursos de ódio direcionados a pessoas com base no seu gênero e raça, Waseem não considerou avaliar viés algorítmico presente nessas bases. Ainda que tenha inferido gênero aos usuários do conjunto de dados, e isso tenha melhorado o coeficiente de concordância entre os anotadores amadores, não foi o suficiente para afirmar que o conjunto de dados propagasse vieses de raça, gênero ou ambos. Necessitando de avaliações mais refinadas.

Em contrapartida, Davidson e colaboradores [29] investigaram viés racial sistêmico em classificadores de discurso de ódio. Usando probabilidade condicional e o

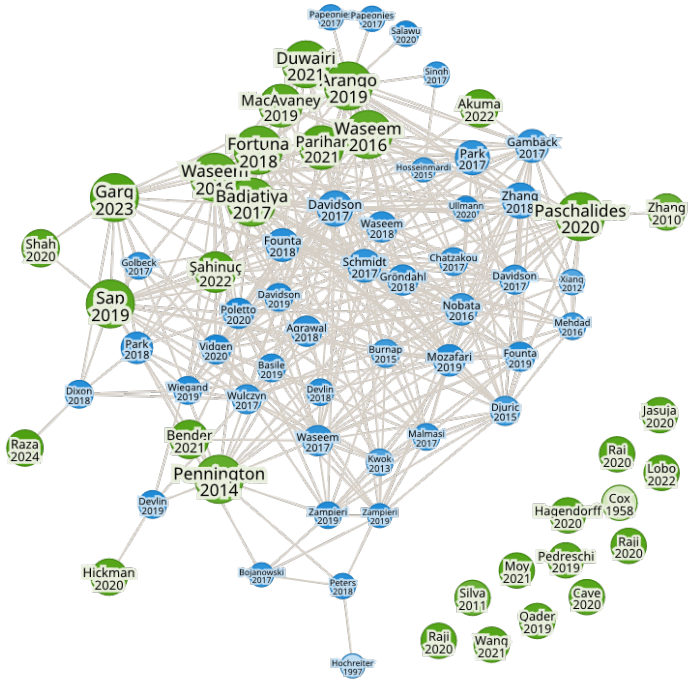


Figura 2.5: A figura mostra um grafo gerado pela ferramenta *Research Rabbit*, com os artigos sobre detecção de discurso de ódio e investigação de discriminação algorítmica, na classificação de discurso de ódio.

método de reamostragem *bootstraping*. Os autores verificaram hipóteses que consistiam em avaliar se, a classificação feita pelos modelos treinados com conjuntos de dados para discurso de ódio, associavam sentenças alinhadas a linguagem usada por minorias com o discurso de ódio. Os resultados deste trabalho mostram disparidades raciais entre o "Inglês Americano Padrão" (*Standard American English - SAE*) e o "Inglês Afro-americano" (*African American English - AAE*). Os autores comprovaram que sentenças no AAE, muito usado por comunidades negras dos Estados Unidos, eram mais prováveis de serem classificadas negativamente, como discurso de ódio ou assédio, do que sentenças no SAE. Para isso, os autores conduziram uma avaliação nos conjuntos de dados de Waseem [24], Waseem e Hovy [56], Davidson e colaboradores [57], Golbeck e colaboradores [90] e Founta e colaboradores [64], retirados do *X*. Um modelo de Regressão Logística com Regularização L2 fora treinado, a partir de incorporações extraídas dos conjuntos de dados com o algoritmo Sacola de Palavras. Cada modelo treinado a partir dos 5 conjuntos de dados fora submetido a um teste usando um novo conjunto de dados, de Blodgett e colaboradores [32][16]. Em um trabalho futuro, Davidson e colaboradores [91] realizaram um estudo para identificar tópicos latentes em conjuntos de dados usando um modelo de tópico estrutural (STM). Demonstrando que tópicos considerados ofensivos estavam associados ao uso de AAE.

[16]Este conjunto de dados é composto por sentenças retiradas de mídias sociais. Ele não possui frases contendo discurso de ódio e tem mais de 50 milhões de exemplos. Ainda mais, é demograficamente vasto, constituído por variações linguísticas de diferentes grupos sociais, incluindo o SAE e o AAE. Por isso, é muito utilizado na investigação de viés racial.

Sap e colaboradores [16] também avaliaram conjuntos de dados de Davidson e colaboradores [57] e Founta e colaboradores [64], para investigar o risco de viés racial. Através do coeficiente de correlação de Pearson, eles mediram a relação entre o uso do dialeto AAE e a presença de discurso de ódio e linguagem ofensiva, utilizando os conjuntos de dados de Blodgett e colaboradores [32] e Preoțiuc-Pietro e Ungar [33]. Para avaliar a propagação de viés sobre os modelos, eles treinaram um classificador e avaliaram as taxas de falsos-positivos (TFP) e falsos-negativos (TFN). 50% das sentenças não ofensivas usando o AAE, eram classificadas incorretamente como ofensivas. Os autores também apresentaram uma metodologia que reduz a probabilidade das sentenças contendo dialetos do AAE serem rotuladas como linguagem ofensiva por anotadores. Com este estudo, os pesquisadores demonstraram que o racismo algorítmico pode ser transmitido pelo viés pessoal de anotadores durante a rotulação de dados, o que influencia o desempenho dos modelos.

Para mitigação de viés racial, Xia e colaboradores [30], desenvolveram uma abordagem que consiste em treinar um modelo adversário para predizer um atributo alvo. Seu intuito era predizer a toxicidade, sem basear essa decisão em atributos protegidos, como os termos usados no dialeto AAE. Para treinar o modelo adversário, uma tupla (x^i, z^i) foi usada. Em que x^i é o texto de entrada e z^i é o rótulo do atributo protegido. Além disso, uma LSTM bidirecional fora usada para a classificação binária. Com isso, os autores conseguiram reduzir a taxa de falsos-positivos para o modelo treinado com o conjunto de dados de Founta e colaboradores [64]. Todavia, o mesmo não foi possível para o modelo treinado com o conjunto de dados de Davidson e colaboradores [57]. Neste trabalho, os autores frisam a importância de reduzir a influência de atributos protegidos durante o treinamento, para que o modelo não generalize os termos como discurso de ódio.

Vargas e colaboradores [92] analisaram a propagação de estereótipos em classificadores para detecção de discurso de ódio, por meio de uma abordagem chamada "Análise de Estereótipo Social" (AES). Com bases de dados no português e inglês, os autores treinaram modelos de aprendizado de máquina e avaliaram vieses com base no gênero e raça. Em seus resultados, foi demonstrado que os classificadores tendem a imputar ofensividade a termos que remetam a grupos minoritários. Assim, reforçando a importância de avaliar a discriminação algorítmica na detecção de discurso de ódio, principalmente em idiomas de baixo recurso, como o português [78].

A Tabela 2.4 reúne os trabalhos relacionados citados nesta seção, com as metodologias e métricas utilizadas para investigar a presença de racismo algorítmico na detecção de discurso de ódio. Nota-se que, dentre as abordagens sugeridas pelos autores, avaliar o viés codificado em bases de dados é uma estratégia importante para atenuação do viés pessoal de anotadores. Além disso, a predição de discurso de ódio sobre uma nova base de dados, alinhada a dialetos de comunidades negras, é uma metodologia comum entre os estudos observados, e traz resultados significativos.

Tabela 2.4: Trabalhos que realizam investigação e mitigação de vieses em bases de dados para detecção de discurso de ódio e suas metodologias implementadas.

#	Autores & ano	Metodologia	Métricas
1	Waseem [24] (2016)	Rotulação com anotadores experientes	Kappa de Cohen
2	Davidson e colaboradores [29] (2019)	Probabilidade condicional e reamostragem *bootstrap*	Teste t
3	Sap e colaboradores [16] (2019)	Estimativa de dialeto e coeficiente de correlação de Pearson	Taxa de falsos positivos e taxa de falsos negativos
4	Davidson e Bhattacharya [91] (2020)	Modelagem de tópicos estrutural e tópicos latentes	Diagnóstico quantitativo
5	Xia e colaboradores [30] (2020)	Atributo alvo e modelo adversário	Métrica F1 e taxa de falsos positivos
6	Vargas e colaboradores [92] (2023)	Análise de estereótipo social	Métrica F1, revocação, precisão e acurácia

2.5 Considerações Finais

A detecção de discurso de ódio é essencial para a manutenção de ambientes amigáveis em plataformas de interação social *online*. Todavia, as ferramentas criadas para este fim podem adquirir os preconceitos inseridos nos dados em diversas etapas do desenvolvimento de sistemas de Inteligência Artificial, como no processo de rotulação [24]. Dessa maneira, acometendo grupos vulneráveis à Discriminação Algorítmica, viabilizando que sejam novamente marginalizados e excluídos.

Em razão dos impactos que podem ser causados por sistemas inteligentes, é imprescindível avaliar quais modelos e combinações de técnicas podem contribuir para a diminuição da perpetuação de estereótipos via tecnologias digitais. Nesse sentido, visto que o Brasil é um país constituído em 55,5% por pessoas negras[17], segundo o Censo Demográfico de 2022, divulgado pelo IBGE[18], é notório que tais ferramentas necessitem de uma avaliação sobre a possibilidade de racismo algorítmico presente na detecção automatizada de discurso de ódio.

Dessa forma, avaliar o viés humano codificado presente nas bases de dados para discurso de ódio, disponíveis no português do Brasil, é uma contribuição importante para a área de Processamento de Linguagem Natural, considerando que ainda existem poucas discussões acerca deste tema [92]. Em adição ao debate sobre Racismo Algorítmico, este trabalho também visa contribuir para o desenvolvimento mais responsável de ferramentas de Inteligência Artificial.

[17] No Brasil, o conceito de "negro" é a união entre pessoas pretas e pardas.

[18] IBGE. Panorama do Censo de 2022. IBGE, 2022. Disponível em <https://censo2022.ibge.gov.br/panorama/>. Acessado em: 26 jul. 2025.

CAPÍTULO

3

Racismo Algorítmico em Tecnologias para Detecção de Discursos de Ódio no Português Brasileiro

3.1 Considerações Iniciais

Este capítulo está organizado entre as Seções 3.2, 3.3 e 3.4. Sendo a primeira delas sobre bases de dados para detecção de discurso de ódio no português do Brasil. Enquanto a segunda, apresenta o conjunto de dados para predição de discurso de ódio, BR-RAPData. Por fim, a terceira seção apresenta uma metodologia para investigação de racismo algorítmico em classificadores treinados para detecção de discurso de ódio.

3.2 Bases de Dados em Português Brasileiro para Classificação de Discursos de Ódio

Conforme discutido na Seção 2.2, os preconceitos podem ser inseridos nas tecnologias digitais em diversas etapas do seu desenvolvimento. Para realizar a detecção automática de discursos de ódio, é necessário treinar ou ajustar um modelo a partir de uma base de dados construída para esta finalidade.

Para encontrar bases de dados no idioma português do Brasil, foi conduzida uma revisão de literatura sistemática, com auxílio da ferramenta *Parsifal*. Sendo projetada em 4 estágios, a revisão abrange as etapas de elaboração, planejamento, condução e relatório, conforme a Figura 3.1.

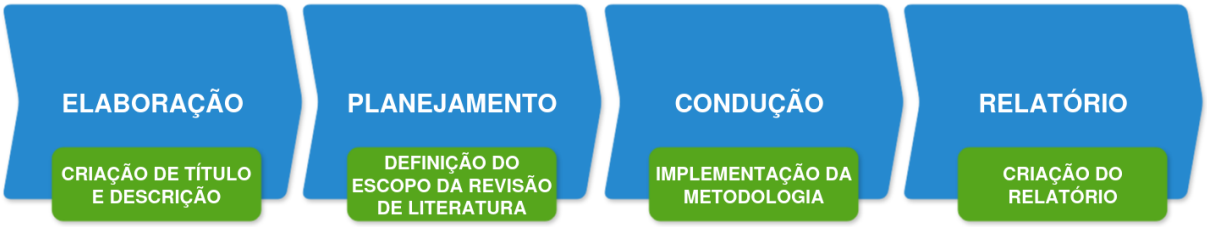


Figura 3.1: A figura ilustra as etapas desenvolvidas durante a revisão de literatura sistemática para encontrar trabalhos que propuseram bases de dados, no idioma português do Brasil, para detecção de discurso de ódio.

3.2.1 Elaboração e Planejamento

Na etapa de Elaboração, foram definidos o título da revisão de literatura e sua descrição. Ambos denominados como “Pesquisa qualitativa sobre bases de dados em português do Brasil para detecção de discurso de ódio”. Na etapa de Planejamento, foram definidos os objetivos da revisão, bem como as variáveis para o método PICOC (População, Intervenção, Comparação, Desfecho ou *Outcome* e Contexto). Foi definido que o objetivo da revisão seria conforme descrito abaixo:

- Encontrar trabalhos que proponham bases de dados para detecção de discurso de ódio em português, para avaliá-los qualitativamente a partir da qualidade da anotação, quantidade de anotadores, experiência dos anotadores, quantidade de sentenças, fonte de coleta de dados e tipos de discriminação abordados na base de dados.

Apenas as variáveis de População, Comparação e Desfecho foram utilizadas nesta revisão. Sendo os valores “base de dados em português”, “modelos de linguagem natural” e “detecção de discurso de ódio”, atribuídos a cada uma delas, respectivamente. Ainda mais, foram definidas as questões de pesquisa abaixo, também na etapa de Planejamento.

Q1 Existe correlação entre a qualidade dos dados e racismo algorítmico?

Q2 Quais características dos conjuntos de dados podem estar relacionadas ao melhor desempenho dos modelos?

As palavras-chave, obtidas através do método PICOC, foram: detecção de discurso de ódio; modelos de linguagem natural; bases de dados em português. Dessa forma, uma sequência de busca foi estabelecida, como mostrado a seguir.

("portuguese database"OR "portuguese corpora"OR "portuguese corpus"OR "portuguese dataset") AND ("language models"OR "large language model"OR "llm"OR "natural language processing"OR "nlp") AND ("hate speech detection"OR "abusive comments"OR "abusive language"OR "hate speech"OR "offensive comments"OR "offensive language")

Por fim, foram definidos os critérios de inclusão e exclusão, o questionário para verificação de qualidade e o questionário para extração de dados.

3.2.2 Condução e Relatório

A sequência de busca, pré-definida, fora utilizada para coletar documentos no *Google* acadêmico. Nesta etapa, 98 documentos foram pré-selecionados. Após ler os títulos e resumos, 25 trabalhos corresponderam aos critérios de inclusão estabelecidos. Após a leitura do artigo completo, o número de trabalhos que não satisfizeram os critérios de exclusão foi igual a 11. Com os questionários de qualidade e extração de dados, foi criado o relatório da revisão e exportado da ferramenta *Parsifal* no formato *.docx*. A Figura 3.2 ilustra os estágios das etapas de Condução e Relatório da revisão de literatura sistemática.

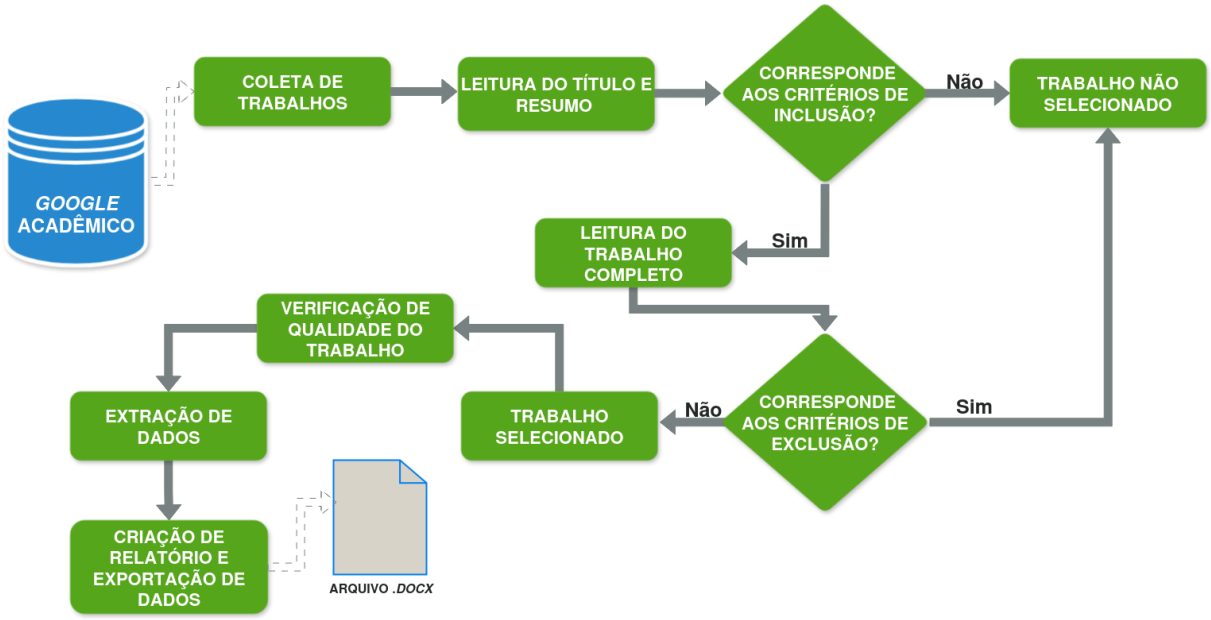


Figura 3.2: A figura ilustra as etapas desenvolvidas durante as fases de condução e relatório da revisão de literatura sistemática.

As bases de dados resultantes do processo de revisão de literatura sistemática estão dispostas na Tabela 3.1, que mostra 11 artigos publicados, cuja principal contribuição fora a criação de uma base de dados para detecção de discurso de ódio, na língua portuguesa do Brasil. Também é possível verificar informações, dispostas pelos autores de cada trabalho, sobre confiabilidade entre anotadores, experiência entre anotadores, quantidade de anotadores e grau de diversidade entre os anotadores. Na Figura 3.3, é possível ver o gráfico gerado pela ferramenta *Parsifal*, mostrando a quantidade de artigos publicados por ano, conforme os trabalhos coletados na revisão de literatura. A partir do ano de 2017, tem-se um aumento significativo na quantidade de trabalhos publicados sobre o tema de discurso de ódio em português. Tal fenômeno deve-se ao fato de que muitos destes trabalhos realizaram a coleta de seus dados entre os anos de 2018 a 2022, como Rosa e colaboradores [61] e Trajano e colaboradores [23]. Neste período, a especulação sobre as eleições presidenciais no Brasil, em plataformas de mídias sociais, ganharam mais notoriedade.

Dentre as bases de dados observadas, apenas “Hate Speech Detection Dataset” e “DOP” não possuem informações sobre os rotuladores. “OFFCOMBR-2” e “HateBR” fa-

Tabela 3.1: A tabela mostra os autores e o ano em que o trabalho foi publicado, junto com o nome das bases de dados, e métricas para avaliação dessas bases de dados. Em que *K* é o percentual de concordância entre anotadores, *QA* a quantidade de anotadores, *GDA* o grau de diversidade entre anotadores, e *E* indica se os anotadores possuem experiência ou não com a rotulação de discursos de ódio.

#	Autores & Ano	Base de Dados	*K*	*QA*	*GDA*	E
1	Pelle e Moreira [8] (2017)	OFFCOMBR-2	0.71	3	Gênero e raça	Não
2	Nascimento e colaboradores [93] (2019)	Hate Speech Detection Dataset				
3	Leite e colaboradores [62] (2020)	ToLD-Br		42	Gênero, orientação sexual e raça	Não
4	Vargas e colaboradores [9] (2021)	HateBR	0.74	3	Raça e regionalidade	Sim
5	Plath e colaboradores [55] (2022)	MINA-BR				Não
6	Guide [94] (2022)	DOP				
7	Oliveira e colaboradores [10] (2023)	TuPy-E		9	Gênero, raça, orientação política	Sim
8	Rosa e colaboradores [61] (2023)	TwitterHateBR		1		Não
9	Trajano e colaboradores [23] (2024)	OLID-BR	0.17		Gênero, faixa etária e nível educacional	Sim
10	Lima e colaboradores [95] (2024)	HEDOS	0.21	3	Gênero, raça e regionalidade	Não
11	Salles e colaboradores [96] (2024)	HateBRXplain		2	Raça	Sim

zem avaliação de concordância entre anotadores, seguindo a métrica Kappa de Fleiss, vista na Subseção 2.3.2, com resultado substancial, conforme a Tabela 2.3. A base de dados "HEDOS", no entanto, obtém resultado justo. Apesar de obter um índice de concordância entre anotadores, a base de dados "OLID-BR" utiliza a métrica Alpha de Krippendorff e, por isso, não é possível emitir uma intensidade para a concordância entre os anotadores deste respectivo conjunto de dados, nem comparar o seu resultado com as métricas Kappa de Fleiss ou Kappa de Cohen. Devido à sua abordagem para rotulação de dados, "ToLD-Br" e "TuPy-E" não usaram o índice de concordância entre anotadores em seus trabalhos. Todavia, estes são os conjuntos de dados com maiores quantidades de anotadores e diversidade entre anotadores. "MINA-BR" e "TwitterHateBR" apresentam apenas poucas informações sobre os rotuladores. Por fim, "HateBRXplain" é a versão explicável do conjunto de dados "HateBR", em que os autores disponibilizaram a justificativa feita por anotadores para cada sentença rotulada como ofensiva.

Os trabalhos de Vargas e colaboradores [9], Guide [94], Oliveira e colaboradores [10], Rosa e colaboradores [61] e Trajano e colaboradores [23], abordam a aquisição do viés humano durante a rotulação dos dados, demonstrando relevância sobre o tema. Os autores de HateBRXplain [96] apresentam um conjunto de dados explicável para detecção de discurso de ódio, além de métodos para avaliação de explicabilidade para classificadores pré-treinados. No entanto, não há comparação entre o desem-

penho de modelos treinados a partir dos conjuntos de dados citados acima, considerando a qualidade dos conjuntos, além da relação deles com racismo algorítmico na detecção de discurso de ódio. Logo, ainda não é possível afirmar se há correlação entre a qualidade do processo de rotulação e a presença de viés.

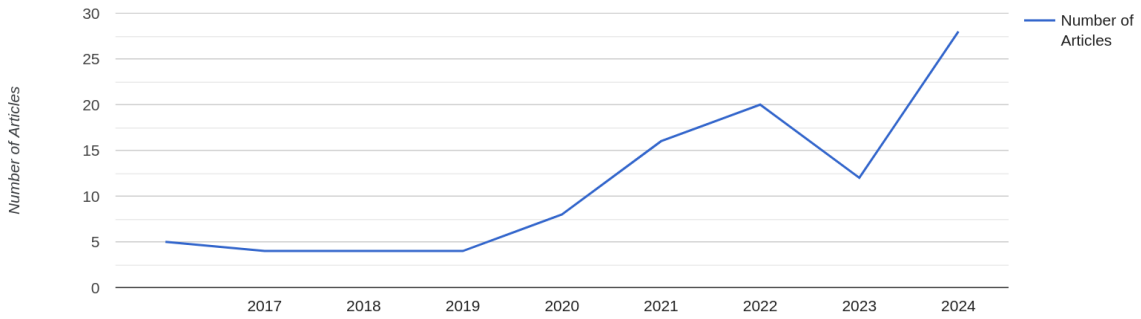


Figura 3.3: Gráfico dos artigos coletados da base de dados *Google* acadêmico, publicados por ano.

3.3 BR-RAPData: Conjunto de Dados para Investigação de Racismo Algorítmico

Conforme feito por Sap e colaboradores [16], Davidson e colaboradores [29] e Xia e colaboradores [30], uma das formas de avaliar racismo algorítmico em conjuntos de dados para detecção de discurso de ódio e modelos treinados a partir deles, é utilizando um conjunto de dados que contenha dialetos característicos de uma comunidade específica. No caso dos trabalhos citados acima, a comunidade negra estadunidense. A base de dados de Blodgett e colaboradores [32], usada para os experimentos dos trabalhos supracitados, é composta por publicações do *X*, em que as sentenças são classificadas entre AAE - *African American English* (inglês afro-americano), e SAE - *Standard American English* (Inglês Americano Padrão). Para isso, os autores criaram um classificador que realiza uma predição sobre a raça do usuário responsável pela publicação da sentença, com base na localização geográfica e nos dialetos utilizados. De modo análogo, Preoțiuc-Pietro e Ungar [33] desenvolveram uma base de dados com publicações da rede social *X*. As publicações eram somente coletadas dos perfis de usuários que realizaram a autodeclaração a partir de um formulário, disponível pelos autores, sobre autodeclaração racial. Posteriormente, os autores criaram um modelo de predição para os maiores grupos étnicos dos Estados Unidos.

Para realizar os experimentos com classificadores treinados com bases de dados no português do Brasil, seguindo a metodologia citada acima, seria necessário um corpus com identificação racial, conforme apresentado por Blodgett e colaboradores ou Preoțiuc-Pietro e Ungar. No entanto, durante esta pesquisa, não fora encontrado, na literatura, um conjunto de dados com essas características. Além disso, traduzir as bases de dados citadas acima não seria viável, porque os dialetos presentes nelas

são específicos da comunidade negra estadunidense.

Considerando os pontos acima, seria necessário haver um dialeto falado por pessoas negras do Brasil, assim, permitindo a criação de uma base de dados para avaliação de racismo algorítmico linguístico. De acordo com David e Nascimento [31], o português falado no Brasil foi influenciado por línguas Bantu[1] durante o processo de colonização, em que africanos foram escravizados no Brasil. Assim, os autores denominam esse fenômeno conforme abaixo:

> [...] definimos pretuguês um conjunto de práticas que os brasileiros, especialmente os de origem afrikanas, passam a utilizar, selecionando repertórios e usos distintos, para falar o português brasileiro. [31, pg.16]

Seguindo essa premissa, é possível afirmar que existe uma variação no português do Brasil, falada principalmente por pessoas negras, mesmo que não exclusivamente. Dessa forma, foi criado um conjunto de dados com músicas de RAP, chamado BR-RAPData. Esse estilo musical foi escolhido, dentre outros, porque é uma cultura criada por pessoas negras, visando relatar a vivência negra e periférica através da arte.

As músicas foram coletadas a partir da discografia de 5 artistas: Negra Li, Emicida, Racionais Mc's, Tasha e Tracie e Atitude Feminina. Para isso, foi feita uma raspagem de dados no portal de músicas brasileiras "Vagalume", com um roteiro desenvolvido na linguagem de programação *Python*, usando a biblioteca "*BeautifulSoup4*". A Tabela 3.2 mostra quais e quantos álbuns de cada artista foram coletados, além da quantidade de músicas totais por cada artista. A Figura 3.4 mostra a nuvem de palavras obtida a partir do conjunto de dados coletado.

Figura 3.4: A figura mostra a nuvem de palavras extraída da base de dados BR-RAPData.

[1]Bantu é um grupo linguístico que engloba diversos idiomas com origem em grupos étnicos da África Subsaariana, como as línguas Kimbundo e Kikongo.

Tabela 3.2: Quantidade de álbuns e músicas coletados de cada artista.

Artistas	#Álbuns	#Músicas	Álbuns
Tasha e Tracie	1	6	Diretoria
Negra Li	3	24	Tudo de Novo Negra Livre Guerreiro, Guerreira
Racionais Mc's	8	81	Cores & Valores Racionais Mc's 25 Tá na Chuva Nada como um dia após o outro dia Sobrevivendo no Inferno Raio X do Brasil Escolha o Seu Caminho Holocausto Urbano
Emicida	9	115	AmarElo Língua Franca Sobre Crianças, Quadris, Pesadelos e Lições de Casa... O Glorioso Retorno De Quem Nunca Esteve Aqui Criolo & Emicida - Ao Vivo Doozicabraba e a Revolução Silenciosa Emicídio Sua Mina Ouve Meu Rep Tamém Pra Quem Já Mordeu Um Cachorro Por Comida, Até Que Eu Cheguei Longe...
Atitude Feminina	1	4	Rosas

Para adaptar as músicas coletadas em um formato semelhante às publicações de plataformas de mídia social, as músicas foram divididas em pedaços de letras. Cada letra de uma música foi separada entre as quebras de linhas. Esse processo resultou em 1.793 pedaços de letras, compondo o BR-RAPData. Além disso, todas as músicas tiveram seu texto convertido para letras minúsculas. Na Tabela 3.3, é mostrado um exemplo de parte de uma música dividida e convertida para letras minúsculas.

É importante ressaltar que para realização dos experimentos conduzidos neste trabalho, foi considerado que não havia na base de dados BR-RAPData sentenças contendo discurso de ódio ou linguagem ofensiva. Ainda mais, segundo Gebru e colaboradores [26], um modelo pode ter um desempenho pior em ambientes de produção no qual os dados são divergentes dos dados utilizados no seu treinamento, principalmente se, nestes dados, houver perpetuação de discriminações.

Tabela 3.3: Exemplo de uma música dividida em pedaços de letras e convertida para letras minúsculas.

Letra sem modificações	Pedaço de letra
Negro drama, Tenta vê, E não vê nada, A não ser uma estrela Longe meio ofuscada	*pedaço 1*: negro drama, tenta vê e não vê nada, a não ser uma estrela longe meio ofuscada
Sente o drama, O preço, a cobrança, No amor, no ódio, A insana vingança	*pedaço 2*: sente o drama, o preço, a cobrança, no amor, no ódio, a insana vingança

3.4 Avaliação de Racismo Algorítmico na Classificação de Discurso de Ódio

3.4.1 Metodologia

Conforme visto na Seção 2.2, os experimentos para verificação de viés seguem estágios bem definidos, que vão desde a coleta de dados até os testes probabilísticos para determinação de discriminação ou não. Sendo assim, as etapas realizadas para concretização de investigação de viés, neste trabalho, têm como referência as metodologias descritas por Davidson e colaboradores [29], Sap e colaboradores [16] e Vargas e colaboradores [92]. A Figura 3.5 mostra os estágios realizados para a investigação de racismo algorítmico, que está presente em bases de dados para detecção de discurso de ódio.

3.4.2 Experimentos

Seleção das bases de dados

Os conjuntos de dados HateBR [9], com 7 mil publicações, TuPy-E [10], com mais de 40 mil publicações e OFFCOMBR-2 [62][2], com 1250 publicações, foram usados para treinar os modelos de classificação de discurso de ódio e linguagem ofensiva. Conforme visto na Seção 3.2, HateBR e OFFCOMBR-2 obtiveram um percentual de concordância entre anotadores substancial. Ainda mais, o conjunto de dados HateBR foi anotado por pessoas experientes, além de ser possível usá-lo, tanto para classificação de linguagem ofensiva, quanto para classificação de discursos de ódio. Além dos dados coletados a partir do *X*, a base de dados TuPy-E oferece uma combinação das bases de dados de Vargas e colaboradores [9], Fortuna e colaboradores [3] e Leite e colaboradores [62]. Sendo assim, as bases de dados HateBR e OFFCOMBR-2 foram

[2]O conjunto de dados de Leite e colaboradores possui uma variação chamada OFFCOMBR-3, em que os dados foram rotulados como ofensivos apenas se houvesse concordância entre todos os anotadores.

Figura 3.5: Metodologia desenvolvida para investigação de racismo algorítmico na classificação de discurso de ódio.

escolhidas em decorrência do seu nível de concordância entre anotadores, enquanto TuPy-E fora escolhida devido a sua quantidade de dados disponíveis.

Pré-processamento

O pré-processamento fora feito apenas para extração de incorporação de palavras com técnicas estáticas. Nesta etapa, foram utilizados os modelos *TFIDFVectorizer*, da biblioteca *scikit-learn*, e Sacola de Palavras Contínua (CBoW), da biblioteca *gensim*, ambas disponíveis na linguagem de programação *Python*. Modelos mais sofisticados, com base na arquitetura de transformadores, implementam técnicas de incorporação de palavras contextualizadas, permitindo maior compreensão sobre os dados. Sendo assim, não foi realizado nenhum processamento prévio nos dados para extração de características.

O pré-processamento consistiu em transformar as sentenças para letras minúsculas, as menções de usuários, caso houvessem, foram removidas, bem como as URLs e palavras de parada[3]. As sentenças foram tokenizadas e tiveram suas respectivas

[3]Palavras de parada ou "Stop Words" são palavras que podem ser retiradas de uma mensagem sem que o contexto seja drasticamente alterado. Por exemplo, "e", "ou" e "um" compõem a lista de palavras de parada do idioma Português.

representações vetoriais extraídas com as técnicas *Continuous Bag of Words* (CBoW) e TF-IDF. Para ambos os treinamentos, os conjuntos de vetores de palavras foram divididos em 80% para treino e 20% para teste. Além disso, para lidar com o desbalanceamento entre as classes dos conjuntos de dados OFFCOMBR-2 e TuPy-E, foram usadas as técnicas de subamostragem e sobreamostragem: *TomekLinks* e Técnica de Sobreamostragem Minoritária Sintética (SMOTE), ambas disponíveis na biblioteca *imbalanced learn*.

Seleção de modelos e treinamento

Seguindo as metodologias de Sap e colaboradores [16], Vargas e colaboradores [92] e Davidson e colaboradores [29], os modelos escolhidos para a classificação de discurso de ódio foram: Regressão Logística (RL), *Multilayer Perceptron* (MLP), BERTimbau base [78] e Tucano-1b1 [79]. Outras técnicas não foram utilizadas, pois, segundo as literaturas usadas como referência, bem como os trabalhos de HateBR [9], OFFCOMBR-2 [8] e TuPy-E [10], estes modelos demonstram bons resultados quanto à classificação de discurso de ódio. Além disso, os modelos BERTimbau$_{base}$ e Tucano-1b1, são GMLs pré-treinados com bases de dados no português, podendo aumentar o desempenho para essa tarefa, que visa avaliar bases de dados na língua portuguesa do Brasil. Os modelos de RL e MLP foram treinados com os vetores de palavras extraídos com CBoW e TF-IDF, de cada conjunto de dados. Devido à quantidade de sentenças presentes em cada base de dados, o modelo de RL fora implementado com o solucionador "*liblinear*" e penalização L1. A rede neural MLP foi treinada com apenas uma camada escondida, seguindo a metodologia de Vargas e colaboradores [9]. A função de ativação padrão "*relu*" foi escolhida, bem como o solucionador padrão, "*adam*". Para os modelos BERTimbau base e Tucano-1b1, foi feito o ajuste fino, uma técnica de aprendizado por transferência, que consiste em realizar um treinamento com uma base de dados menor, para uma tarefa específica. Neste caso, a detecção de discurso de ódio e linguagem ofensiva. A arquitetura proveniente dos Transformadores nestes modelos consegue capturar elementos do texto que contribuem para o contexto, sendo desnecessárias modificações nas bases de dados. Para seus treinamentos, os dados foram divididos em 80% para treino, 10% para teste e 10% para validação. O tamanho de lote escolhido foi 16, o otimizador foi o "*adam*" e taxa de aprendizado foi igual a $2e-4$. Para lidar com o desbalanceamento nas bases de dados OFFCOMBR-2 e TuPy-E, a função de custo foi modificada para considerar pesos diferentes para as classes durante o treinamento, forçando o modelo a ter maior penalidade para erros na classe minoritária. Para o modelo BERTimbau base o número de épocas foi igual a 20, enquanto para o modelo Tucano-1b1 este número foi igual a 5. Ambos modelos pré-treinados tiveram o treinamento realizado com a Unidade de Processamento Gráfico (GPU) *Tesla* T4, disponível para utilização através do *Google Colab*.

Avaliação de desempenho dos modelos

Com intuito de avaliar o desempenho dos modelos, algumas métricas de aprendizado de máquina foram observadas, sendo elas: a Acurácia, Média F1, Revocação e Área sob a Curva Característica de Operação do Receptor (AUC). Cada modelo treinado com diferentes técnicas de incorporação de palavras e bases de dados teve a matriz de confusão computada e, a partir dela, foram retiradas as métricas, conforme a Tabela 3.4.

Tabela 3.4: Métricas de aprendizado de máquina para os modelos Regressão Logística, Multlayer Perceptron, BERTimbau$_{base}$ e Tucano-1b1, treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E.

	Base de Dados											
	OFFCOMBR-2				HateBR				TuPy-E			
Modelo	f1	acc	recall	auc	f1	acc	recall	auc	f1	acc	recall	auc
TF-IDF+RL	0.67	0.74	0.56	0.73	0.85	0.85	0.81	0.94	0.73	0.75	0.66	0.75
CBoW+RL	0.55	0.57	0.57	0.57	0.77	0.77	0.77	0.87	0.70	0.71	0.68	0.71
TF-IDF+MLP	**0.83**	**0.83**	**0.93**	**0.84**	0.83	0.83	0.81	0.84	**0.74**	**0.76**	**0.69**	**0.77**
CBoW+MLP	0.56	0.58	0.55	0.59	0.80	0.79	0.83	0.80	0.73	0.74	0.69	0.74
BERTimbau$_{base}$	0.63	0.84	0.54	0.86	**0.91**	**0.90**	**0.92**	**0.96**	0.69	0.73	0.65	0.81
Tucano-1b1	0.52	0.71	0.43	0.80	0.82	0.83	0.86	0.92	0.65	0.69	0.62	0.78

Predição de discurso de ódio

Cada modelo treinado realizou a predição de discurso de ódio sobre a base de dados BR-RAPData, sendo 0, para "Não discurso de ódio" e 1 para "Discurso de ódio". As probabilidades emitidas pelos modelos também foram armazenadas com as classificações. Neste cenário, foi considerado que esse conjunto de dados continha apenas sentenças não pertencentes à classe de discurso de ódio ou linguagem ofensiva.

Teste de significância estatística

Nesta etapa, foram feitas avaliações considerando as predições dos modelos de aprendizado de máquina sobre o conjunto de dados BR-RAPData. A métrica Kappa de Fleiss foi utilizada para verificar a concordância entre os classificadores treinados com uma mesma base de dados. Por exemplo, com os rótulos de cada sentença emitidos para o conjunto de dados BR-RAPData, pelos modelos TF-IDF+RL, TF-IDF+MLP, CBoW+RL, CBoW+MLP, BERTimbau$_{base}$ e Tucano-1b1, treinados com a base de dados OFFCOMBR-2, foi feito o cálculo de confiabilidade entre os classificadores. O mesmo se repetiu para as bases de dados HateBR e TuPy-E. O intuito deste experimento era avaliar a concordância entre os classificadores de uma mesma base de dados, similar ao que ocorre entre os anotadores durante a rotulação de uma base de dados.

Ademais, fora realizado o teste de normalidade para verificar se as probabilidades emitidas pelos classificadores de discurso de ódio, para o conjunto de dados BR-RAPData, seguiam uma distribuição normal. Para tal, foram consideradas as hipóteses descritas abaixo:

H_N Os dados seguem distribuição normal

H_A Os dados **não** seguem distribuição normal

O teste para partida da normalidade, disponível na biblioteca *SciPy*, proposto por D'Agostino e Pearson [97], foi aplicado sobre as probabilidades emitidas por cada um dos 18 modelos treinados. O nível de significância $\alpha = 0,05$ fora escolhido como limiar para verificar que, se $p \leq \alpha$, então a Hipótese Nula é rejeitada, do contrário, se $p > \alpha$, então a Hipótese Nula não é rejeitada, pois não há indícios de que os dados não seguem uma distribuição normal.

Devido a não normalidade dos dados, foram aplicados apenas testes de significância não-paramétricos, para analisar as classificações feitas pelos modelos de aprendizado de máquina. O qui-quadrado, conforme a Equação 3.1, é um teste não paramétrico usado para encontrar um valor de dispersão para duas variáveis categóricas nominais, neste caso, discurso de ódio e não discurso de ódio. Sendo o_i a proporção observada em determinada categoria e e_i a proporção esperada sob a mesma categoria. Sendo assim, com a biblioteca *SciPy*, disponível na linguagem de programação *Python*, o valor de χ^2 foi computado sobre cada uma das amostragens de predição, realizadas pelos modelos de aprendizado de máquina, comparando-as entre si. Ou seja, apenas modelos treinados com uma mesma base de dados foram comparados mutualmente, como feito anteriormente no teste de Kappa de Fleiss.

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} \tag{3.1}$$

O resultado do teste foi comparado com um valor de χ^2_c crítico, conforme o grau de liberdade obtido e o nível de significância adotado. Neste caso, 1 e 5%, respectivamente, obtendo-se o valor crítico $\chi^2_c = 3,841$. Para avaliar os resultados do teste, foram consideradas as seguintes hipóteses:

H_N A distribuição observada de sentenças classificadas como discurso de ódio, por modelos treinados com uma mesma base de dados, é semelhante à distribuição esperada

H_A A distribuição observada de sentenças classificadas como discurso de ódio, por modelos treinados com uma mesma base de dados, **não** é semelhante à distribuição esperada

Se $\chi^2 > \chi^2_c$, então H_N é rejeitada e H_A é aceita como verdadeira. Do contrário, H_N é mantida.

Na última etapa dos experimentos para investigação de racismo algorítmico, foi feito o teste de Kruskal-Wallis [98]. Esta é uma estatística alternativa não-paramétrica para a Análise de Variância (ANOVA), generalizada a partir do teste Willcoxon-Mann-Whitney (WMW), e serve para comparar duas ou mais amostras [99]. A Equação 3.2 mostra como calcular a medida de diferença entre os grupos comparados, representada pela letra H.

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \tag{3.2}$$

Em que n_i é o tamanho de cada amostra, $N = \sum n_i$ é o número total de indivíduos e R_i é a soma dos postos[4] em cada amostra. Se o número de indivíduos por amostra for grande, então H tem uma distribuição próxima ao qui-quadrado. Sendo necessário comparar H_{calc} com χ^2_c, com $(k-1)$ graus de liberdade, em que k é a quantidade de grupos. Outrossim, caso haja empate entre os postos em cada amostra, o valor de H é corrigido por um fator de correção, conforme a Equação 3.3.

$$FC = 1 - \frac{CE}{N^3 - N} \tag{3.3}$$

Em que $CE = \sum(t^3 - t)$, sendo t o número de postos empatados, e N o número total de indivíduos em cada posto. Ao dividir H por FC, é obtido o valor de H corrigido, que pode ser usado para testar as hipóteses.

O teste H, ou teste de Kruskal-Wallis, foi computado para as classificações feitas por cada modelo, treinado com diferentes bases de dados. O *Listing* 3.1, escrito na linguagem de programação *Python*, mostra como calcular o valor de H. O algoritmo recebe como entrada da função "$compute\_ri\_ni(df)$" um *dataframe* contendo os postos de cada indivíduo da amostragem. A função '$kruskal\_wallis(ri, ni, r)$" recebe como entrada os valores de R_i, n_i e t, computados na etapa anterior.

```
import pandas as pd
import numpy as np

df = pd.DataFrame(columns=['offcombr', 'hatebr', 'tupy'])
df['offcombr'] = [5, 7, 14, 17, 12, 8.5]
df['hatebr'] = [8.5, 10, 6, 18, 15, 16]
df['tupy'] = [2.0, 3, 1, 13, 11, 4]

def compute_ri_ni(df):
    c_names = df.columns
    ri = []
    ni = []
    t = dict()
    visto = set()

    for i in range(len(c_names)):
        ri.append(np.sum(df[c_names[i]]))
        ni.append(len(df[c_names[i]].loc[df[c_names[i]].apply(lambda x:
x > 0.0)]))

        for number in df[c_names[i]]:
            c = 1
            if number in visto:
                c += 1
                if number in t.keys():
                    t[number] += 1
```

[4]Os postos são atribuídos a partir da ordenação dos valores observados de cada amostra. Caso haja repetição entre os valores observados, o posto atribuído é a média entre o valor dos postos.

```
                    else:
                        t[number] = c
                else: visto.add(number)
        return ri,ni,t

    def kruskal_wallis(ri,ni,t):
        N = np.sum(ni)
        H = 12/(N*(N+1))*(np.sum(np.square(ri)/ni)) - 3*(N+1)
        FC = 1
        if len(t) > 0:
            t_values_list = list(t.values())
            CE = np.sum(np.power(t_values_list, 3) - t_values_list)
            FC = 1 - CE/(np.power(N, 3) - N)
            H = H/FC
        return H
    ri,ni,t = compute_ri_ni(df)
    H = kruskal_wallis(ri,ni,t)
    X2 = 5.991
    if H > X2:
        print("A Hipotese nula pode ser rejeitada. Hipotese alternativa e
    aceita.")
    else:
        print("A hipotese nula NAO pode ser rejeitada.")
```

Listing 3.1: Código em python para calcular o valor de H através da fórmula de Kruskal-Wallis.

A Tabela 3.5 mostra os postos e as proporções de discurso de ódio para a base de dados BR-RAPData, emitidas através dos modelos treinados com OFFCOMBR-2, HateBR e TuPy-E. Através dela, o valor de H foi computado com o código presente no Listing 3.1.

Tabela 3.5: Proporções de sentenças classificadas como discurso de ódio do conjunto de dados BR-RAPData. Na coluna P estão as proporções de discurso de ódio. Enquanto na coluna Posto, estão os postos obtidos através da ordenação das proporções. R_i é a soma dos postos, n_i é o número de postos em cada amostra, e R_i^2/n_i é o quadrado dos postos divido pela soma de postos em cada amostra.

	OFFCOMBR-2		HateBR		TuPy-E	
	P	Posto	P	Posto	P	Posto
TF-IDF+RL	404	5	547	8.5	163	2
TF-IDF+MLP	534	7	638	10	250	3
CBoW+RL	877	14	495	6	104	1
CBoW+MLP	1156	17	1516	18	720	13
BERTimbau$_{base}$	699	12	976	15	681	11
Tucano-1b1	547	8.5	1052	16	297	4
R_i	63.5		73.5		34	
n_i	6		6		6	
R_i^2/n_i	672.04		900.37		192.66	

Com o resultado de H, foi feita a comparação com $\chi^2_c = 5,991$, sendo $\alpha = 0,05$ e grau de liberdade igual a 2, as hipóteses abaixo foram testadas.

H_N A quantidade de sentenças classificadas como discurso de ódio varia entre os modelos treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E.

H_A A quantidade de sentenças classificadas como discurso de ódio **não** varia entre os modelos treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E.

Novamente, se $H > \chi^2_c$, então a hipótese nula é rejeitada. Do contrário, a hipótese nula é aceita.

3.4.3 Resultados

A Figura A.1 do Apêndice A mostra as proporções de cada classificação feita sobre o conjunto de dados BR-RAPData, com os modelos de aprendizado de máquina treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E. É possível notar que, com o avanço das técnicas de incorporação de palavras, inicialmente estáticas, como TF-IDF, e posteriormente contextualizadas, com BERTimbau e Tucano, as proporções para a classe de discurso de ódio ficaram maiores. Também é possível notar um aumento para a proporção de sentenças preditas como discurso de ódio à medida que a complexidade dos modelos também aumenta. De modo geral, as proporções para a classe de discurso de ódio ficaram maiores ao utilizar ambas técnicas de processamento de linguagem natural e aprendizado de máquina, mais complexas. Como os modelos com base na arquitetura de Transformadores: BERTimbau$_{base}$ e Tucano-1b1.

Conforme discutido na seção anterior, foi consentido que os experimentos seriam realizados sob o pressuposto de que não havia discurso de ódio entre as sentenças do conjunto de dados BR-RAPData. Este tópico será debatido com mais detalhes na Seção 4.1. Sendo assim, foram calculadas as taxas de falsos positivos (TFP) para cada um dos modelos, dispostas na Tabela 3.6. As maiores taxas de falsos positivos foram obtidas entre os classificadores treinados com a base de dados HateBR [9] e, consequentemente, estes foram os classificadores que mais rotularam sentenças de BR-RAPData como discurso de ódio. Também é possível notar, através da Figura 3.6, que o modelo CBoW+MLP, alcançou a maior TFP entre todas as bases de dados. Em contrapartida, o modelo TF-IDF+RL alcançou a menor TFP entre os modelos treinados com a base de dados OFFCOMBR-2, enquanto o modelo CBoW+RL alcançou a menor TFP entre os modelos treinados com as bases de dados HateBR e TuPy-E.

Outra métrica usada para avaliar a predição dos modelos sob o conjunto de dados BR-RAPData foi a Kappa de Fleiss. Observou-se que, mesmo entre as bases de dados que alcançaram uma intensidade de concordância entre anotadores substancial, como OFFCOMBR-2 e HateBR, o valor de K, atingido pelos classificadores na rotulação de BR-RAPData, foi abaixo do esperado. Para os classificadores treinados com OFFCOMBR-2, o resultado de K foi igual a 17%. Para os classificadores treinados com HateBR, o resultado de K foi igual a 25%. Enquanto isso, para a base de dados

Tabela 3.6: Taxa de falsos positivos e proporções das "classes discurso de ódio" e "não discurso de ódio" para a base de dados BR-RAPData emitidas por cada modelo treinado com as bases de dados OFFCOMBR-2, HateBR e TuPy-E.

	Bases de dados								
	OFFCOMBR-2			HateBR			TuPy-E		
Modelo	TFP	Ódio	N. Ódio	TFP	Ódio	N. Ódio	TFP	Ódio	N. Ódio
TFI-IDF+RL	0.22	1389	404	**0.30**	1246	547	0.09	163	1630
TF-IDF+MLP	0.29	1259	534	**0.35**	1155	638	0.13	250	1543
CBoW+RL	**0.48**	916	877	0.27	1298	495	0.05	104	1689
CBoW+MLP	0.64	637	1156	**0.84**	277	1516	0.40	720	1073
BERTimbau$_{base}$	0.38	1094	699	**0.54**	817	976	0.37	1112	681
Tucano-1b1	0.30	1246	547	**0.58**	741	1052	0.16	297	1496

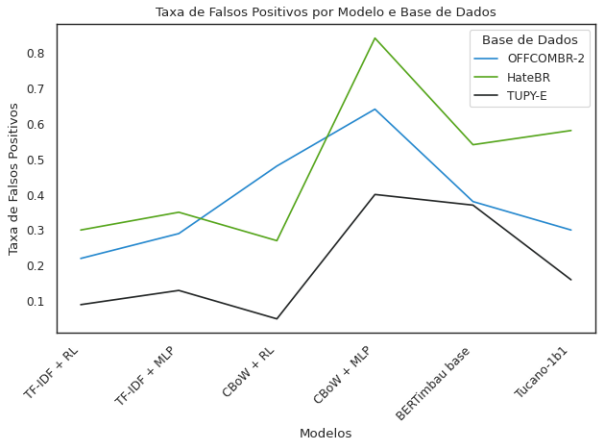
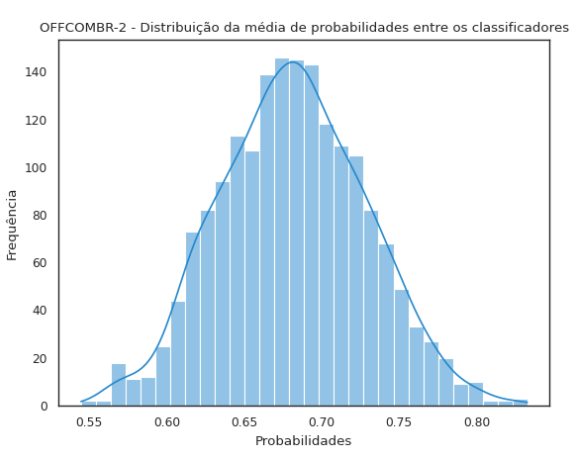


Figura 3.6: Taxas de falsos positivos sobre a base de dados BR-RAPData, obtidas através das predições realizadas pelos modelos treinados com OFFCOMBR-2, HateBR e TuPy-E.

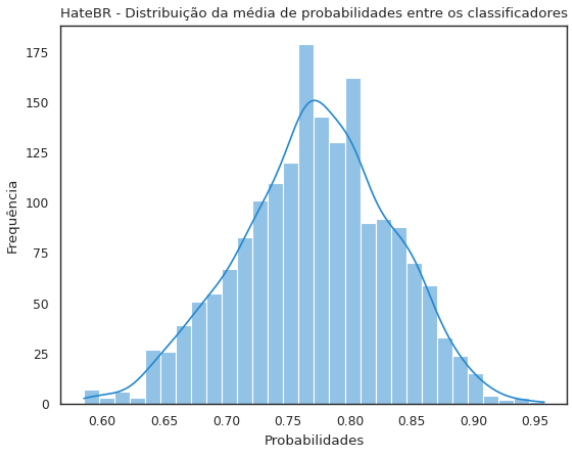
TuPy-E, o valor foi igual a 19%. Todas essas porcentagens estão abaixo da intensidade de concordância aceitável, que seria igual ou superior a 40%, segundo Vargas e colaboradores [9].

Para avaliação sobre as distribuições dos dados, foram feitos testes de normalidade sobre todas as amostras de BR-RAPData rotuladas pelos classificadores. O teste de normalidade também foi realizado sobre as médias de probabilidades, conforme a Figura 3.7. A única distribuição normal, segundo o teste, foi a média de probabilidades dos modelos treinados com o conjunto de dados OFFCOMBR-2. Isso mostra a necessidade de realizar um teste estatístico para normalidade, pois, com apenas o histograma não é possível obter um resultado confiável. No entanto, como resultado do teste para as 18 amostras avaliadas, nenhuma teve um valor de significância acima de 5%, rejeitando-se, portanto, a hipótese nula de que os dados seguem uma distribuição normal. No Apêndice A, as Figuras A.2, A.3 e A.4 mostram as distribuições de probabilidades para cada amostragem, emitidas por modelos treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E, respectivamente.

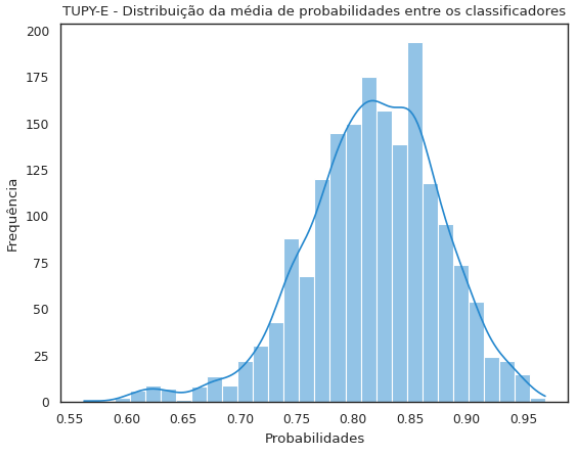
Ainda mais, o teste não-paramétrico qui-quadrado, fora realizado para verificar se as proporções de sentenças classificadas como discurso de ódio, emitidas por modelos



OFFCOMBR-2



HateBR



TuPy-E

Figura 3.7: Distribuição das médias de probabilidades de discurso de ódio, emitidas por classificadores treinados OFFCOMBR-2, HateBR ou TuPy-E, para o conjunto de dados BR-RAPData.

treinados com a mesma base de dados, divergem entre si. Foi constatado que, as distribuições eram semelhantes à distribuição esperada. Ou seja, os valores de χ^2 foram todos menores ou iguais ao valor de $\chi^2_c = 3,841$. Portanto, sendo considerada como verdadeira a hipótese nula. A Tabela A.2 do Apêndice A, mostra os resultados obtidos para X^2, para cada comparação entre modelos treinados com a mesma base de dados.

Com o teste de *Kruskal-Wallis*, foi verificado se havia diferença significativa entre as proporções de discurso de ódio emitida com os modelos treinados via diferentes bases de dados. O valor de $H_c = 4,937$ foi menor que $\chi^2_c = 5,991$. Dessa maneira, mantendo a hipótese nula de que há variação significativa entre as amostras rotuladas com os modelos treinados por diferentes bases de dados.

Dentre os modelos treinados com diferentes técnicas de *word embeddings* e PLN, alguns alcançaram resultados satisfatórios, comparados às literaturas dos conjuntos de dados OFFCOMBR-2 [8], HateBR [9] e TuPy-E [10]. A combinação de vetores de palavras extraídos com a técnica de TF-IDF e o modelo *Multilayer Perceptron* (MLP) atingiu 83% e 74% com a métrica F1, para os conjuntos de dados OFFCOMBR-2 e

TuPy-E, respectivamente. Para o conjunto de dados HateBR, o modelo BERTimbau$_{base}$ alcançou o valor de $F1 = 91\%$. A acurácia e revocação ou *recall* também tiveram bons resultados para estes modelos, bem como a métrica AUC. Essas medidas indicam se o modelo tem um bom desempenho ao detectar discursos de ódio ou não. Por exemplo, quanto maior for a métrica F1 de um modelo, melhores foram os resultados de precisão e revocação, indicando diminuição no número de falsos positivos e falsos negativos. Além disso, a acurácia, que mede a taxa de acertos de um modelo, e a métrica AUC, indicam a certeza de acertos do modelo e sua habilidade em distinguir as classes. É importante ressalter que, para conjuntos de dados desbalanceados, como OFFCOMBR-2 e TuPy-E, a acurácia pode ser aumentada apenas por predizer as sentenças a favor da classe majoritária, o que é incorreto. Por fim, é possível afirmar que estes modelos conseguem identificar discurso de ódio. Assim, respondendo a pergunta 1, da Seção 1.1.

CAPÍTULO

4

Considerações Finais

4.1 Discussão

Conforme discutido na Seção 3.3, David e Nascimento [31] afirmam que o idioma português foi influenciado por línguas de origem Bantu durante os mais de 300 anos de escravização de povos africanos no Brasil. Assim, denominou-se o "pretuguês", termo cunhado pela pesquisadora e socióloga Lélia Gonzalez, como o português falado principalmente por pessoas negras e marginalizadas [100]. Seguindo essa premissa, a metodologia deste trabalho fora aplicada com base nas metodologias de Sap e colaboradores [16], bem como Davidson e colaboradores [29], que consistem em analisar o preconceito racial com base em dialetos da comunidade negra estadunidense. No contexto brasileiro, optou-se por analisar letras de RAP, um estilo musical difundido entre comunidades negras no Brasil. Sendo assim, a Tabela 4.1 mostra algumas partes de letras de músicas, rotuladas como discurso de ódio e não discurso de ódio, por classificadores treinados com os conjuntos de dados OFFCOMBR-2, HateBR e TuPy-E.

Como mostrado na Subseção 3.4.3, todos os classificadores treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E, identificaram sentenças do conjunto de dados BR-RAPData como discurso de ódio. Ainda mais, houve amostragem de predições que mostraram aumento significativo, dependendo das técnicas de incorporação de palavras e modelos de aprendizado de máquina usados. Com o resultado do teste de Kappa de Fleiss, foi possível notar que houvera pouca concordância entre os classificadores treinados com a mesma base de dados, apontando divergências sobre o que deve ser considerado como discurso de ódio ou não, mesmo não havendo uma definição precisa sobre discurso de ódio, de acordo com Vargas e colaboradores [9]. Em contrapartida, Poletto e colaboradores [89] afirmam que, discurso de ódio é uma fração de linguagem ofensiva ou tóxica [23]. Sendo assim, os modelos apresentados

Tabela 4.1: Exemplos de músicas da base de dados BR-RAPData rotuladas como discurso de ódio e não discurso de ódio.

Letra	Rótulo	Modelo	Base de dados	Música	Artista
D-D-D-Drip de negona D-D-D-Drip de negona Nasci com a boca que elas compra Cacho de colombiana Unha de brasileirinha D-D-D-Drip de negona Nasci com a boca que elas compra Cacho de colombiana	Discurso de ódio	TF-IDF+RL	OFFCOMBR-2	Amarrou	Tasha e Tracie
Um grande exercito do rap quando eu olho Os doido vira o zóio vamo virar também Atenção muita fé sem tira ninguém Periferia resiste	Discurso de ódio	BERTimbau$_{base}$	HateBR	Exército do rap	Negra Li
Hoje eu sou ladrão artigo 157 As cachorra me amam Os playboy se derretem Hoje eu sou ladrão artigo 157 A policia bola um plano Sou herói dos pivete	Não discurso de ódio	Tucano-1b1	TuPy-E	Eu sou 157	Racionais *MC's*
Holofotes fortes purpurina E os sorrisos dessas "mina"só me lembram cocaína Em cinco abrem-se cortinas Estáticas retinas brilham garoa fina	Não discurso de ódio	CBoW+MLP	HateBR	Hoje cedo	Emicida
Hoje o meu amor veio me visitar E trouxe rosas para me alegrar E com lágrimas pede pra voltar Hoje o perfume eu não sinto mais O meu amor já não me bate mais Infelizmente eu descanso em paz!	Não discurso de ódio	CBoW+RL	TuPy-E	Rosas	Atitude Feminina
Chega de festejar a desvantagem E permitir que desgastem a nossa imagem Descendente negro atual meu nome é Brown Não sou complexado e tal Apenas Racional É a verdade mais pura Postura definitiva A juventude negra Agora tem voz ativa	Discurso de ódio	TF-IDF+MLP	OFFCOMBR-2	Voz ativa	Racionais *MC's*

na Tabela 3.4 deveriam mostrar rotulações similares, considerando que, neste experimento, foram comparados apenas classificadores treinados com a mesma base de dados. Para Davidson e colaboradores [57], é necessário haver uma limitação sobre a proibição de determinadas linguagens na Internet, para não serem consideradas como discurso de ódio ou linguagem ofensiva. Por exemplo, silenciando usuários que façam citações de letras de RAP, sob a premissa de que seria linguagem ofensiva. Ou em casos em que haja tonalidade de denúncia sobre violações de direitos humanos, como ocorre com usuários que denunciam o genocídio palestino nas redes sociais *Instagram* e *Facebook*[1].

Ainda mais, as taxas de falsos positivos, mostradas na Tabela 3.6, em adição aos resultados da métrica Kappa de Fleiss, evidenciam que: 4 entre os 16 classificadores não foram concisos nas suas decisões ao rotular mais de 50% da base de dados BR-RAPData como discurso de ódio. Logo, respondendo à questão de pesquisa $Q1$, definida Seção 3.2, não é possível afirmar se existe correlação entre a qualidade das bases de dados e racismo algorítmico.

Por mais que os resultados da Tabela 3.4 sejam promissores, visto que pelo menos um modelo treinado com cada base de dados atingiu um valor para métrica F1 acima de 70%, é importante analisar os indícios de discriminação algorítmica. Outrossim, os modelos com melhor desempenho foram treinados com a base de dados HateBR,

[1]BROWN, Deborah. *Meta's Broken Promises Systemic Censorship of Palestine Content on Instagram and Facebook*. *Human Rights Watch*, 2023. Disponível em <https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and>. Acessado em: 26 jul. 2025.

que demonstrou qualidade no sentido de balanceamento entre os dados, anotadores com experiência, diversidade entre anotadores, além de intensidade de confiabilidade substancial. Portanto, respondendo à questão de pesquisa Q2, definida na Seção 3.2, estas são características de um conjunto de dados que podem estar relacionadas ao bom desempenho de um modelo.

Em adição, após a predição de discurso de ódio sobre o conjunto de dados BR-RAPData, foram testadas as seguintes hipóteses:

1. Teste de normalidade

   H_N As probabilidades emitidas pelos modelos de classificação, treinados com OFFCOMBR-2, HateBR e TuPy-E, seguem uma distribuição normal.

   H_A As probabilidades emitidas pelos modelos de classificação, treinados com OFFCOMBR-2, HateBR e TuPy-E, **não** seguem uma distribuição normal.

2. Teste de qui-quadrado

   H_N A distribuição observada de sentenças classificadas como discurso de ódio, por modelos treinados com uma mesma base de dados, é semelhante à distribuição esperada.

   H_A A distribuição observada de sentenças classificadas como discurso de ódio, por modelos treinados com uma mesma base de dados, **não** é semelhante à distribuição esperada.

3. Teste de *Kruskal-Wallis*

   H_N A quantidade de sentenças classificadas como discurso de ódio varia entre os modelos treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E.

   H_A A quantidade de sentenças classificadas como discurso de ódio **não** varia entre os modelos treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E.

Devido a não normalidade dos dados, a hipótese alternativa fora adotada para o teste de normalidade. Enquanto os testes de qui-quadrado e *Kruskal-Wallis*, tiveram as hipóteses nulas adotadas. Estes resultados evidenciam que:

1. Os modelos treinados com uma mesma base de dados, OFFCOMBR-2, HateBR ou TuPy-E, emitiram proporções semelhantes entre si ao detectar discurso de ódio na base de dados BR-RAPData. Portanto, notou-se que houve aquisição de viés com base no pretuguês, para os modelos treinados com as respectivas bases de dados.

2. Os modelos treinados com diferentes bases de dados revelam uma variação nas proporções de discurso de ódio emitidas por eles. Ou seja, mesmo que todos os modelos treinados com OFFCOMBR-2, HateBR e TuPy-E tenham detectado

discurso de ódio no conjunto de dados BR-RAPData, as proporções, emitidas por modelos treinados com diferentes bases de dados, não foram semelhantes entre si.

É importante ressaltar que, segundo Gebru e colaboradores [26], se um modelo for testado no ambiente de produção com dados que não são similares aos seus dados de treinamento, é provável que ele tenha um mau desempenho. Ainda mais, se, nestes conjuntos de treinamento, tenha a presença de viés codificado, podendo ser adquirido durante a etapa de rotulação do conjunto de dados, segundo Sap e colaboradores [16]. Além disso, como discutido na Seção 2.2, a aquisição de discriminação algorítmica também pode ocorrer em outros ciclos de desenvolvimento de um modelo de IA, como no treinamento. Dessa maneira, respondendo à questão 2, definida no Capítulo 1, é possível investigar a discriminação algorítmica propagada pela detecção de discurso de ódio mediante metodologias para verificação de viés, implementadas a partir de predições de discurso de ódio em bases de dados que contenham dialetos de alguma comunidade específica, como a comunidade negra do Brasil, através de testes de significância estatística, para validação de resultados.

Mais uma vez, é ressaltada a importância de avaliar, além do desempenho do modelo, a sua capacidade de amplificar discriminações codificadas, como o racismo algorítmico. Portanto, adotar práticas para avaliação de viés em modelos e bases de dados contribui para que as ferramentas de IA sejam desenvolvidas fundamentadas na transparência, responsabilização, explicabilidade e justiça.

4.1.1 Dificuldades encontradas

Durante o desenvolvimento deste trabalho, algumas limitações foram encontradas e serão discutidas a seguir.

Desbalanceamento entre as classes de discurso de ódio

Para realizar a classificação de discurso de ódio automatizada, os modelos são treinados a partir de conjuntos de dados construídos através da coleta de textos via plataformas de interação social na Internet. Mesmo em ambientes considerados como prováveis de se encontrar discursos de ódio ou linguagem ofensiva, por exemplo, em redes sociais, a quantidade de publicações não ofensivas é muito maior que a quantidade de publicações ofensivas. Este fenômeno gera desbalanceamento entre as classes, como nas bases de dados OFFCOMBR-2 [8] e TuPy-E [10], fazendo com que os modelos classifiquem mais sentenças como não discurso de ódio, o que, por sua vez, aumenta a taxa de falsos negativos.

Base de dados demográfica

A metodologia proposta por Davidson e colaboradores [29] realiza a investigação de racismo algorítmico através da base de dados de Blodgett e colaboradores [32], que

possui sentenças escritas no AAE e SAE. No entanto, não foi encontrada, na literatura, uma base de dados no idioma português do Brasil que apresentasse as mesmas características. Sendo assim, foi optado por utilizar letras de RAP, um estilo musical predominante entre as periferias e logo, pelos falantes de pretuguês. No entanto, Davidson e colaboradores [57] explicam que, uma das limitações da detecção de discurso de ódio é a censura de termos considerados ofensivos, o que não é incomum de se encontrar em letras de RAP. Dessa maneira, seguindo a definição de linguagem ofensiva ou tóxica de Leite e colaboradores [62], algumas partes de letras de RAP podem de fato ser consideradas como ofensivas.

Idioma com poucos recursos

O desempenho de modelos durante o aprendizado por transferência pode ser impactado pela falta de dados em diferentes línguas. Isso ocorre com idiomas com poucos recursos, que são sub-representados nos conjuntos de dados de treinamento de modelos de linguagem, como o português [78]. Segundo Fortuna e colaboradores [3], a maioria das bases de dados encontradas na literatura para detecção de discurso de ódio estão no idioma inglês, fazendo-se necessária não somente a construção de novas bases de dados em português, mas também de abordagens para avaliação de discriminação algorítmica neste idioma.

Limitação de recursos

A partir da arquitetura de Transformadores [2], os modelos de IA passaram a ser treinados com milhares, e atualmente bilhares de parâmetros. Devido ao seu treinamento, com volumosas bases de dados, estas tecnologias precisam de equipamentos com grande poder de processamento, como Unidades de Processamento Gráfico (GPU) e Unidades de Processamento de Tensor (TPU), limitando vários recursos como dinheiro, tempo, água e energia [101].

4.2 Conclusão e Trabalhos Futuros

Neste trabalho, foi proposta uma metodologia para investigação de racismo algorítmico, organizada em onze estágios: seleção das bases de dados; pré-processamento; seleção de modelos; treinamento de modelos; avaliação de desempenho; predição de discurso de ódio; métrica Kappa de Fleiss; teste de significância estatística; e apresentação de resultados. As bases de dados foram escolhidas com base na intensidade de concordância entre anotadores, emitida pela métrica Kappa de Fleiss, a quantidade de anotadores, o grau de diversidade entre anotadores e nível de experiência entre anotadores, visando conjuntos de dados em que a diminuição de viés fosse pautada. Também foram selecionados diferentes modelos e técnicas de processamento de linguagem natural para verificar as suas capacidades de detectar automaticamente discursos de ódio. Ainda mais, com os modelos treinados, foi avaliada a presença de discriminação algorítmica através da predição de discurso de ódio sobre a base

Tabela 4.2: Tabela com as principais contribuições deste trabalho

Contribuições	Resumo
Metodologia para investigação de racismo algorítmico	Esta metodologia tem como base os trabalhos de [29] [16] [30], e realiza investigação de racismo algorítmico no português do Brasil.
Base de dados BR-RAPData	Base de daos alinhada racialmente, através do dialeto pretuguês [31]
Modelos para detecção de discurso de ódio	18 modelos para detecção de discurso de ódio, treinados com diferentes bases de dados e técnicas de representação de palavras.
Revisão de literatura sistemática	RLS desenvolvida para encontrar trabalhos que contribuíssem com conjuntos de dados para detecção de discurso de ódio no português do Brasil
Métricas para avaliação de qualidade ed bases de dados para detecção de discurso de ódio	Métricas extraídas através da RLS, sendo: coenficiente de concordância entre anotadores (K), quantidade de anotadores (QA), grau de diversidade entre anotadores (GDA) e experiência dos anotadores (E).

de dados BR-RAPData, proposta neste trabalho com intuito de avaliar a discriminação de raça com base no dialeto pretuguês. Sendo assim, conforme discutido na Seção 4.1, os resultados apresentados na Subseção 3.4.3, evidenciam a presença de racismo algorítmico nos modelos treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E. Ainda mais, os modelos TF-IDF+MLP, treinados com as bases de dados OFFCOMBR-2 e TuPy-E, e o modelo BERTimbau$_{base}$, treinado com a base de dados HateBR, alcançaram valores para a métrica F1 igual a 83%, 74% e 91%, respectivamente. Mostrando-se resultados substanciais comparados às literaturas [8], [9] e [10].

As contribuições deste trabalho foram desenvolvidas entre os anos de 2021 a 2025. Ainda na graduação, foi desenvolvido o trabalho intitulado "Discriminação de raça em sistemas computacionais: Estudo de casos no contexto brasileiro", via iniciação científica voluntária. Este projeto fora apresentado no evento Integra UFMS, no ano de 2022. Ainda mais, no Apêndice B, está o trabalho intitulado "Natural language processing techniques for hate speech evaluation for Brazilian Portuguese" [61], publicado na Conferência Internacional sobre Ciência Computacional e suas Aplicações (ICCSA), no ano de 2023. Em 2024, foi apresentada a metodologia de pesquisa desenvolvida para investigação de discriminação algorítmica no 7º Workshop do Projeto InterSCity (INCT da Internet do Futuro para Cidades Inteligentes) e 1º Workshop do Projeto EcoSustain (Ciência de Dados e Computação para o Meio Ambiente). Em 2025, a resenha do livro de Benjamin [13], "Raça depois da Tecnologia: Ferramentas Abolicionistas para o Novo Código de Jim", foi publicada na 6ª edição da Revista

Outrora[2]. Ainda neste ano, o artigo "*Black drama: between hate speech and algorithmic racism*" fora submetido à Conferência Internacional de Aprendizado de Máquina e Aplicações (ICMLA), e está sob avaliação.

Em trabalhos futuros, pretende-se ampliar a base de dados BR-RAPData para ter mais composições de artistas que se identifiquem com o gênero feminino, e observar se há variação na detecção de discurso de ódio com base no gênero, além de raça; observar se as proporções de discurso de ódio para outros estilos musicais como sertanejo ou MPB se mantêm semelhantes às proporções emitidas para o conjunto de dados BR-RAPData; e por fim, construir uma base de dados demográfica, seguindo a metodologia de Blodgett e colaboradores [32], comparando o idioma português formal com o pretuguês, para avaliar se o racismo algorítmico persiste em uma base de dados que possua sentenças retiradas de redes sociais como *Instagram* e *X*.

[2]ROSA, C. C. S.; TASO, F. T. S. Raça depois da Tecnologia: Ferramentas Abolicionistas para o Novo Código de Jim. Revista Outrora, 2025. Disponível em <https://revistaoutrora.wixsite.com/revistaoutrora/numero-atual>. Acessado em: 6 ago. 2025.

Bibliografia

[1] Madelena Y Ng, Supriya Kapur, Katherine D Blizinsky, and Tina Hernandez-Boussard. The ai life cycle: a holistic approach to creating ethical ai for health decisions. *Nature medicine*, 28(11):2247–2249, 2022. Citado nas páginas vii, 8, e 9.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. Citado nas páginas vii, 19, 20, e 51.

[3] Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104, 2019. Citado nas páginas ix, 11, 12, 36, e 51.

[4] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977. Citado nas páginas ix, 15, e 16.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. Citado nas páginas xiii, 1, 2, e 9.

[6] Rosane Leal da Silva and Fernanda dos Santos Rodrigues da Silva. Reconhecimento facial e segurança pública: os perigos do uso da tecnologia no sistema penal seletivo brasileiro. In *Congresso Internacional de Direito e Contemporaneidade, Santa Maria, RS, Brasil*, volume 5, 2019. Citado nas páginas xiii, 1, e 10.

[7] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018. Citado nas páginas xiii, 1, 2, e 10.

[8] Rogers Prates De Pelle and Viviane P Moreira. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC, 2017. Citado nas páginas xiii, 5, 6, 32, 38, 45, 50, 52, e 73.

[9] Francielle Alves Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Fabrício Benevenuto, and Thiago Alexandre Salgueiro Pardo. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. *arXiv preprint arXiv:2103.14972*, 2021. Citado nas páginas xiii, 5, 6, 13, 14, 20, 25, 32, 36, 38, 43, 44, 45, 47, e 52.

[10] Felipe Oliveira, Victoria Reis, and Nelson Ebecken. Tupy-e: detecting hate speech in brazilian portuguese social media with a novel dataset and comprehensive analysis of models. *arXiv preprint arXiv:2312.17704*, 2023. Citado nas páginas xiii, 5, 6, 32, 36, 38, 45, 50, e 52.

[11] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018. Citado na página 2.

[12] Evgeny Morozov. *Big tech*. Ubu Editora LTDA-ME, 2018. Citado na página 2.

[13] Ruha Benjamin. *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons, 2019. Citado nas páginas 2, 7, e 52.

[14] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022. Citado nas páginas 2 e 8.

[15] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381, 2020. Citado nas páginas 2 e 8.

[16] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019. Citado nas páginas 2, 4, 5, 6, 14, 24, 27, 28, 33, 36, 38, 47, 50, e 52.

[17] Safiya Umoja Noble. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press, 2018. Citado na página 3.

[18] Tarcízio Silva. *Racismo algorítmico: inteligência artificial e discriminação nas redes digitais*. Edições Sesc SP, 2022. Citado nas páginas 3, 4, e 10.

[19] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017. Citado na página 3.

[20] Natasha Duarte, Emma Llanso, and Anna Loup. Mixed messages? the limits of automated social media content analysis. washington, dc: Center for democracy & technology, 2017. Citado na página 3.

[21] Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE, 2021. Citado na página 4.

[22] Jacquelyn Rahman. The n word: Its history and use in the african american community. *Journal of English Linguistics*, 40(2):137–171, 2012. Citado na página 4.

[23] Douglas Trajano, Rafael H Bordini, and Renata Vieira. Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, 58(4):1263–1289, 2024. Citado nas páginas 4, 14, 15, 24, 31, 32, e 47.

[24] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016. Citado nas páginas 4, 9, 25, 26, e 28.

[25] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32, 2023. Citado nas páginas 4 e 24.

[26] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. Citado nas páginas 4, 6, 35, e 50.

[27] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54, 2019. Citado na página 4.

[28] Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023. Citado na página 4.

[29] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019. Citado nas páginas 5, 6, 25, 28, 33, 36, 38, 47, 50, e 52.

[30] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*, 2020. Citado nas páginas 5, 27, 28, 33, e 52.

[31] Makosa Tomás David and Gabriel Nascimento. As influências das línguas bantu no português do brasil: Origens e trajetórias rumo ao pretoguês. *Mandinga-Revista de Estudos Linguísticos (ISSN: 2526-3455)*, 7(1):7–20, 2023. Citado nas páginas 5, 34, 47, e 52.

[32] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016. Citado nas páginas 5, 26, 27, 33, 50, e 53.

[33] Daniel Preoţiuc-Pietro and Lyle Ungar. User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th international conference on computational linguistics*, pages 1534–1545, 2018. Citado nas páginas 5, 27, e 33.

[34] Stephen Cave and Kanta Dihal. The whiteness of ai. *Philosophy & Technology*, 33(4):685–703, 2020. Citado na página 8.

[35] Steve Lohr et al. The age of big data. *New York Times*, 11(2012):12–16, 2012. Citado na página 8.

[36] Dias Oliva Thiago, Antonialli Dennys Marcelo, and Alessandra Gomes. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture*, 25(2):700–732, 2021. Citado na página 8.

[37] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*, 2021. Citado na página 8.

[38] You Chen, Ellen Wright Clayton, Laurie Lovett Novak, Shilo Anders, and Bradley Malin. Human-centered design to address biases in artificial intelligence. *Journal of medical Internet research*, 25:e43251, 2023. Citado na página 9.

[39] Susan Leavy, Barry O'Sullivan, and Eugenia Siapera. Data, power and bias in artificial intelligence. *arXiv preprint arXiv:2008.07341*, 2020. Citado na página 9.

[40] Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the fourth workshop on online abuse and harms*, pages 150–161, 2020. Citado na página 9.

[41] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1):2053951715622512, 2016. Citado na página 9.

[42] Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141, 2020. Citado nas páginas 9 e 10.

[43] Tarcízio da Silva. Visão computacional e racismo algorítmico: branquitude e opacidade no aprendizado de máquina. *Revista da Associação Brasileira de Pesquisadores/as Negros/as (ABPN)*, 12(31), 2020. Citado na página 9.

[44] Arivaldo Santos de Souza. Racismo institucional: para compreender o conceito. *Revista da Associação Brasileira de Pesquisadores/as Negros/as (ABPN)*, 1(3):77–88, 2011. Citado na página 9.

[45] Maria Aparecida da Silva Bento. *Pactos narcísicos no racismo: branquitude e poder nas organizações empresariais e no poder público.* PhD thesis, Universidade de São Paulo, 2002. Citado na página 10.

[46] Pablo Nunes. Novas ferramentas, velhas práticas: reconhecimento facial e policiamento no brasil. *Retratos da Violência–cinco meses de monitoramento, análises e descobertas*, pages 67–70, 2019. Citado na página 10.

[47] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021. Citado na página 10.

[48] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and engineering ethics*, 26(6):3333–3361, 2020. Citado na página 10.

[49] Stefan Larsson and Fredrik Heintz. Transparency in artificial intelligence. *Internet Policy Review*, 9(2), 2020. Citado na página 10.

[50] Mark Coeckelbergh. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, 26(4):2051–2068, 2020. Citado na página 10.

[51] Allan Dafoe. Ai governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443, 2018. Citado na página 11.

[52] Nathalie A Smuha. The eu approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 20(4):97–106, 2019. Citado na página 11.

[53] Brasil. Projeto de lei nº 2.338, de 2023. *Diário Oficial da República Federativa do Brasil*, 2023. Citado na página 11.

[54] Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018. Citado nas páginas 11 e 13.

[55] Hannah O Plath, Maria Estela O Paiva, Danielle L Pinto, and Paula DP Costa. Detecçao de discurso de odio contra mulheres em textos em português brasileiro: Construçao da base mina-br e modelo de classificaçao. *Revista Eletrônica de Iniciação Científica em Computação*, 20(3), 2022. Citado nas páginas 11 e 32.

[56] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016. Citado nas páginas 11, 13, 14, 15, 25, e 26.

[57] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017. Citado nas páginas 11, 12, 13, 14, 24, 26, 27, 48, e 51.

[58] Michel Rosenfeld. Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo L. Rev.*, 24:1523, 2002. Citado na página 12.

[59] Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. Categorizing sexism and misogyny through neural approaches. *ACM Transactions on the Web (TWEB)*, 15(4):1–31, 2021. Citado na página 12.

[60] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*, 2021. Citado na página 13.

[61] Cássia CS Rosa, Fábio V Martinez, and Renato Ishii. Natural language processing techniques for hate speech evaluation for brazilian portuguese. In *International Conference on Computational Science and Its Applications*, pages 104–117. Springer, 2023. Citado nas páginas 13, 14, 17, 20, 31, 32, 52, e 70.

[62] Joao A Leite, Diego F Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*, 2020. Citado nas páginas 13, 17, 24, 32, 36, e 51.

[63] Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*, 2020. Citado nas páginas 13 e 73.

[64] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018. Citado nas páginas 13, 14, 26, e 27.

[65] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017. Citado na página 14.

[66] Rehab Duwairi, Amena Hayajneh, and Muhannad Quwaider. A deep learning framework for automatic detection of hate speech embedded in arabic tweets. *Arabian Journal for Science and Engineering*, 46:4001–4014, 2021. Citado na página 14.

[67] Matthijs J Warrens. Five ways to look at cohen's kappa. *Journal of Psychology & Psychotherapy*, 5, 2015. Citado na página 15.

[68] Julius Sim and Chris C Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005. Citado na página 15.

[69] Joseph L Fleiss, Jacob Cohen, and Brian S Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological bulletin*, 72(5):323, 1969. Citado na página 15.

[70] Julie Beth Lovins. Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, 11(1-2):22–31, 1968. Citado na página 17.

[71] Singh Jasmeet and Gupta Vishal. Text stemming: Approaches, applications, and challenges. *ACM Comput. Surv*, 49(3):1–46, 2016. Citado na página 17.

[72] Viviane Moreira Orengo and Christian R Huyck. A stemming algorithmm for the portuguese language. In *spire*, volume 8, pages 186–193, 2001. Citado na página 17.

[73] Reinaldo Viana Alvares, Ana Cristina Bicharra Garcia, and Inhaúma Ferraz. Stembr: A stemming algorithm for the brazilian portuguese language. In *Portuguese conference on artificial intelligence*, pages 693–701. Springer, 2005. Citado na página 17.

[74] João Ferreira, Hugo Gonçalo Oliveira, and Ricardo Rodrigues. Improving nltk for processing portuguese. In *8th Symposium on Languages, Applications and Technologies (SLATE 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019. Citado na página 17.

[75] H. M. Caseli and M. G. V. Nunes, editors. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition, 2024. Citado na página 18.

[76] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Citado na página 18.

[77] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. Citado nas páginas 20 e 22.

[78] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer, 2020. Citado nas páginas 20, 22, 27, 38, e 51.

[79] Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. Tucano: Advancing neural text generation for portuguese. *arXiv preprint arXiv:2411.07854*, 2024. Citado nas páginas 20, 22, e 38.

[80] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009. Citado na página 21.

[81] Hind Saleh, Areej Alhothali, and Kawthar Moria. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719, 2023. Citado na página 22.

[82] Stephen Akuma, Tyosar Lubem, and Isaac Terngu Adom. Comparing bag of words and tf-idf with different models for hate speech detection from live tweets. *International Journal of Information Technology*, 14(7):3629–3635, 2022. Citado na página 22.

[83] Adilson Moreira. *Racismo recreativo*. Pólen Produção Editorial LTDA, 2019. Citado na página 24.

[84] Luc Steels. Human language is a culturally evolving system. *Psychonomic bulletin & review*, 24:190–193, 2017. Citado na página 24.

[85] Simona Frenda et al. The role of sarcasm in hate speech. a multilingual perspective. In *Proceedings of the doctoral symposium of the xxxivinternational conference of the spanish society for natural language processing (sepln 2018)*, pages 13–17. Lloret, E.; Saquete, E.; Mart ınez-Barco, P.; Moreno, I., 2018. Citado na página 24.

[86] Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398, 2022. Citado na página 24.

[87] Dennys Antonialli. Drag queen vs. david duke: Whose tweets are more 'toxic'. *Wired. Retrieved (July/August 2019)*, 2019. Citado na página 24.

[88] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. Citado na página 24.

[89] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021. Citado nas páginas 24 e 47.

[90] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233, 2017. Citado na página 26.

[91] Thomas Davidson and Debasmita Bhattacharya. Examining racial bias in an online abuse corpus with structural topic modeling. *arXiv preprint arXiv:2005.13041*, 2020. Citado nas páginas 26 e 28.

[92] Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th international conference on recent advances in natural language processing*, pages 1187–1196, 2023. Citado nas páginas 27, 28, 36, e 38.

[93] Gabriel Nascimento, Flavio Carvalho, Alexandre Martins da Cunha, Carlos Roberto Viana, and Gustavo Paiva Guedes. Hate speech detection using brazilian imageboards. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 325–328, 2019. Citado na página 32.

[94] Bruno Ferrari Guide. *Detecção automática de discurso de ódio punitivista em redes sociais*. PhD thesis, Universidade de São Paulo, 2022. Citado na página 32.

[95] Guilherme Lima, Amanda Oliveira, Felix Silva, Luana Pinheiro, Eduardo Luz, and Larissa Freitas. Desafios sobre a detecção de discurso de ódio em português brasileiro: Construção de um novo corpus e reflexões sobre o processo. *Americas Conference on Information Systems (AMCIS).*, 2024. Citado na página 32.

[96] Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. Hatebrxplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in brazilian portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669, 2025. Citado na página 32.

[97] RALPH D'agostino and Egon S Pearson. Tests for departure from normality. empirical results for the distributions of b 2 and b. *Biometrika*, 60(3):613–622, 1973. Citado na página 40.

[98] Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010. Citado na página 40.

[99] Sidia M Callegari-Jacques. *Bioestatística: princípios e aplicações*. Artmed Editora, 2009. Citado na página 40.

[100] Cláudia Pons Cardoso. Amefricanizando o feminismo: o pensamento de lélia gonzalez. *Revista Estudos Feministas*, 22(03):965–986, 2014. Citado na página 47.

[101] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021. Citado na página 51.

APÊNDICE

A

Resultados Detalhados da Pesquisa

Este apêndice reúne os resultados dos testes de significância estatística realizados nas etapas de 09 e 10 da metodologia de investigação de racismo algorítmico.

As Figuras A.2, A.3 e A.4 mostram as probabilidades emitidas por cada modelo para a classe discurso de ódio do conjunto de dados BR-RAPData. A Figura A.1 mostra a quantidade de sentenças classificadas como discurso de ódio por modelos treinados com OFFCOMBR-2, HateBR e TuPy-E. Na Tabela A.1, estão os resultados dos testes de normalidade, identificado como p, para cada amostragem de BR-RAPData, classificada como discurso de ódio pelos modelos. A Tabela A.2 mostra os resultados do teste qui-quadrado para cada tupla de amostragem de BR-RAPData, classificada como discurso de ódio pelos modelos treinados com OFFCOMBR-2, HateBR e TuPy-E.

Tabela A.1: Resultados dos testes de normalidade para cada amostragem do conjunto de dados BR-RAPData, que fora classificada como discurso de ódio pelos modelos treinados com OFFCOMBR-2, HateBR e TuPy-E.

	Bases de dados		
	OFFCOMBR-2	HateBR	TuPy-E
Modelo	p	p	p
TFI-IDF+RL	8.476389269669449e-15	2.133669115373844e-20	1.954204532877015e-05
TF-IDF+MLP	5.8796380215978055e-06	1.5199028463701202e-31	5.909271552178307e-19
CBoW+RL	6.249775632045007e-29	1.333718015311639e-33	6.484031565276153e-06
CBoW+MLP	2.624040326012902e-45	4.3502031117887456e-35	1.2066903980000186e-132
BERTimbau$_{base}$	1.1127575041967457e-54	2.6384334233355906e-27	4.2707533342928017e-10
Tucano-1b1	1.3328023939085917e-10	9.161294059305301e-42	8.690264521584161e-08

Tabela A.2: Resultados dos testes qui-quadrado de cada par de amostragem de BR-RAPData, rotulados como discurso de ódio por modelos treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E.

OFFCOMBR-2						
	TFIDF + RL	TFIDF + MLP	CBoW + RL	CBoW + MLP	BERTimbau$_{base}$	Tucano-1b1
TFIDF + RL	x	9.507794451193171e-07	8.539185685935149e-61	3.476954495772707e-141	1.9748928428423216e-26	7.799154139298625e-08
TFIDF + MLP		x	1.4231808296981769e-31	7.335993058732268e-96	8.12911345720559e-09	0.6623387863429553
CBoW + RL			x	7.318881377269188e-21	2.5962501960447334e-09	2.957579841767358e-29
CBoW + MLP				x	1.9556052062452316e-52	6.72620200185691e-92
BERTimbau$_{base}$					x	1.1862651880210016e-07
Tucano-1b1						x
HateBR						
	TFIDF + RL	TFIDF + MLP	CBoW + RL	CBoW + MLP	BERTimbau$_{base}$	Tucano-1b1
TFIDF + RL	x	0.0013975410178407302	0.06068489924255563	1.439120745486107e-234	2.1879388964157325e-47	2.6061167363917098e-64
TFIDF + MLP		x	3.384000651213691e-07	1.815105066199872e-196	1.1477630442874596e-29	2.0316885011505712e-43
CBoW + RL			x	3.825047739825756e-258	1.051179307622755e-59	2.0575972365926738e-78
CBoW + MLP				x	4.256585664091125e-85	6.499744472066273e-66
BERTimbau$_{base}$					x	0.011514930045704915
Tucano-1b1						x
TuPy-E						
	TFIDF + RL	TFIDF + MLP	CBoW + RL	CBoW + MLP	BERTimbau$_{base}$	Tucano-1b1
TFIDF + RL	x	6.8348613090966724e-06	0.00022464748003565269	5.131104997398355e-103	4.5273789487358926e-92	3.0995287043716014e-11
TFIDF + MLP		x	4.748937124753481e-16	1.4280416296090914e-69	2.736692903276112e-60	0.03263825481298226
CBoW + RL			x	1.2728255375547832e-131	1.0901803876246146e-119	2.598126886453557e-24
CBoW + MLP				x	0.19339520063675059	4.259326733844849e-55
BERTimbau$_{base}$					x	9.107900161464051e-47
Tucano-1b1						x

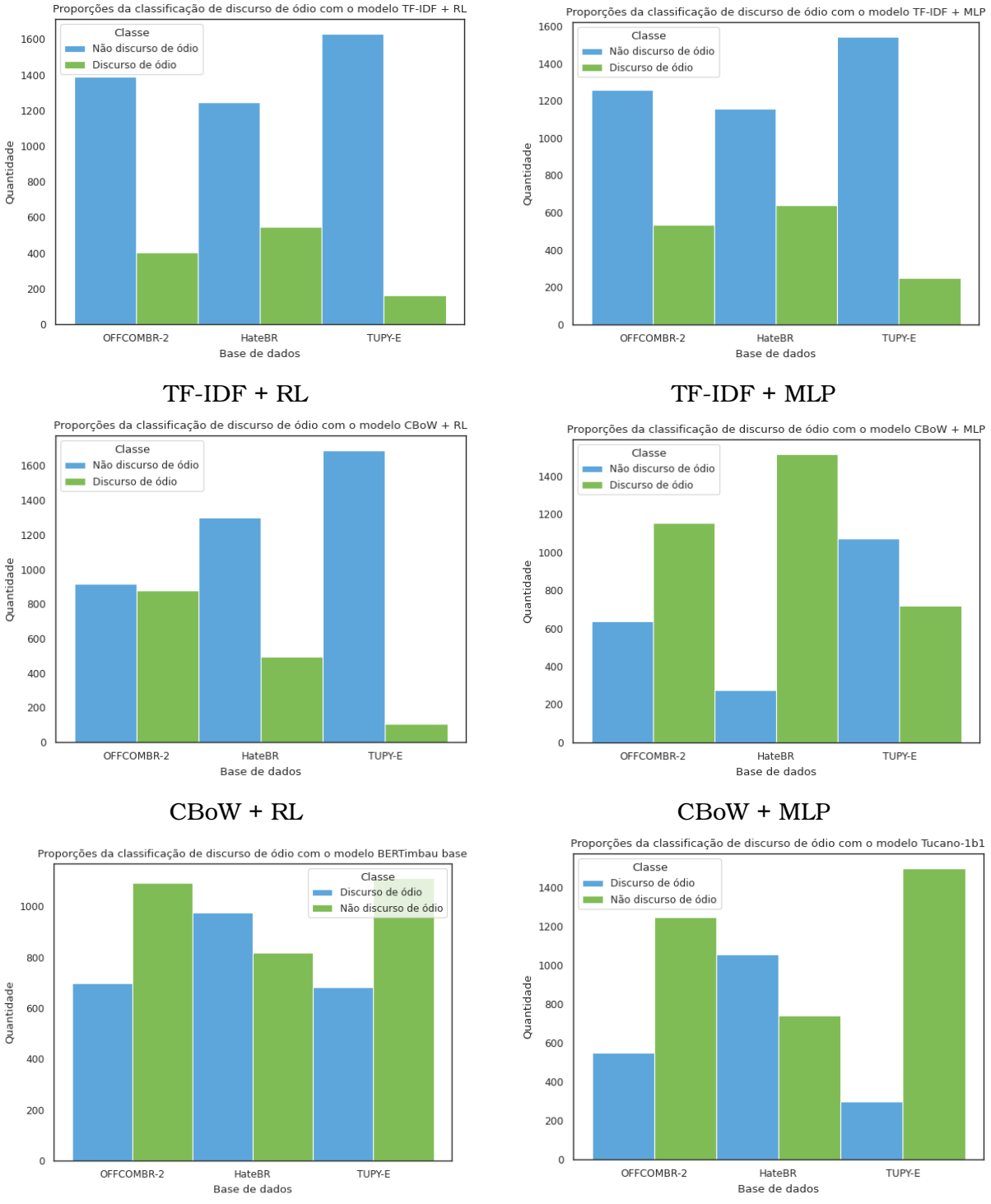
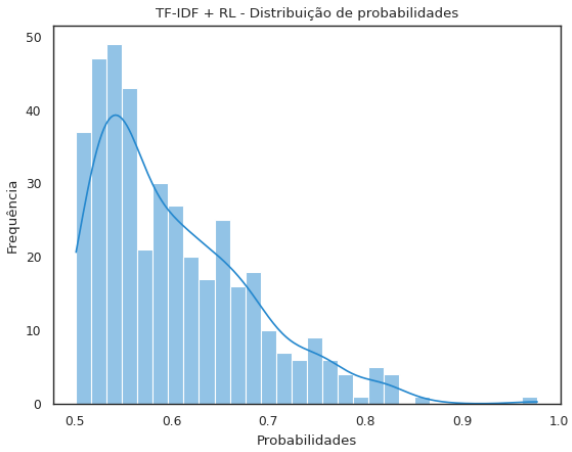
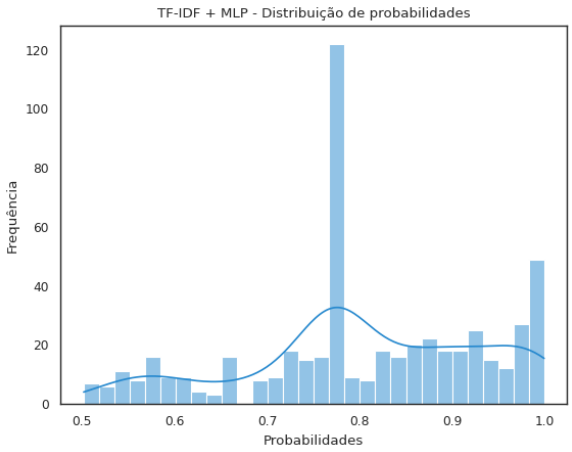


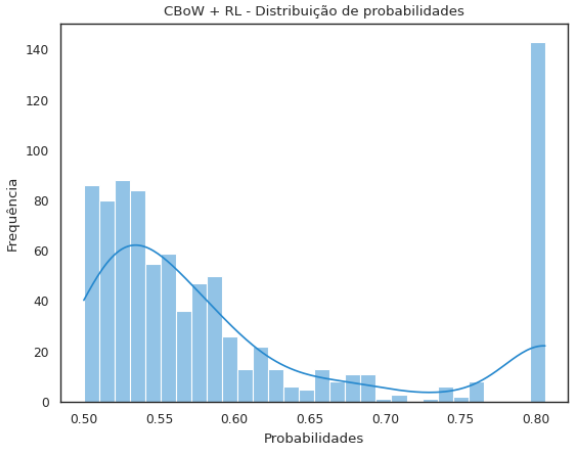
Figura A.1: Quantidade de sentenças do conjunto de dados BR-RAPData, preditas como "Discurso de ódio" e "Não discurso de ódio" pelos classificadores TF-IDF+RL, TF-IDF+MLP, CBoW+RL, CBoW+MLP, BERTimbau$_{base}$ e Tucano-1b1, treinados com as bases de dados OFFCOMBR-2, HateBR e TuPy-E.



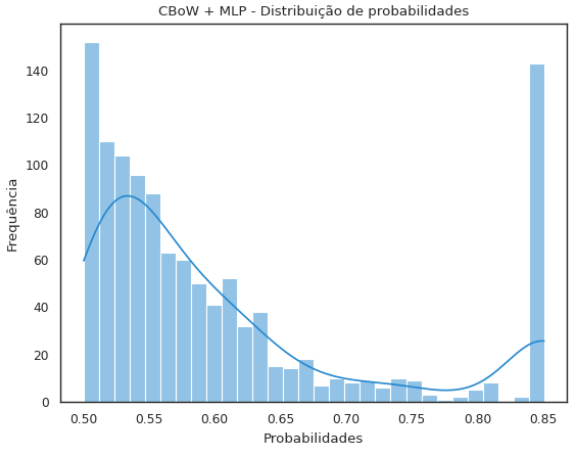
TFIDF + RL



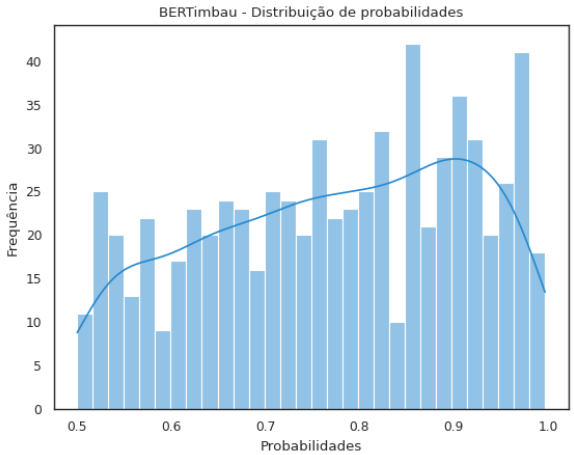
TFIDF + MLP



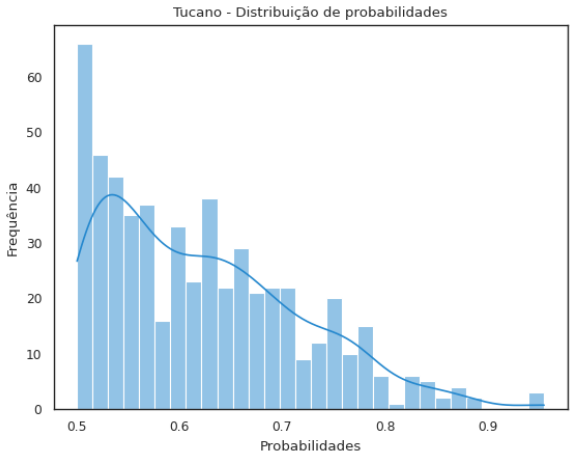
CBoW + RL



CBoW + MLP

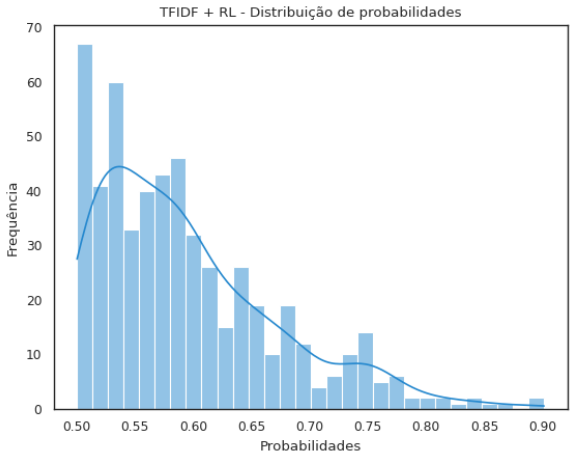


BERTimbau$_{base}$

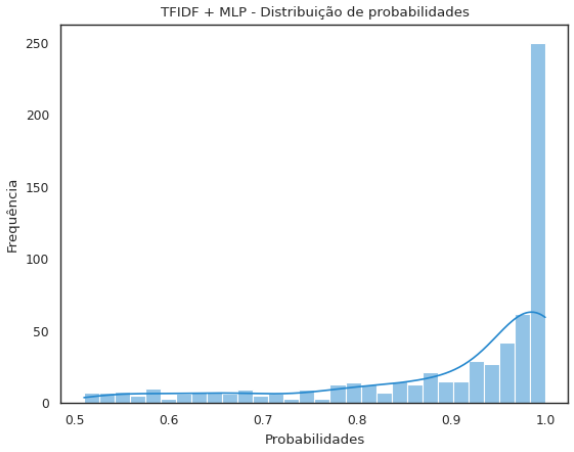


Tucano-1b1

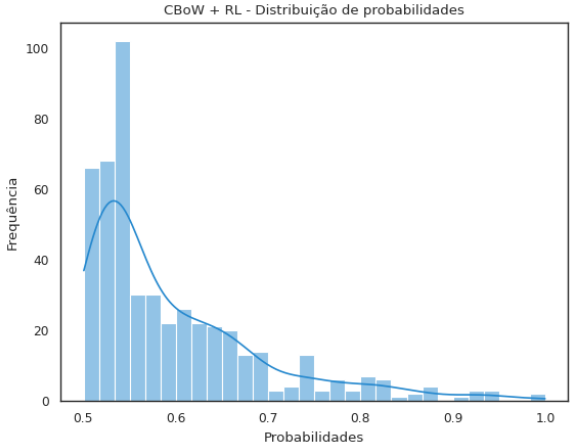
Figura A.2: Distribuição de probabilidades das sentenças de BR-RAPData classificadas como discurso de ódio por modelos treinados com OFFCOMBR-2.



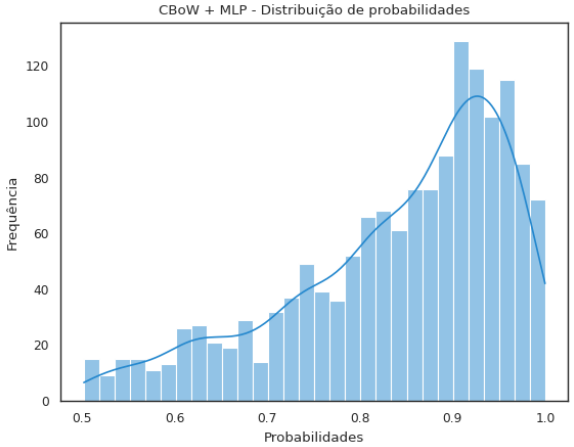
TFIDF + RL



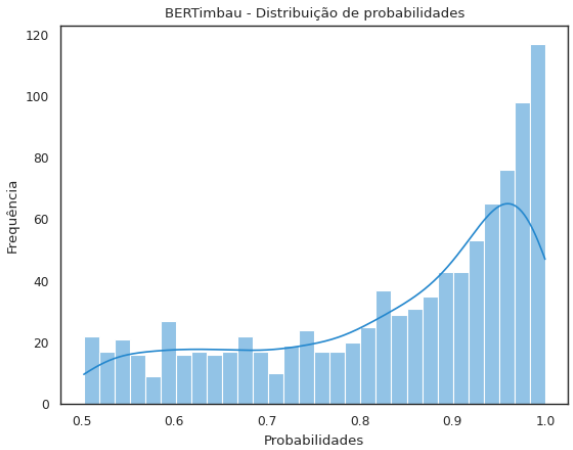
TFIDF + MLP



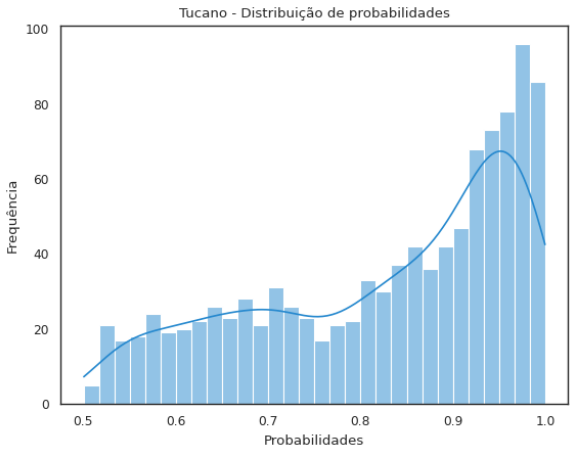
CBoW + RL



CBoW + MLP

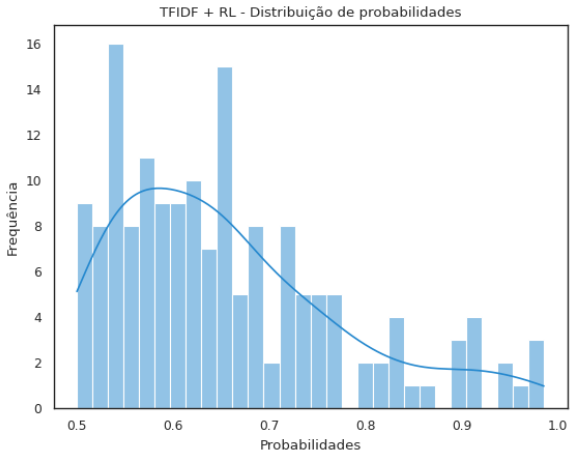


BERTimbau$_{base}$

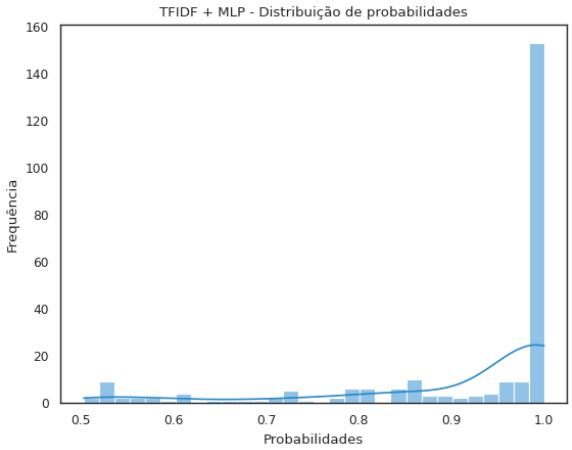


Tucano-1b1

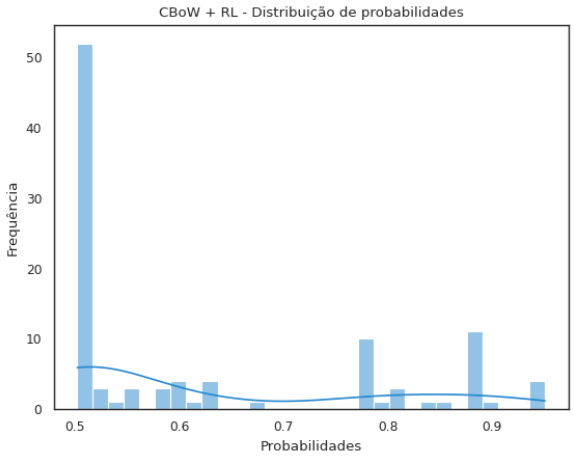
Figura A.3: Distribuição de probabilidades das sentenças de BR-RAPData classificadas como discurso de ódio por modelos treinados com HateBR.



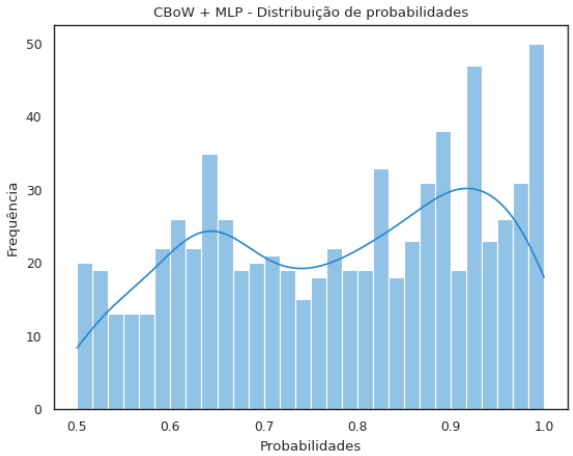
TFIDF + RL



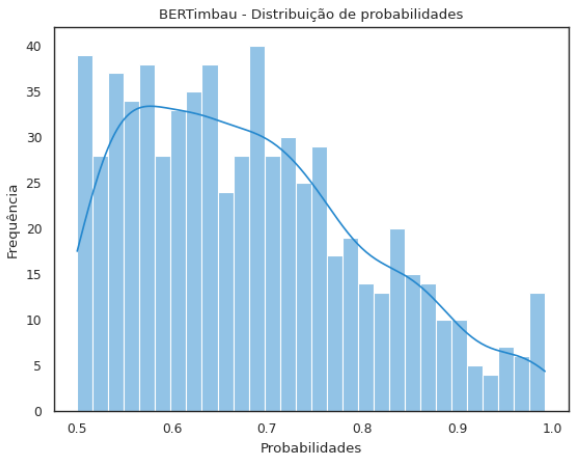
TFIDF + MLP



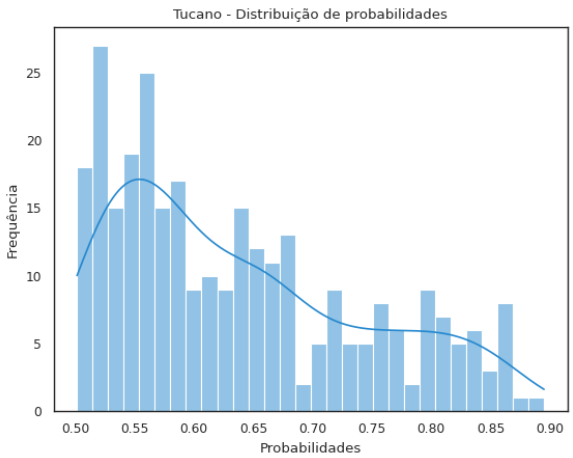
CBoW + RL



CBoW + MLP



BERTimbau$_{base}$



Tucano-1b1

Figura A.4: Distribuição de probabilidades das sentenças de BR-RAPData classificadas como discurso de ódio por modelos treinados com TuPy-E.

APÊNDICE

B

Técnicas de Processamento de Linguagem Natural para Avaliação de Discurso de Ódio para o Português Brasileiro

O artigo publicado em junho de 2023, durante a "*24th International Conference on Computational Science and Its Applications*" (ICCSA) [61], contribuiu para área de detecção de discurso de ódio com o conjunto de dados TwitterHateBR[1]. Além do ajuste fino de um modelo treinado com o português do Brasil, o BERTimbau. Com isso, foi possível demonstrar resultados iniciais sobre o funcionamento e eficácia de um modelo, bem como as estratégias adotadas para coleta e rotulação dos dados.

Como experimento inicial, uma raspagem de dados com a ferramenta Snscrape foi feita na mídia social Twitter (X), em 2022, durante as eleições presidenciais no Brasil. Para isso, um *script* buscou termos que poderiam identificar estereótipos utilizados para atingir uma pessoa ou a comunidade a qual ela pertence, com base no seu gênero, raça, etnia, nacionalidade, orientação sexual, entre outras características. Também foram buscadas postagens que estavam marcadas com "#eleições2022", procurando coletar dados para analisar o cenário político daquele momento. Na tabela B.1, estão dispostos os termos de identificação escolhidos para compor os tipos de discriminação abordados no conjunto de dados para avaliação de discurso de ódio. A escolha dos termos teve como base palavras usadas para proferir ofensas. Todavia, elas, por si só, não são consideradas odiosas, visto que a depender do contexto, podem ser empregadas de maneira não ofensiva entre os membros de algumas comunidades.

[1]Disponível em <https://github.com/cassiasilvaR/TwitterHateBR>. Acessado em 21 ago. 2025.

Período	Termos de Identificação	Tipo de discriminação
Outubro a Novembro de 2022	“favelado”, “nordestino”, “índio”, “africano”	Xenofobia
Outubro a Novembro de 2022	“piranha”, “vagabunda”, “feminazi”, “macumbeira”	Misoginia
Agosto a Novembro de 2022	“humor negro”, “negro”, “escravo”	Racismo
Setembro a Novembro de 2022	“retardado”, “débil”, “demente”	Capacitismo
Setembro a Novembro de 2022	“obeso”, “gordo”	Gordofobia
Janeiro de 2019 a Novembro de 2022	“traveco”, “afeminado”, “opção sexual”	LGBTFobia

Tabela B.1: Mostra quais termos foram usados para realizar a raspagem de dados no Twitter. Também é mostrado que tipo de preconceito era abordado a partir do termo escolhido e em qual período essa coleta foi realizada.

A partir dos dados coletados, a anotação foi feita somente pela autora do artigo, uma mulher negra de 21 anos, nascida no Brasil. O conjunto final, composto por 848 sentenças, estava divida entre as classes “discurso_de_odio”, “neutro”, e “ofensa”. Sendo 108 discurso de ódio, 370 neutro e 370 ofensa. Alguns exemplos coletados e anotados podem ser vistos na tabela B.2. Nota-se que na 5ª linha de “Texto”, a palavra “piranha”, usada para a raspagem de dados por ser potencialmente ofensiva, está em uma palavra rotulada como neutra, porque o contexto da frase não apresenta conotação ofensiva ou odiosa.

Texto	Rótulo
Que negros de merda	Discurso de Ódio
Sua loira vagabunda do caralho morte p vc é pouco...	Discurso de Ódio
Essa enquete prova que o brasileiro é um povo MT retardado...	Ofensa
tenho muito nojo de gordo suado cheio de banha kkk	Ofensa
Amiga mas nem todo mundo tem o sonho de ser piranha	Neutro
Homens negros falem sobre amor, isso motiva outros negros	Neutro

Tabela B.2: Exemplos de sentenças coletadas pela raspagem de dados. Os três tipos de classe estão presentes: Discurso de Ódio, Ofensa e Neutro.

Também foi feita uma análise exploratória nos dados a fim de obter maior compreensão do contexto sobre o corpus coletado. A partir de bigramas, foi possível observar a predominância dos pares de palavras em cada uma das classes, assim como em todas as sentenças. Na figura B.1, o bigrama da classe de discurso de ódio mostra uma predominância de bigramas com conotação misógina, evidenciando que essas palavras foram associadas ao discurso de ódio. No gráfico com bigramas de pares

ofensivos, palavras com conotação misógina também são vistas, além de palavras com conotação capacitista e palavras de baixo calão. Com a associação de termos misóginos ao discurso de ódio, há potencialidade de existir um viés direcionado ao uso dessas palavras. Todavia, o contexto em que elas foram aplicadas, nesse caso o bigrama que avalia os pares, mostra a ressignificação desses termos pela comunidade de mulheres, e sim como uma forma de propagar discursos sexistas. Comparando os bigramas do gráfico de sentenças de discurso de ódio com o exemplo de ressignificação, apresentado na tabela B.2, é possível notar que antes do termo "piranha", não há palavras que induzem ao discurso de ódio ou ofensa, como visto nos pares de palavras. Por isso, além do contexto, a correlação entre as palavras é um elemento muito importante para a detecção de discurso de ódio. O bigrama mais frequente em todas as classes é o de "opção sexual". Esse termo foi buscado no maior intervalo temporal na coleta dados, evidenciando um desbalanceamento. Assim como os outros, o termo somente não pode ser classificado como discurso de ódio ou ofensivo. Entretanto, militantes da causa LGBTQIA+ informam que o seu uso é incorreto, e que "orientação sexual" seria mais adequado.

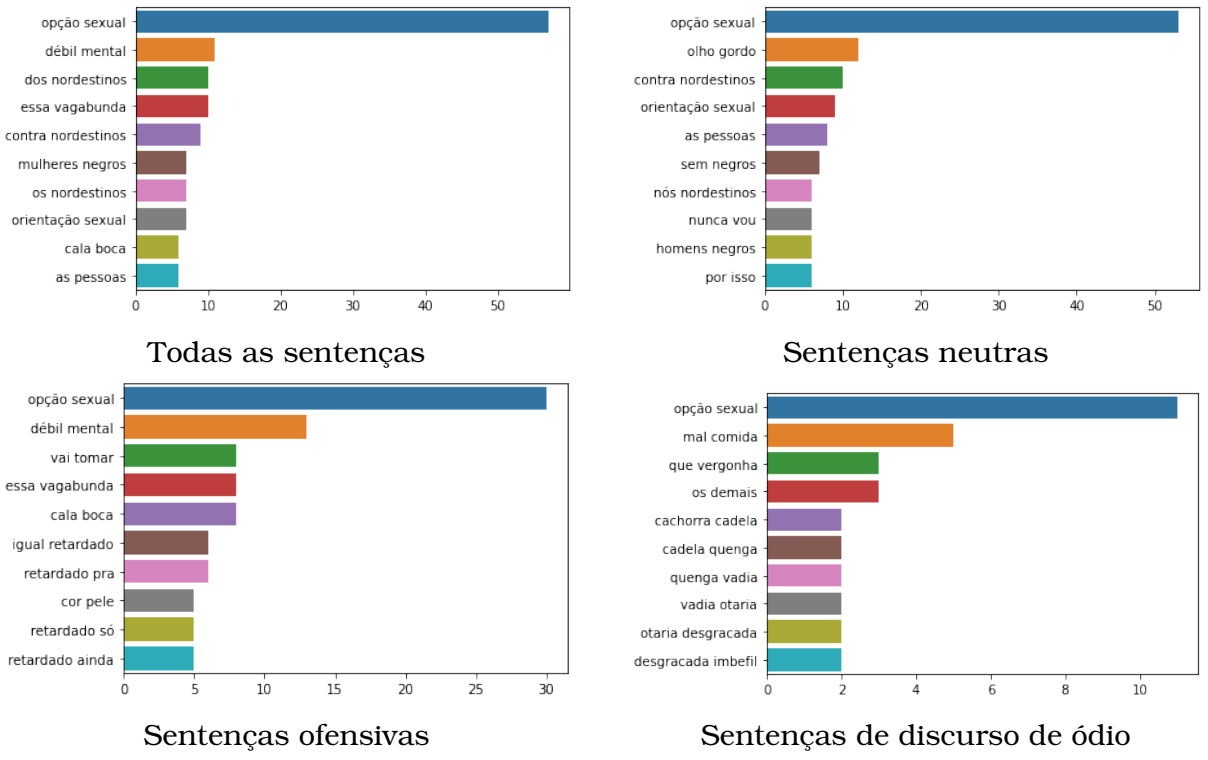


Figura B.1: Na parte superior: na esquerda, é mostrado o gráfico com os bigramas de palavras mais frequentes em todas as sentenças; e na direita, é mostrado o gráfico com os bigramas de palavras mais frequentes na classe de sentenças neutras. Na parte inferior: na esquerda, é mostrado o gráfico com os bigramas de palavras mais frequentes na classe de sentenças ofensivas; e na direita, é mostrado o gráfico com os bigramas mais frequentes na classe de sentenças odiosas.

O modelo BERTimbau foi usado para o ajuste fino com 3 conjuntos de dados distintos. O primeiro modelo foi ajustado com o conjunto de dados *Multilingual Hate-*

Check (MHC), de Röttger e seus colaboradores [63], desenvolvido para realizar testes funcionais. O segundo modelo foi ajustado com o conjunto de dados "OFFCOMBR-2", de Pelle e Moreira [8]. O último modelo foi ajustado com o conjunto de dados construído neste trabalho, o TwitterHateBR. Todos as bases de dados foram modificadas para conter as classes "neutro" e "discurso_de_odio". A divisão para o ajuste ocorreu da seguinte forma: 80% para treino e 20% para teste. A execução foi feita em 4 épocas, com tamanho de lote igual a 16 e taxa de aprendizagem iniciando em 0.00001. O primeiro modelo obteve uma média F1 de 68% e acurácia de 66%, enquanto o segundo modelo obteve 74% de média F e 73% de acurácia. O modelo ajustado com a base de dados desenvolvida nesse trabalho obteve média F e acurácia de 74%.

B.1 Considerações finais

Detectar discurso de ódio online é fundamental, principalmente nos períodos em que são discutidos os direitos humanos fundamentais. Por exemplo, durante as eleições presidenciais. O ajuste do modelo BERTimbau com conjuntos de dados em português contribuiu para identificação de comportamento ofensivos em ambientes de interação virtual. Além disso, a análise sobre o contexto e o cuidado ao rotular sentenças potencialmente odiosas, retratados neste trabalho, serão vistas com mais cautela em trabalhos futuros.