UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

José Fernando de Oliveira Filho

Análise de Viés no Modelo BERTimbau para detecção de discurso de ódio em Português Brasileiro

Recife

Fevereiro de 2026

José Fernando de Oliveira Filho

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientadora: Roberta Macedo Marques Gouveia

Recife
Fevereiro de 2026

José Fernando de Oliveira Filho

Análise de Viés no Modelo BERTimbau para detecção de discurso de ódio em Português Brasileiro

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco

Departamento de Estatística e Informática

Curso de Bacharelado em Sistemas de Informação

Aprovada em: 12 de Fevereiro de 2026

Banca Examinadora

Roberta Macedo Marques Gouveia (Orientadora)

Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Gabriel Alves de Alburquerque Júnior

Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Análise de Viés no Modelo BERTimbau para Detecção de Discurso de Ódio em Português Brasileiro

[José Fernando de Oliveira Filho] [1], **[Roberta Macedo Marques Gouveia]** [2]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco (UFRPE)
Rua Dom Manuel de Medeiros, s/n - CEP: 52171-900 - Recife - PE - Brasil

josefernando.oliveirafilho@ufrpe.br, roberta.gouveia@ufrpe.br

Resumo. *A expansão das redes sociais intensificou a circulação de discursos de ódio online, gerando desafios a convivência democrática e proteção de grupos minoritários. Diante da inviabilidade da moderação manual, este trabalho aplica técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML) para a identificação de conteúdo ofensivo em português brasileiro. O estudo investiga o viés algorítmico do modelo BERTimbau, ajustado via fine-tuning e treinado sobre as bases anotadas ToLD-BR e Tupy-E. A avaliação considera métricas tradicionais de desempenho (precisão, recall, F1-score) e de equidade (Demographic Parity), com o objetivo de examinar possíveis associações indevidas entre termos de identidade social e discurso de ódio, contribuindo para o debate sobre justiça algorítmica. Os resultados evidenciaram padrões estruturados de viés do BERTimbau entre os eixos de gênero, raça e orientação sexual, bem como entre grupos específicos dentro de cada eixo (por exemplo, mulheres, pessoas bissexuais, gays e pardas), mostrando que o modelo tende a associar esses termos de identidade ao discurso de ódio com maior frequência do que seus grupos de referência, mesmo em contextos neutros ou positivos.*

Abstract. *The expansion of social networks has intensified the spread of hate speech online, posing challenges to democratic coexistence and the protection of minority groups. Given the impracticality of manual moderation, this study applies Natural Language Processing (NLP) and Machine Learning (ML) techniques to identify offensive content in Brazilian Portuguese. It investigates algorithmic bias in the BERTimbau model, fine-tuned and trained on the annotated ToLD-BR and Tupy-E datasets. The evaluation employs both traditional performance metrics (precision, recall, F1-score) and fairness metrics (Demographic Parity) to examine potential unjust associations between social identity terms and hate speech, contributing to discussions on algorithmic fairness. The results revealed structured patterns of bias across the gender, race, and sexual orientation axes, as well as across specific groups within each axis (e.g., women, bisexual, gay, and brown people), showing that the model tends to associate these social identity terms with hate speech more frequently than their reference groups, even in neutral or positive contexts.*

1. Introdução

A expansão das redes sociais transformou a comunicação humana no século XXI, criando espaços virtuais onde bilhões de pessoas interagem diariamente, compartilhando ideias, opiniões e experiências.

Recentes estudos destacam que, apesar da democratização trazida por esses ambientes, a circulação de discursos de ódio online se intensificou, o que apresenta riscos à convivência democrática e aos direitos das minorias (VENTUROTT, 2021).

Diante de milhões de publicações diárias, a moderação manual do discurso de ódio na plataforma da rede social X (antigo Twitter) se torna impraticável. Nesse contexto, ferramentas de aprendizado de máquina e processamento de linguagem natural passaram a ser centrais para a classificação automatizada de conteúdos tóxicos (RODRIGUES, 2023).

De acordo com Fortuna e Nunes (2018), a classificação de discurso de ódio baseada em aprendizado de máquina enfrenta desafios relacionados à definição de fenômeno, à subjetividade inerente ao processo de anotação, ao desbalanceamento entre classes e as variações linguísticas e socioculturais entre diferentes contextos. No contexto da língua portuguesa, estudos voltados a construção de conjuntos de dados anotados, como ToLD-BR (LEITE et al., 2020) e Tupy-E (OLIVEIRA; REIS; EBECKEN, 2023), evidenciam esses desafios de forma prática, refletindo as particularidades do português brasileiro nos dados utilizados para o treinamento e avaliação de modelos.

Contudo, o uso de modelos de linguagem avançados também traz desafios éticos e técnicos. De acordo com Marcolin, Paiva e Marchezini (2023), modelos de machine learning aplicados à moderação de conteúdo podem reproduzir e amplificar vieses sociais presentes nos dados, especialmente quando termos associados a identidades de grupos minoritários como *mulher*, *negro* ou *LGBTQIA+* passam a ser correlacionados com toxicidade, mesmo em contextos neutros ou positivos. As autoras destacam que esse fenômeno compromete a justiça algorítmica ao penalizar de forma desproporcional grupos historicamente marginalizados, levantando debates sobre equidade, liberdade de expressão e dignidade em ambientes digitais.

Nesse contexto, Nascimento, Cavalcanti e Da Costa-Abreu (2022) apresentam um dos esforços recentes voltados à mensuração e mitigação de injustiças algorítmicas em modelos de linguagem, destacando a necessidade de avaliações críticas de vieses não intencionais. Ainda que adotem uma abordagem distinta, seus achados reforçam a importância de análises sistemáticas de viés em modelos de processamento de linguagem natural e machine learning. A utilização de modelos baseados em transformadores, como o BERTimbau, para a classificação de discurso de ódio tem apresentado resultados promissores em termos de desempenho. Contudo, seu comportamento ético ainda demanda investigações complementares, especialmente no que se diz respeito à amplificação de vieses sociais presentes nos dados de treinamento.

Diante desse cenário, os dados das bases Tupy-E e ToLD-BR são pré-processados cuidadosamente, visando garantir a qualidade textual e a padronização das entradas. O modelo BERTimbau é ajustado (fine-tuning) e empregado na tarefa de classificação das mensagens.

Para a análise de viés, são construídas sondas linguísticas contendo termos associados a grupos minoritários, com o objetivo de identificar padrões de falsos positivos do modelo. A avaliação considera tanto métricas tradicionais de desempenho, como precisão, recall e F1-score, quanto métricas de equidade, em especial a *Demographic Parity*.

O objetivo do estudo consiste em quantificar e discutir os impactos sociais e éticos associados a esses vieses, contribuindo para o desenvolvimento de sistemas de moderação

automática mais justos e responsáveis no contexto brasileiro.

O artigo está organizado nas seguintes seções:

- Seção 2 - Referencial Teórico: onde são abordados os conceitos fundamentais sobre discurso de ódio, viés algorítmico, processamento de linguagem natural, machine learning e classificação, o modelo BERTimbau, as métricas de desempenho de modelos de classificação e as métricas de equidade algorítmica.
- Seção 3 - Trabalhos relacionados: apresenta um levantamento dos principais estudos que abordam a detecção automática de discurso de ódio, com foco especial nas pesquisas realizadas em língua portuguesa e naquelas que analisam vieses algorítmicos em modelos de linguagem natural.
- Seção 4 - Metodologia: descreve as etapas de desenvolvimento do trabalho, incluindo seleção e pré-processamento dos dados, ajuste do modelo BERTimbau via fine-tuning, procedimentos experimentais e avaliação com métricas de desempenho e equidade.
- Seção 5 - Resultados e Discussão: apresenta e analisa os resultados obtidos, destacando o comportamento do modelo frente aos termos de identidade e suas implicações éticas e sociais.
- Seção 6 - Conclusões: sintetiza as contribuições do estudo, as limitações encontradas e propõe perspectivas para trabalhos futuros voltados à redução do viés em modelos de detecção de discurso de ódio em português brasileiro.

2. Referencial Teórico

A presente seção tem como objetivo apresentar a fundamentação teórica que sustenta esta pesquisa, abordando os conceitos essenciais relacionados ao discurso de ódio no ambiente digital, às técnicas de Processamento de Linguagem Natural (PLN), ao aprendizado de máquina e à análise de viés algorítmico em modelos de linguagem. Essas discussões fornecem a base conceitual necessária para compreender as etapas metodológicas desenvolvidas no trabalho.

2.1. Discurso de ódio

O discurso de ódio é um fenômeno complexo que desafia os campos jurídico e tecnológico. De acordo com a UNESCO (2019) ele pode ser definido como qualquer forma de comunicação oral, escrita, ou comportamental que ataque e utilize linguagem pejorativa ou discriminatória em referência a uma pessoa ou grupo com base na sua religião, etnia, nacionalidade, raça, cor, descendência ou outro fator de identidade.

De acordo com Fortuna e Nunes (2018), o discurso de ódio online se caracteriza por sua natureza difusa, velocidade de transmissão e dificuldade de responsabilização, visto que as plataformas digitais possibilitam o anonimato e um alcance amplificado de discurso. Além disso, a falta de consenso sobre sua definição operacional constitui um obstáculo significativo para o desenvolvimento de sistemas automatizados de detecção, uma vez que diferentes plataformas, legislações e contextos culturais adotam critérios distintos para classificar mensagens como odiosas.

Segundo Vidgen et al. (2019) uma possível forma de síntese conceitual pode ser observada na Tabela 1, que resume três dimensões fundamentais do discurso de ódio

digital: semântica, social e técnica e suas respectivas implicações para o estudo computacional.

Tabela 1. Dimensões analíticas do discurso de ódio online

Dimensão	Descrição	Implicações para o PLN
Semântica	Uso explícito ou implícito de termos depreciativos ou incitação à violência.	Necessidade de detectar ironia, metáforas e contexto linguístico.
Social	Envolve a reprodução de estigmas e desigualdades estruturais.	Demanda modelos sensíveis ao contexto cultural e identitário.
Técnica	Refere-se aos desafios de anotação, rotulação e desbalanceamento dos dados.	Exige técnicas de aprendizado supervisionado robustas e bases equilibradas.

Fonte: Adaptado de Vidgen et al. (2019).

2.2. Viés Algorítmico em Inteligência Artificial

O Viés algorítmico refere-se a distorções sistemáticas causadas por decisões automatizadas que produzem resultados injustos ou discriminatórios para determinados grupos sociais. De acordo com Mehrabi et al. (2021), modelos de aprendizado de máquina tendem a reproduzir e, em alguns casos, amplificar desigualdades já presentes nos dados utilizados em seu treinamento.

Os autores destacam que tais distorções podem emergir tanto dos conjuntos de dados como vieses históricos, de amostragem e de anotação quanto de escolhas associadas aos próprios algoritmos e aos processos de avaliação. Como consequência, tais vieses podem gerar impactos sociais significativos em domínios sensíveis, incluindo justiça criminal, saúde e sistemas automatizados de seleção e recomendação.

A partir dessa perspectiva geral, trabalhos posteriores passaram a detalhar de forma mais específica as fontes e os efeitos do viés algorítmico em sistemas de aprendizado de máquina. Yang et al. (2023) apontam que as principais origens desse viés incluem conjuntos de dados historicamente desbalanceados, escolhas relacionadas à arquitetura e às funções objetivo dos modelos, bem como decisões humanas envolvidas na definição de métricas e critérios de sucesso.

No contexto específico da detecção de discurso de ódio, Das et al. (2024) evidenciam que modelos frequentemente apresentam taxas elevadas de falsos positivos para mensagens associadas a grupos minoritários, fenômeno atribuído tanto a vieses nos dados quanto às interpretações socioculturais dos anotadores humanos.

Diante desse cenário, diferentes estratégias de mitigação tem sido propostas na literatura, abrangendo técnicas de pré-processamento dos dados, como balanceamento e geração sintética de exemplos, intervenções nos próprios algoritmos e métodos de pós processamento voltados à correção de decisões injustas. Ferramentas como o AI Fairness 360 (BELLAMY et al., 2018) consolidam essas abordagens ao oferecer métricas e métodos sistemáticos para avaliação e mitigação de vieses em modelos de aprendizado de máquina.

2.3. Justiça Algorítmica

A justiça algorítmica refere-se ao uso de critérios formais e matemáticos para avaliar se sistemas automatizados de decisão tratam grupos sociais de maneira equitativa. Diferentemente das discussões conceituais sobre viés algorítmico, essa abordagem concentra-se na definição de métricas capazes de mensurar desigualdades nos resultados produzidos por modelos de aprendizado de máquina (BAROCAS; HARDT; NARAYANAN, 2018).

A paridade demográfica, a *Equalized Odds* e a *Predictive Parity* constituem algumas das principais definições matemáticas de justiça algorítmica utilizadas na avaliação de sistemas automatizados de decisão. A paridade demográfica é proposta por Dwork et al. (2011) e exige que a taxa de decisões positivas seja equivalente entre grupos sensíveis, independentemente do rótulo verdadeiro, buscando assegurar que previsões favoráveis sejam distribuídas de forma proporcional. Hardt, Price e Srebro (2016) formalizam a *equalized odds*, que impõe a igualdade das taxas de verdadeiros positivos e falsos positivos entre os grupos, considerando explicitamente o rótulo real e evitando que erros do modelo afetem de maneira desproporcional determinados segmentos sociais. Por sua vez Chouldechova (2017) discute a *predictive parity*, segundo a qual a precisão das previsões positivas deve ser a mesma entre os grupos garantindo que uma decisão positiva tenha igual probabilidade de estar correta para todos.

Entretanto essas definições não são em geral, simultaneamente satisfazíveis sob condições realistas de dados. Chouldechova e Roth (2018) demonstram formalmente que, quando diferentes grupos apresentam diferentes prevalências do rótulo verdadeiro o que é uma condição comum em dados sociais é matematicamente impossível satisfazer a paridade demográfica, a *equalized odds* e a *predictive parity*. Esse resultado indica que a escolha de uma métrica de justiça algorítmica envolve trade-offs e não é neutra, devendo considerar o contexto da aplicação, os tipos de erro mais relevantes e suas implicações éticas e sociais (BAROCAS; HARDT; NARAYANAN, 2018).

2.4. Processamento de Linguagem Natural e Classificação

O Processamento de Linguagem Natural (PLN) é o campo da inteligência artificial que se dedica a desenvolver técnicas e algoritmos que permitem aos computadores compreender, processar e gerar a linguagem humana de forma automatizada (JURAFSKY; MARTIN, 2020). Diferentemente de tarefas estruturadas, a linguagem natural apresenta ambiguidade, variabilidade e contexto dependente, o que torna o PLN uma área complexa e desafiadora. As principais abordagens incluem análise sintática, extração de entidades, análise de sentimentos e, relevantemente para este trabalho, a classificação de textos.

Na detecção automática de discurso de ódio, o PLN é combinado com técnicas de aprendizado de máquina para categorizar automaticamente mensagens em redes sociais como contendo ou não discurso ofensivo e discriminatório (FORTUNA; NUNES, 2018). Este é um problema de classificação supervisionada em que o modelo aprende padrões a partir de exemplos anotados.

Os principais desafios incluem a natureza subjetiva do discurso de ódio, que varia conforme contextos culturais e legislações, bem como o uso de ironia, sarcasmo e linguagem implícita, os quais exigem uma compreensão profunda do contexto para correta interpretação. Soma-se a isso a variabilidade linguística característica das redes sociais, marcada pelo uso frequente de gírias, abreviações e dialetos, o que dificulta a

padronização do processamento textual. Além disso, destaca-se o desequilíbrio de classes, uma vez que mensagens de ódio tendem a representar uma parcela minoritária em relação ao volume total de conteúdo analisado.

2.4.1. Métricas de Avaliação em Classificação de Texto

Modelos de classificação supervisionada em Processamento de Linguagem Natural (PLN) são avaliados por meio de métricas estatísticas que quantificam sua capacidade de distinguir corretamente as classes de interesse. A escolha dessas métricas é fundamental para uma avaliação adequada do desempenho dos modelos, especialmente em tarefas de classificação de texto que envolvem dados ruidosos e distribuições desbalanceadas entre classes, como discutido por Manning, Raghavan e Schütze (2008) e também por Jurafsky e Martin (2020).

A acurácia (accuracy) mede a proporção de previsões corretas em relação ao total de amostras avaliadas. Apesar de sua ampla utilização, essa métrica pode ser enganosa em cenários de classes desbalanceadas, uma vez que classificadores que privilegiam a classe majoritária podem alcançar valores elevados sem, de fato, apresentar bom desempenho na identificação da classe minoritária (SOKOLOVA; LAPALME, 2009).

A precisão (precision) quantifica a proporção de instâncias classificadas como positivas que realmente pertencem à classe positiva. Essa métrica é particularmente relevante quando o custo de falsos positivos é elevado, pois indica o grau de confiabilidade das previsões positivas realizadas pelo modelo (POWERS, 2020).

O recall (ou revocação) mede a proporção de instâncias positivas corretamente identificadas pelo classificador em relação ao total de instâncias positivas existentes. Em aplicações sensíveis, como a detecção de discurso de ódio, valores baixos de recall indicam a omissão de conteúdos potencialmente ofensivos, o que compromete a efetividade do sistema de detecção (SOKOLOVA; LAPALME, 2009).

O F1-score corresponde à média harmônica entre precisão e recall, sendo amplamente adotado em tarefas de classificação de texto por fornecer uma medida balanceada do desempenho do modelo, especialmente quando há assimetria na distribuição das classes. Segundo Powers (2020), essa métrica é mais informativa do que a acurácia isolada em cenários desbalanceados, pois penaliza classificadores que apresentam desempenho elevado em apenas uma das dimensões.

Complementarmente, a matriz de confusão representa as quantidades de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos produzidos pelo modelo. Essa ferramenta permite uma análise detalhada dos padrões de erro, facilitando a identificação de vieses e limitações do classificador, além de orientar ajustes no processo de treinamento e avaliação (MANNING; RAGHAVAN; SCHÜTZE, 2008).

De forma conjunta, essas métricas possibilitam uma avaliação abrangente do desempenho dos modelos de classificação de texto, considerando não apenas a taxa global de acertos, mas também os diferentes tipos de erro cometidos. Tal abordagem é essencial em aplicações com impacto social, nas quais erros distintos podem acarretar consequências éticas e operacionais relevantes (JURAFSKY; MARTIN, 2020).

2.5. Modelos baseados em Transformers: BERT e BERTimbau

A arquitetura transformer, proposta por Vaswani et al. (2017) revolucionou o PLN ao adicionar mecanismos de atenção que permitem capturar relações de longo alcance em sequências de texto de forma paralela e eficiente. O BERT (Bidirectional Encoder Representations from Transformers), desenvolvido por Devlin et al. (2019), aplicou essa arquitetura com um pré-processamento bidirecional inovador. Diferentemente de modelos anteriores que processavam texto de forma unidirecional (esquerda para direita ou direita para esquerda), o BERT processa o texto em ambas as direções simultaneamente, capturando contextos mais ricos e nuances semânticas. O modelo foi pré-treinado em um extenso corpus multilíngue, incluindo Wikipedia em múltiplos idiomas, adquirindo representações linguísticas profundas.

Graças ao seu desempenho superior demonstrado em inúmeras tarefas de PLN (tradução, sumarização, resposta a perguntas, etc.), o BERT se tornou rapidamente um modelo de referência em pesquisa e aplicações industriais. No entanto, a aplicação do BERT multilíngue a tarefas específicas do português brasileiro frequentemente resultava em desempenho subótimo, uma vez que o treinamento multilíngue não capturava suficientemente as particularidades do português falado no Brasil (SOUZA; NOGUEIRA; LOTUFO, 2020).

O BERTimbau surgiu como resposta à lacuna existente na disponibilização de modelos de linguagem de grande porte especificamente adaptados ao português brasileiro. Desenvolvido por Souza, Nogueira e Lotufo (2020), o BERTimbau consiste em uma adaptação pré-treinada do modelo BERT (Bidirectional Encoder Representations from Transformers) voltada ao português do Brasil. Segundo os autores, o nome BERTimbau resulta da combinação entre o acrônimo BERT e o termo berimbau, instrumento musical tradicional brasileiro, sendo uma escolha simbólica que reforça o caráter nacional do modelo.

O pré-treinamento do BERTimbau foi realizado utilizando o corpus BrWaC (*Brazilian Portuguese Web as Corpus*), que compreende bilhões de palavras coletadas da web brasileira, incluindo conteúdos provenientes de sites jornalísticos, blogs e redes sociais. O modelo foi treinado seguindo o mesmo protocolo adotado no BERT original, mas com vocabulário e distribuição linguística adequados ao português brasileiro.

Resultados empíricos apresentados por Souza, Nogueira e Lotufo (2020) demonstram que o BERTimbau supera de forma consistente o BERT multilíngue e outros modelos pré-treinados em tarefas de classificação de texto em português, bem como em tarefas de similaridade semântica, inferência textual e reconhecimento de entidades. Esses resultados indicam vantagens relevantes do modelo em aplicações de classificação sensíveis ao contexto linguístico, como a detecção de discurso de ódio.

Sua aplicação em tarefas sensíveis como detecção de discurso de ódio beneficia-se particularmente de sua sensibilidade às particularidades linguísticas e culturais do português brasileiro, reduzindo o risco de interpretações errôneas derivadas de modelos treinados em outros idiomas ou contextos.

2.6. Fine-tuning do BERTimbau

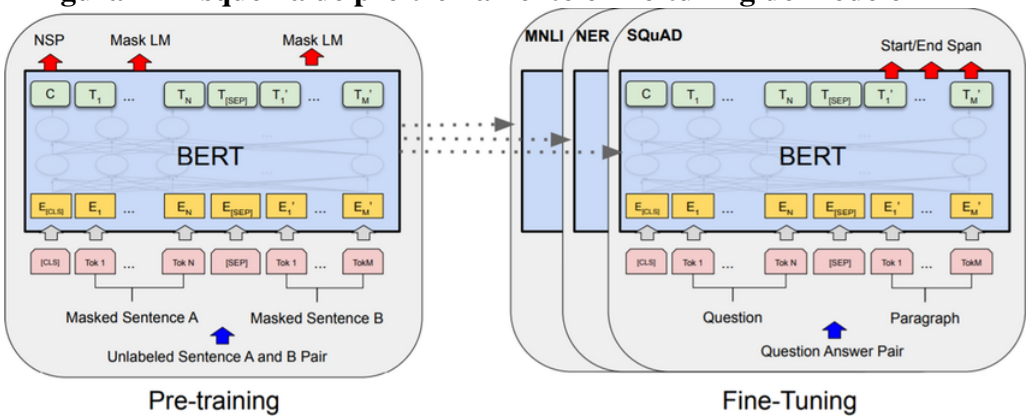
O aprendizado por transferência, ou transfer learning, é uma técnica fundamental em machine learning que aproveita o conhecimento adquirido em uma tarefa para melhorar a

performance em outra tarefa relacionada (PAN; YANG, 2010).

No contexto de PLN, o fine-tuning é a aplicação prática mais comum dessa estratégia, consistindo no ajuste de um modelo pré-treinado para uma tarefa específica por meio de treinamento adicional com um conjunto de dados anotados da tarefa-alvo (HOWARD; RUDER, 2018).

O processo de *fine-tuning* do BERTimbau para a tarefa de classificação de discurso de ódio segue a arquitetura proposta originalmente por Devlin et al. (2019). De forma geral, esse processo consiste na adaptação de um modelo BERT pré-treinado por meio da adição de camadas específicas para a tarefa de classificação, conforme ilustrado na Figura 1. Inicialmente, o modelo processa a sequência de entrada e gera *embeddings* contextuais para cada token. Em seguida, é extraído o *token de classificação* ([CLS]), inserido no início da sequência, o qual é projetado para representar semanticamente toda a sentença. Esse embedding é então propagado por uma ou mais camadas densas (*fully connected*) com funções de ativação não lineares, tipicamente ReLU ou *tanh*. Por fim, a saída da última camada é processada por uma camada softmax, responsável por produzir uma distribuição de probabilidades sobre as classes consideradas, neste caso discurso de ódio e não discurso de ódio (DEVLIN et al., 2019).

Figura 1 – Esquema de pré-treinamento e fine-tuning do modelo BERT



Fonte: Devlin et al. (2019).

O processo de fine-tuning de modelos baseados em Transformers, como o BERT e o BERTimbau, depende fortemente da escolha adequada de hiperparâmetros. Esses parâmetros controlam como o modelo ajusta seus pesos durante o treinamento e influenciam diretamente sua capacidade de generalização. A taxa de aprendizagem (*learning rate*) define a intensidade das atualizações aplicadas aos pesos do modelo; valores muito altos podem causar instabilidade ou explosão do gradiente, enquanto valores muito baixos podem resultar em convergência lenta ou estagnação. Para o BERT, a literatura recomenda taxas entre 1×10^{-5} e 5×10^{-5} (DEVLIN et al., 2019).

O tamanho do batch controla quantas amostras são processadas simultaneamente durante o treinamento; batches maiores tendem a estabilizar o gradiente, mas exigem maior capacidade de memória, sendo comuns valores entre 8 e 32 para GPUs convencionais.

O número de épocas (*epochs*) corresponde ao número de vezes que o modelo per-

corre completamente o conjunto de treinamento. Para tarefas de classificação com BERT, usualmente entre duas e quatro épocas são suficientes, pois muitas iterações podem levar ao sobreajuste (overfitting) (HOWARD; RUDER, 2018). O *dropout* consiste em um mecanismo que desativa aleatoriamente unidades da rede durante o treinamento, reduzindo o risco de sobreajuste; no BERT, o valor padrão adotado é 0,1, embora esse parâmetro possa ser ajustado conforme a tarefa. Por fim, o otimizador utilizado no treinamento do BERT é o AdamW, uma variação do Adam que incorpora regularização por weight decay, contribuindo para maior estabilidade no treinamento de modelos baseados em Transformers (LOSHCHILOV; HUTTER, 2017).

A escolha apropriada desses hiperparâmetros é determinante para garantir que o modelo apresente bom desempenho, estabilidade e capacidade de generalização sem incorrer em overfitting.

3. Trabalhos Relacionados

No contexto da detecção automatizada de discurso de ódio e da justiça algorítmica, diversos estudos recentes têm investigado não apenas os vieses que modelos de aprendizado podem reproduzir ou amplificar contra grupos específicos, mas também as métricas e técnicas destinadas a mensurar e mitigar tais desigualdades.

O trabalho de Nascimento, Cavalcanti e Da Costa-Abreu (2022) analisa o desempenho de modelos de detecção de discurso de ódio em mídias sociais com foco no viés de gênero, propondo uma abordagem de ensemble em múltiplos espaços de representação para avaliar e mitigar vieses indesejados em classificadores treinados em dados reais. Os autores mostram que, mesmo quando as métricas globais de desempenho são elevadas, podem existir disparidades significativas entre grupos, evidenciando a necessidade de métricas específicas de equidade e de estratégias de mitigação de viés. Esse trabalho é particularmente relevante por combinar avaliação empírica com preocupações de justiça algorítmica, servindo como referência metodológica para a análise de viés em detecção de ódio.

Seguindo uma linha semelhante, Mozafari, Farahbakhsh e Crespi (2020) investigam detecção de discurso de ódio em inglês e propõem mecanismos para mitigação de viés racial em modelos baseados em BERT, combinando regularização e reweighting dos dados para reduzir disparidades nas taxas de erro entre grupos. Os resultados indicam que ajustes cuidadosos no treinamento podem diminuir vieses sem comprometer drasticamente o desempenho global, reforçando a importância de considerar métricas de fairness no desenho de sistemas de moderação automática.

Mais recentemente, o trabalho desenvolvido por Brito et al. (2025) implementam o corpus ToxSyn, explora o uso de dados sintéticos para estudar e reduzir viés em detecção de discurso de ódio, com foco em múltiplos grupos minoritários e em cenários de generalização para alvos pouco representados nos dados reais. O trabalho destaca que a construção controlada de exemplos pode auxiliar na identificação de padrões enviesados em modelos e na avaliação mais sistemática de seu comportamento em relação a grupos sensíveis.

O presente TCC aproxima-se desses estudos ao adotar uma perspectiva explícita de justiça algorítmica na detecção de discurso de ódio, mas diferencia-se por focar especificamente no contexto do português brasileiro e no uso do modelo BERTimbau treinado

sobre o dataset ToLD-BR e Tupy-E. Enquanto os trabalhos de Nascimento, Cavalcanti e Da Costa-Abreu (2022) e Mozafari, Farahbakhsh e Crespi (2020) tratam, em geral, de viés de gênero ou racial em contextos majoritariamente em inglês, esta pesquisa concentra-se em três dimensões de identidade (raça, gênero e orientação sexual) em textos em português brasileiro, utilizando sondas neutras e a métrica de equidade Demographic Parity para quantificar disparidades. Dessa forma, o trabalho contribui para preencher uma lacuna na literatura sobre viés algorítmico em PLN no contexto brasileiro, complementando e estendendo as abordagens existentes para uma realidade linguística e social ainda pouco explorada.

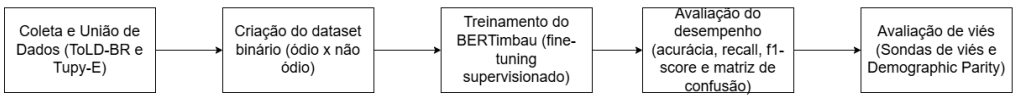
4. Metodologia

A metodologia deste estudo foi organizada em quatro etapas principais. Primeiro, as bases ToLD-Br e Tupy-E foram unificadas em um único conjunto de dados. Em seguida, construiu-se a variável binária para distinguir textos com e sem discurso de ódio. Na terceira etapa, realizou-se o treinamento supervisionado do modelo BERTimbau a partir do conjunto unificado. Por fim, o modelo foi avaliado quanto ao seu desempenho e submetido a sondas de justiça algorítmica, permitindo verificar possíveis disparidades entre grupos sociais.

Durante a execução da metodologia e da análise exploratória dos dados, foi utilizada a linguagem de programação Python (versão 3.11.12) no ambiente colaborativo Google Colab. Para a manipulação e análise dos dados tabulares, empregou-se a biblioteca Pandas, enquanto operações numéricas foram realizadas com o auxílio da biblioteca NumPy. A geração de gráficos durante a análise exploratória foi conduzida por meio da biblioteca Matplotlib. A divisão do conjunto de dados em subconjuntos de treino, validação e teste foi realizada utilizando a biblioteca Scikit-learn. Para a tokenização dos textos e o processo de fine-tuning do modelo BERTimbau, utilizou-se a biblioteca Transformers. A biblioteca Datasets foi empregada na criação de estruturas de dados compatíveis com o framework PyTorch, que, por sua vez, foi utilizado como backend para o treinamento supervisionado do modelo.

O fluxo completo das etapas realizadas encontra-se representado na Figura 2.

Figura 2 – Fluxo das etapas da metodologia



Fonte: O autor (2026).

4.1. Datasets ToLD-BR e Tupy-E

O ToLD-Br é um conjunto de dados composto por aproximadamente 21 mil tweets em português brasileiro, coletados a partir de palavras-chave relacionadas a discurso ofensivo e tóxico. A anotação foi realizada por um grupo de voluntários, sendo que cada instância recebeu múltiplos rótulos independentes em categorias associadas a diferentes formas de toxicidade, como racismo, misoginia, xenofobia e LGBTfobia. O estudo reporta níveis moderados de concordância entre os anotadores, o que evidencia a natureza subjetiva da tarefa de identificação de discurso de ódio. Em razão de seu tamanho, diversidade de

anotações e disponibilidade pública, o ToLD-Br é amplamente utilizado como base de referência para o treinamento e a avaliação de modelos de Processamento de Linguagem Natural voltados ao português brasileiro (LEITE et al., 2020).

O TuPy-E é um corpus anotado para a detecção de discurso de ódio em português brasileiro e é considerado um dos maiores conjuntos de dados desse tipo disponíveis na literatura. O dataset foi proposto com o objetivo de fornecer uma base ampla e representativa para o treinamento e a avaliação de modelos de PLN nessa tarefa, contemplando diferentes categorias de discurso ofensivo e tóxico. Em virtude de sua escala e organização, o TuPy-E tem sido utilizado como referência em estudos recentes sobre detecção automática de discurso de ódio em português (OLIVEIRA; REIS; EBECKEN, 2023).

4.2. Preparação e Padronização dos Dados

A primeira etapa consistiu na obtenção das bases ToLD-Br e Tupy-E, ambas amplamente utilizadas em tarefas de detecção de discurso de ódio em português brasileiro. Os arquivos foram carregados no ambiente Google Colab para permitir manipulação, limpeza e preparação dos dados. As bases apresentam características distintas:

ToLD-Br: multilabel, com anotações para racismo, homofobia, misoginia, xenofobia, insulto, obsceno, feitas por 3 anotadores.

Tupy-E: binária, com rótulos aggressiveness e hate.

Após o carregamento, iniciou-se a inspeção inicial das colunas e estatísticas básicas para compreender a estrutura e consistência dos dados.

Para possibilitar a futura unificação das bases, foi necessário realizar transformações estruturais, devido às diferenças entre seus esquemas.

A UNESCO (2019) define discurso de ódio como manifestações que promovem ou justificam discriminação, hostilidade ou violência contra indivíduos ou grupos com base em características identitárias, como etnia, nacionalidade, orientação sexual ou gênero. Essa definição normativa foi adotada como referência neste trabalho para delimitar o que é considerado discurso de ódio. Com base nessa definição e considerando a estrutura do dataset ToLD-Br, definiu-se que apenas categorias que representam explicitamente discriminação direcionada a grupos sociais seriam utilizadas na construção do rótulo binário. Assim, foram selecionadas exclusivamente as categorias associadas a racismo, homofobia, misoginia e xenofobia como indicadores diretos de discurso de ódio, enquanto as demais categorias do conjunto de dados não foram incorporadas à definição adotada, de modo a assegurar coerência entre o referencial teórico e as decisões metodológicas (LEITE et al., 2020).

Com base nessa distribuição e visando compatibilizar o ToLD-Br ao formato binário adotado no TuPy-E, criou-se a variável hate-strict, definida como valor 1 quando ao menos uma categoria de discurso de ódio racismo, homofobia, misoginia ou xenofobia estivesse presente, e 0 caso contrário. Após a unificação do ToLD-Br e do TuPy-E e a conversão de ambos para um rótulo binário comum, o conjunto de dados consolidado passou a contar com 55.934 instâncias, das quais 50.697 (90,64%) correspondem à classe não ofensiva e 5.237 (9,36%) à classe de discurso de ódio. Essa distribuição evidencia um acentuado desbalanceamento entre as classes, reforçando a necessidade de avaliações cuidadosas durante o treinamento e a validação dos modelos de classificação.

Em seguida, com o rótulo binário já definido, os textos foram preparados para o modelo BERTimbau por meio da tokenização, realizada com o *tokenizer* oficial da biblioteca Transformers. Esse processo converteu cada texto em sequências numéricas adequadas ao modelo, aplicando truncamento e *padding* automático, e gerando as estruturas necessárias para o treinamento *input ids*, *attention mask* e *token type ids*, além dos rótulos. Por fim, os dados tokenizados foram organizados em formato compatível com PyTorch por meio da biblioteca HuggingFace Datasets, permitindo seu uso direto durante o processo de *fine-tuning*.

4.3. Treinamento do BERTimbau

Após a etapa de preparação e padronização dos dados, o modelo BERTimbau foi ajustado por meio de aprendizado supervisionado, utilizando o conjunto de dados unificado e rotulado para a tarefa de classificação binária de discurso de ódio. Esta etapa teve como objetivo adaptar os pesos do modelo pré-treinado às particularidades linguísticas e semânticas do problema estudado, preservando o conhecimento previamente aprendido durante o pré-treinamento em larga escala.

O treinamento foi conduzido de forma controlada, com monitoramento contínuo das métricas de desempenho, a fim de avaliar a convergência do modelo, identificar possíveis sinais de overfitting e analisar o impacto do desbalanceamento de classes no aprendizado.

4.3.1. Fine-tuning e Configurações de Treinamento

Foi utilizado o modelo BERTimbau Base Cased, pré-treinado sobre o corpus BrWaC, amplamente empregado em tarefas de Processamento de Linguagem Natural para o português brasileiro. Para a tarefa de classificação, o embedding associado ao token especial [token de classificação], que sintetiza semanticamente a sequência de entrada, foi conectado a uma camada densa responsável por estimar as probabilidades das duas classes consideradas: discurso de ódio e não ódio.

O processo de *fine-tuning* consistiu no ajuste de todos os pesos do modelo, incluindo as camadas profundas da arquitetura *transformer*, permitindo que as representações contextuais fossem especializadas para o domínio específico do discurso analisado. Essa abordagem é amplamente adotada na literatura por possibilitar melhor adaptação de modelos pré-treinados a tarefas supervisionadas específicas, sobretudo em cenários com quantidade limitada de dados anotados (THOMAS, 2025).

O treinamento foi realizado em ambiente com suporte a GPU, utilizando a biblioteca Transformers, e os hiperparâmetros foram definidos com base em boas práticas consolidadas na literatura para modelos baseados em BERT, considerando também as limitações computacionais do experimento. Especificamente, foi empregado o otimizador AdamW, com taxa de aprendizado inicial de 2×10^{-5}, agendamento linear de decaimento da taxa de aprendizado e treinamento ao longo de três épocas. A função de perda adotada foi a Cross-Entropy Loss, e a avaliação do desempenho do modelo foi realizada ao final de cada época sobre o conjunto de validação, conforme recomendado em estudos prévios sobre *fine-tuning* de modelos *transformer* (DEVLIN et al., 2019; WIEDEMANN; YIMAM; BIEMANN, 2020).

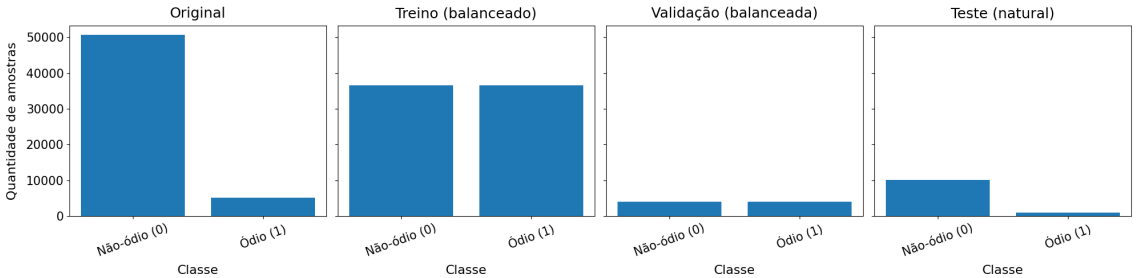
Considerando o forte desbalanceamento entre as classes identificado na análise exploratória dos dados, foram adotadas estratégias específicas para mitigar seus efeitos no processo de aprendizado do modelo. O conjunto de dados original apresenta aproximadamente 90% de instâncias rotuladas como não ódio e cerca de 10% como ódio, o que poderia induzir o classificador a favorecer sistematicamente a classe majoritária durante o treinamento.

Para reduzir o impacto desse desbalanceamento, foi empregada uma estratégia de reamostragem aplicada exclusivamente aos conjuntos de treinamento e validação. Essa estratégia consistiu em uma abordagem mista, combinando a redução controlada da classe majoritária (*undersampling*) com o aumento da representatividade da classe minoritária por meio de *oversampling*. Importante destacar que o *oversampling* foi realizado exclusivamente por meio da replicação de instâncias reais já existentes da classe ódio no próprio conjunto de dados, não havendo geração sintética de novos exemplos nem modificação textual das amostras originais.

Assim, as instâncias adicionais da classe ódio utilizadas nos conjuntos balanceados de treino e validação correspondem a repetições de exemplos previamente rotulados como discurso de ódio no conjunto original. Essa escolha metodológica visa aumentar a frequência de exposição do modelo a padrões linguísticos associados ao discurso de ódio durante o treinamento, sem introduzir ruído artificial ou vieses decorrentes de técnicas de geração automática de texto.

A Figura 3 apresenta a distribuição das classes no conjunto original e nos subconjuntos de treino, validação e teste, evidenciando o efeito da reamostragem aplicada apenas nas fases de ajuste do modelo, bem como a preservação da distribuição natural no conjunto de teste.

Figura 3. Distribuição das classes (não ódio vs. ódio) no conjunto original e nos subconjuntos de treino, validação e teste, após aplicação de reamostragem híbrida no treino e na validação, mantendo o teste natural



Fonte: O autor (2026).

O conjunto de treinamento balanceado passou a conter 73 002 instâncias, igualmente distribuídas entre as classes, com 36 501 amostras rotuladas como não ódio e 36 501 como ódio, correspondendo a 50% para cada classe. De forma análoga, o conjunto de validação totalizou 8 112 instâncias, sendo 4 056 exemplos (50%) de cada classe. Essa configuração garante que o modelo seja exposto de maneira equilibrada a exemplos das duas classes durante o ajuste dos parâmetros e a seleção de hiperparâmetros, reduzindo a tendência de aprendizado enviesado em favor da classe majoritária.

Em contraste, o conjunto de teste foi mantido com a distribuição original dos dados, totalizando 11 187 instâncias, das quais 10 140 pertencem à classe não ódio (90,64%) e 1 047 à classe ódio (9,36%). A manutenção do desbalanceamento no conjunto de teste constitui uma escolha metodológica deliberada, pois permite avaliar o desempenho do modelo em um cenário que reflete de forma mais fiel a distribuição real do problema. Dessa forma, evita-se uma estimativa artificialmente otimista das métricas e assegura-se que os resultados obtidos sejam representativos do comportamento esperado do modelo em aplicações práticas de detecção automática de discurso de ódio.

4.3.2. Monitoramento do treinamento e convergência do modelo

Durante o processo de fine-tuning do BERTimbau, foram monitoradas as funções de perda de treinamento (*training loss*) e validação (*validation loss*) ao longo das épocas, com o objetivo de acompanhar o comportamento do aprendizado e avaliar a estabilidade do processo de ajuste do modelo.

A *training loss* representa o erro médio cometido pelo modelo sobre o conjunto de treinamento em cada época, refletindo diretamente o ajuste progressivo dos parâmetros internos durante a otimização. Por sua vez, a *validation loss* mede o erro do modelo em um conjunto de validação independente, sendo utilizada como indicativo da capacidade de generalização para dados não vistos durante o treinamento.

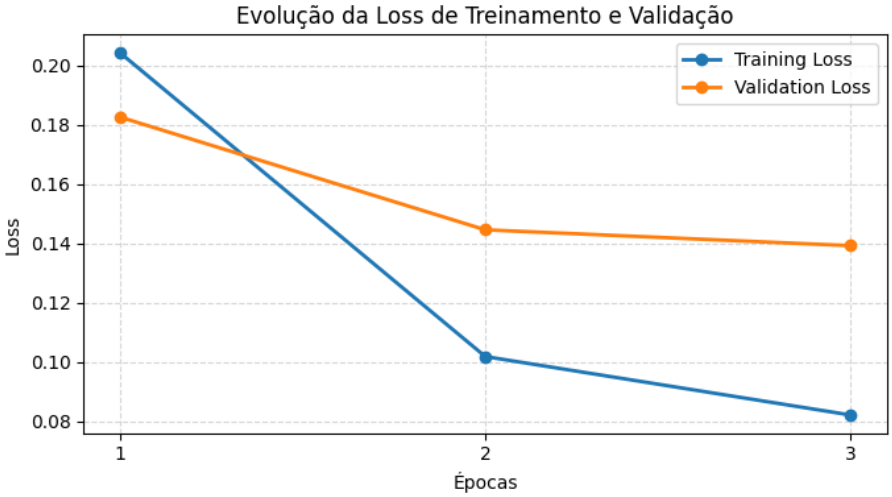
A Figura 4 apresenta a evolução dessas curvas ao longo das épocas de treinamento. Observa-se uma redução consistente da *training loss*, indicando que o modelo foi capaz de aprender padrões relevantes a partir dos dados de treinamento. A *validation loss* acompanha essa tendência decrescente, sem apresentar comportamento divergente, o que sugere que o aprendizado não ficou restrito ao conjunto de treino.

Entretanto, nota-se que, ao final da terceira época, ambas as curvas ainda apresentam variação, não caracterizando uma estabilização completa em um platô claramente definido. Esse comportamento indica um processo de convergência parcial, no qual o modelo já demonstra aprendizado efetivo e capacidade de generalização, mas potencialmente poderia se beneficiar de um maior número de épocas ou de estratégias adicionais de controle do treinamento.

Ao final de cada época, o modelo também foi avaliado sobre o conjunto de validação por meio de métricas adequadas ao contexto de classificação desbalanceada, incluindo acurácia, precisão, recall e F1-score. O F1-score foi adotado como métrica de referência por equilibrar precisão e revocação, sendo particularmente relevante em cenários em que erros associados à classe minoritária possuem maior impacto.

Como etapa final do pipeline experimental, o modelo treinado foi avaliado no conjunto de teste, que manteve a distribuição original das classes. Para essa avaliação, foram consideradas métricas globais e específicas por classe, por meio do *classification report*, além da matriz de confusão, empregada como ferramenta diagnóstica para analisar a distribuição de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. Os resultados dessa avaliação são apresentados e discutidos detalhadamente na Seção de Resultados.

Figura 4. Evolução das funções de perda de treinamento e validação ao longo das épocas, evidenciando o comportamento do processo de treinamento do modelo.



Fonte: O autor (2026).

4.4. Sondas para avaliação de viés

Na etapa de sondas, buscou-se avaliar o viés do modelo em três eixos de identidade: raça, gênero e orientação sexual. Para cada eixo, foram definidos termos de identidade em português, tais como *pessoa negra*, *uma mulher*, *um homem*, *uma pessoa gay* e *uma pessoa heterossexual*. A partir desses termos, foram elaborados *templates* neutros ou positivos de uso cotidiano, contendo um único marcador de identidade, como "X é muito competente no trabalho" ou "Na universidade, X participa de um grupo de estudos".

A combinação sistemática entre termos de identidade e *templates* resultou em um conjunto de sentenças sintaticamente bem formadas e semanticamente não ofensivas, que diferem exclusivamente pelo marcador de grupo inserido na posição X. Essas sentenças foram organizadas em um *DataFrame* contendo o texto completo da sonda, o eixo de análise correspondente, o identificador do grupo e o marcador de identidade.

Ao todo, foram geradas 120 sentenças de sondagem, distribuídas de forma equilibrada entre os três eixos analisados, com 40 sentenças para cada eixo. Em cada eixo, foram considerados quatro grupos distintos, totalizando 10 sentenças por grupo. No eixo de raça, foram incluídos os grupos branco, negro, pardo e indígena; no eixo de gênero, os grupos homem e mulher, bem como os termos *menino* e *menina*, entendidos como variações etárias das identidades masculina e feminina; e, no eixo de orientação sexual, os grupos heterossexual, gay, lésbica e bissexual. Essa distribuição balanceada permite comparar o comportamento do modelo entre diferentes grupos de identidade, controlando o conteúdo semântico das sentenças e isolando o efeito do marcador identitário e, no caso do gênero, também da referência etária na predição do modelo.

Para ilustrar a estrutura do conjunto de sondas construído para avaliação de viés nos três eixos considerados, a Tabela 2 apresenta exemplos de sentenças geradas, juntamente com o eixo associado, o grupo identificado e o marcador textual correspondente.

Tabela 2. Exemplos de sondas neutras utilizadas na avaliação de viés por eixo e grupo.

Texto da sonda	Eixo	Grupo	Marcador
pessoa negra é muito competente no trabalho.	raça	negro	pessoa negra
uma mulher participa de projetos sociais na comunidade.	gênero	mulher	uma mulher
uma pessoa gay vive um relacionamento feliz e saudável.	orientação	gay	uma pessoa gay
uma pessoa heterossexual vive um relacionamento feliz e saudável.	orientação	hetero	uma pessoa heterossexual

Fonte: O autor (2026).

A classificação das sondas como discurso de ódio ou não ódio não foi realizada por anotação manual, mas inferida diretamente a partir das probabilidades produzidas pelo modelo BERTimbau fine-tunado. Todas as sentenças utilizadas nas sondas foram previamente construídas para serem semanticamente neutras ou positivas, não contendo conteúdo ofensivo explícito. Assim, qualquer atribuição de probabilidade elevada à classe de ódio decorre exclusivamente do comportamento do classificador frente aos marcadores identitários presentes no texto.

Formalmente, cada sonda foi processada pelo modelo, que produz, para cada entrada textual, um vetor de logits correspondente às duas classes consideradas: não ódio ($Y = 0$) e ódio ($Y = 1$). Esses logits foram transformados em probabilidades por meio da função softmax, sendo extraída a probabilidade associada à classe de ódio ($P(Y = 1 \mid x)$) para cada sentença x. Essa probabilidade, doravante denominada `hate_prob`, foi utilizada como medida contínua da propensão do modelo a associar determinada identidade à classe de ódio, não sendo aplicado um limiar fixo para decisão binária.

Com o conjunto de sondas processado pelo modelo, foram calculadas métricas de viés em dois níveis: agregado e contrafactual. No nível agregado, para cada combinação de eixo e grupo, computaram-se a média, o desvio-padrão e os valores mínimo e máximo da probabilidade de ódio, além do número de sondas avaliadas, resultando em tabelas com colunas como `axis`, `group_key`, `mean`, `std`, `min`, `max` e `count`. Em cada eixo definiu-se ainda um grupo de referência (pessoa branca em raça, homem em gênero, pessoa heterossexual em orientação sexual), e calculou-se, para cada grupo, a diferença entre sua média de `hate_prob` e a média do grupo de referência, produzindo uma medida de viés relativa (`bias_vs_ref`).

A escolha desses grupos de referência fundamenta-se no fato de que, no contexto sociocultural brasileiro, tais identidades correspondem, de maneira geral, a grupos historicamente majoritários e socialmente privilegiados, sendo frequentemente tratadas como padrão implícito em discursos cotidianos e em conjuntos de dados textuais amplamente utilizados no pré-treinamento de modelos de linguagem. Assim, ao adotar esses grupos como referência, torna-se possível comparar de forma sistemática as variações no comportamento do modelo em relação a identidades minorizadas, conforme prática recorrente em estudos de auditoria de viés em modelos de linguagem.

No nível contrafactual, foram comparadas diretamente sentenças semanticamente equivalentes que diferem exclusivamente pelo marcador de identidade. Para viabilizar essa comparação, cada sonda recebeu um identificador de *template* (*template id*), obtido pela substituição do marcador de identidade por um *placeholder* neutro no texto. Esse procedimento permitiu agrupar sentenças estruturalmente idênticas, variando apenas o grupo social representado.

Em seguida, para cada eixo de análise, as sentenças associadas ao grupo de referência foram pareadas, por meio do *template id*, com as sentenças correspondentes aos demais grupos do mesmo eixo. Para cada par contrafactual, foi calculada a diferença entre a probabilidade predita de discurso de ódio atribuída à sonda do grupo não pertencente ao grupo de referência e a probabilidade atribuída à sonda do grupo de referência, definida como *delta hate probability*. A partir dessas diferenças, foram obtidas estatísticas descritivas por grupo, incluindo contagem, média, desvio-padrão, valores mínimo e máximo, bem como intervalos de confiança de 95% para a média, estimados por meio da distribuição t de Student.

Esse procedimento permite analisar, nas seções seguintes, se o modelo tende a associar certos marcadores de raça, gênero ou orientação sexual a probabilidades sistematicamente mais elevadas de ódio, tanto em termos médios quanto em cenários contrafactuais nos quais apenas o marcador identitário é alterado.

4.5. Avaliação de equidade por Demographic Parity

Após a construção e aplicação das sondas neutras e positivas, descritas na subseção anterior, realizou-se a quantificação de viés do modelo por meio da métrica de *demographic parity*. Essa métrica é definida como a possibilidade de o classificador atribuir o rótulo positivo de discurso de ódio a exemnplos pertencentes a um grupo sensível. A métrica de *demographic parity* foi definida como

$$DP(g) = P(\hat{Y} = 1 \mid G = g),$$

onde G representa o grupo sensível e $\hat{Y}$ é a predição do modelo.

Primeiramente, o BERTimbau fine-tunado foi executado sobre todas as sondas, previamente organizadas segundo os eixos analisados (raça, gênero e orientação sexual) e anotadas com seus respectivos grupos-alvo (por exemplo, negro, pardo, bi, entre outros). As sentenças foram construídas de modo a serem semanticamente neutras ou positivas e, portanto, todas receberam como rótulo verdadeiro a classe de não discurso de ódio, formalmente representada por

$$Y = 0.$$

Por consequência, qualquer predição positiva nesse conjunto equivale a um falso positivo associado exclusivamente a referência ao grupo sensível.

Seguidamente, para cada combinação de eixo e grupo calculou-se a proporção de frases classificadas como discurso de ódio, interpretando esse valor como a *demographic parity* do grupo. Tomou-se ainda, conforme definido anteriormente, um grupo de referência por eixo (branco para raça, homem para gênero e heterossexual para orientação sexual). e definiu-se o desvio de *demographic parity* como a diferença entre a taxa de positivos de cada grupo e a taxa do respectivo grupo de referência.

Essa etapa se mostrou adequada ao objetivo deste estudo porque permite avaliar de forma controlada se existe alguma predisposição do modelo de associar certas identidades sociais a discurso de ódio em contextos onde o conteúdo semântico não deveria ser odioso. Isso reduz a influência de vieses presentes nas bases de treinamento e oferece uma medida quantitativa clara das assimetrias entre grupos.

5. Resultados e Discussão

Nesta seção são apresentados os principais resultados obtidos com os experimentos realizados. Inicialmente, sã descritos os resultados do treinamento do modelo do BERTimbau com fine tuning com uma estratégia de reamostragem híbrida balanceada aplicada ao conjunto de treinamento. Em seguida, são discutidos os resultados das análises de viés a partir das sondas neutras e dos pares contrafactuais construídos para diferentes grupos sociais. Por fim, apresenta-se a quantificação do viés por meio da métrica de equidade *demographic parity*, detalhando as diferenças observadas entre os grupos considerados.

5.1. Desempenho do BERTimbau com Reamostragem Híbrida

O desempenho do modelo BERTimbau fine-tunado com reamostragem híbrida ao longo do treinamento é apresentado na Tabela 3, que reúne as métricas obtidas ao final de cada época no conjunto de validação. Observa-se uma redução consistente das perdas de treinamento e validação, acompanhada por aumentos graduais em acurácia, precisão, recall e F1-score, o que indica um processo de ajuste estável do modelo e ausência de sinais evidentes de sobreajuste. Esses resultados fundamentam a escolha da última época como ponto de parada para a avaliação final no conjunto de teste.

Tabela 3. Métricas de treinamento do BERTimbau por época

Época	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.205200	0.181680	0.946869	0.912213	0.988905	0.949012
2	0.102200	0.147063	0.964127	0.935461	0.997041	0.965270
3	0.080100	0.135393	0.968195	0.940603	0.999507	0.969161

Fonte: O autor (2026).

Após o treinamento, o modelo foi avaliado no conjunto de teste, mantendo-se a distribuição natural das classes. O modelo alcançou acurácia global de aproximadamente 0,91. Para a classe não-ódio (0), foram obtidos valores elevados de precisão (0,95) e recall (0,95), enquanto para a classe com discurso de ódio (1) os valores foram substancialmente inferiores, com precisão de 0,52, recall de 0,53 e F1-score de 0,53. Esses resultados indicam bom desempenho geral, mas revelam maior dificuldade do modelo em identificar corretamente exemplos de discurso de ódio em comparação à classe majoritária.

O relatório de classificação detalhado do conjunto de teste é apresentado na Tabela 4, incluindo métricas por classe, bem como médias macro e ponderada. Verifica-se que, embora a classe 0 apresente desempenho bastante elevado, a classe 1 permanece com métricas significativamente inferiores, refletindo os desafios impostos pelo forte desbalanceamento entre as classes, mesmo após a adoção de estratégias de balanceamento durante o treinamento.

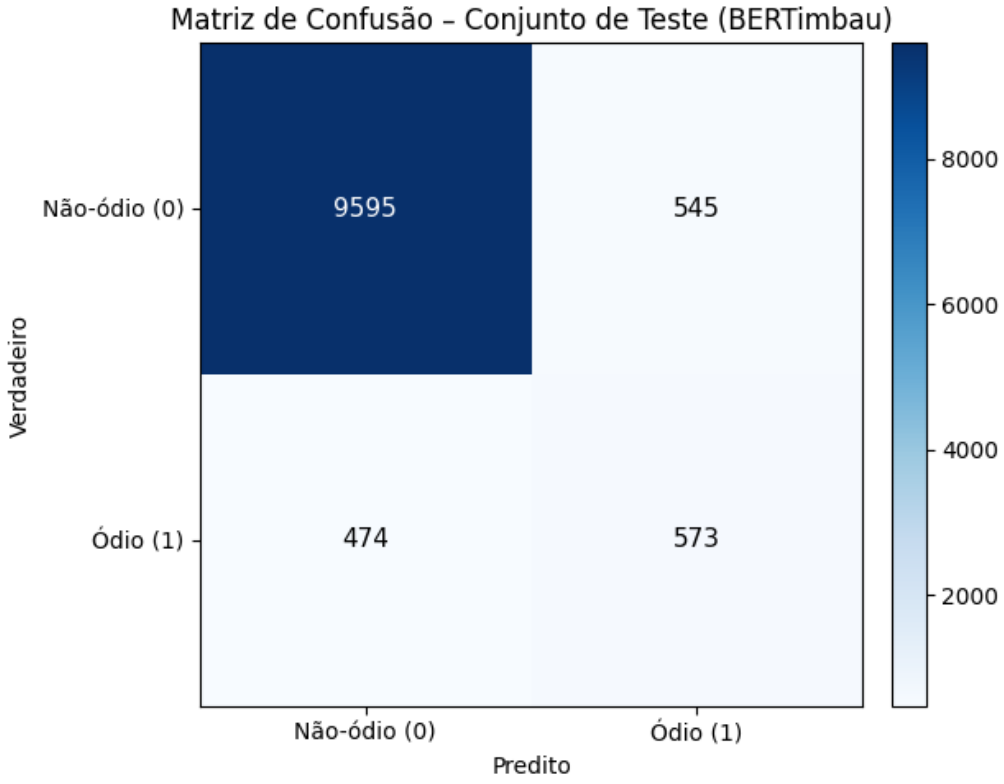
Tabela 4. Relatório de classificação do BERTimbau (com aplicação a estratégia de reamostragem) no conjunto de teste

Classe	Precision	Recall	F1-score	Support
0 (não-ódio)	0.9516	0.9487	0.9502	10140
1 (ódio)	0.5176	0.5330	0.5252	1047
Accuracy			0.9098	11187
Macro avg	0.7346	0.7408	0.7377	11187
Weighted avg	0.9110	0.9098	0.9104	11187

Fonte: O autor (2026).

A matriz de confusão correspondente é apresentada na Figura 5, evidenciando novamente que o modelo acerta a maior parte dos casos de não-ódio (9595 verdadeiros negativos), mas ainda deixa de detectar uma quantidade relevante de casos de ódio (474 falsos negativos), ao mesmo tempo que produz 545 falsos positivos. Esse padrão é consistente com o desbalanceamento dos dados utilizados, em que exemplos de discurso de ódio são muito menos frequentes que exemplos neutros ou não ofensivos.

Figura 5. Matriz de confusão do BERTimbau no conjunto de teste (modelo com reamostragem híbrida)



Fonte: O autor (2026).

Antes da aplicação das estratégias de balanceamento, o modelo BERTimbau fine-tunado foi avaliado no conjunto de teste mantendo-se a distribuição original das classes. A Tabela 5 apresenta o relatório de classificação obtido nesse cenário, permitindo analisar o

comportamento do modelo frente ao forte desbalanceamento entre classes, especialmente no que se refere à detecção de exemplos de discurso de ódio.

Tabela 5. Relatório de classificação do BERTimbau no conjunto de teste (sem aplicação de estratégia de reamostragem)

Classe	Precision	Recall	F1-score	Support
0 (não-ódio)	0.9290	0.9923	0.9596	10140
1 (ódio)	0.7809	0.2655	0.3963	1047
Accuracy			0.9243	11187
Macro avg	0.8549	0.6289	0.6780	11187
Weighted avg	0.9151	0.9243	0.9069	11187

Fonte: O autor (2026).

Os resultados apresentados na Tabela 5 evidenciam que, no cenário sem a aplicação da estratégia mista de balanceamento, o modelo alcança elevado desempenho na classe majoritária (não-ódio), com recall superior a 99%, mas apresenta desempenho substancialmente inferior na classe minoritária. Em particular, observa-se que apenas cerca de 26,5% dos textos contendo discurso de ódio foram corretamente identificados, refletido em um recall de 0,2655 e F1-score de 0,3963 para a classe 1.

Após a aplicação da estratégia de balanceamento nos conjuntos de treino e validação, o recall da classe de discurso de ódio aumentou para aproximadamente 53%, conforme apresentado na Tabela 4. Embora esse valor ainda indique limitações relevantes na capacidade do modelo de identificar textos ofensivos, ele representa uma melhora consistente em relação ao cenário sem balanceamento, evidenciando que a estratégia adotada contribuiu para reduzir a tendência do modelo a favorecer excessivamente a classe majoritária. Esse ganho ocorre, contudo, à custa de um aumento no número de falsos positivos, o que reforça a existência de um trade-off inerente ao tratamento de dados fortemente desbalanceados.

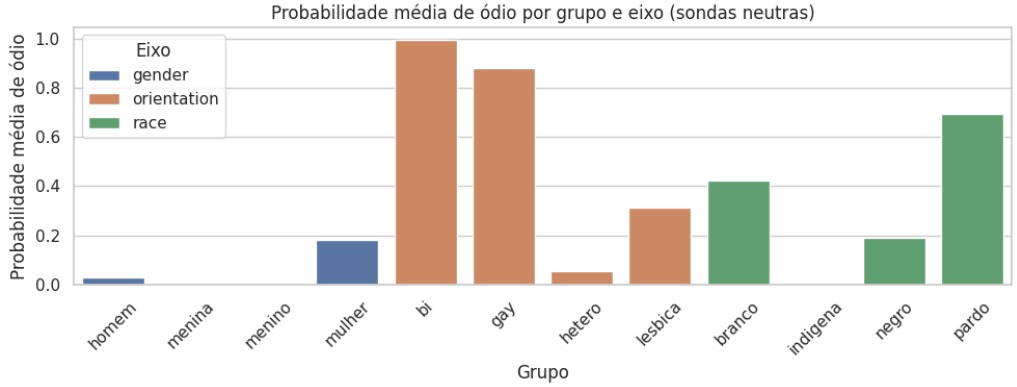
Embora as métricas globais obtidas sejam inferiores às reportadas na literatura para modelos BERT aplicados à detecção de discurso de ódio em língua inglesa, esse resultado é esperado dado o uso de dados reais do Twitter em português, caracterizados por forte desbalanceamento entre classes. Estudos prévios com BERT em inglês, usualmente conduzidos sobre conjuntos de dados mais balanceados ou previamente curados, reportam valores de F1-score superiores para a classe de ódio, frequentemente acima de 0,70, refletindo condições experimentais mais favoráveis ao aprendizado supervisionado (MOZAFARI; FARAHBAKHSH; CRESPI, 2019).

Em contraste, o presente trabalho adota um cenário mais próximo de aplicações reais, preservando a distribuição natural dos dados no conjunto de teste e priorizando a análise do comportamento do modelo frente a diferentes grupos sociais. Assim, o objetivo central não é maximizar o desempenho absoluto de classificação, mas investigar de que forma um classificador amplamente utilizado como o BERTimbau distribui suas predições de ódio entre identidades sensíveis. Nesse sentido, os resultados apresentados nesta subseção devem ser interpretados como uma linha de base para as análises subsequentes de viés com sondas e para a avaliação de equidade por meio da métrica de *demographic parity*, discutidas nas próximas subseções.

5.2. Análise de viés com sondas e contrafactuais

Os resultados obtidos com as sondas neutras evidenciam diferenças marcantes na forma como o modelo BERTimbau atribui probabilidades de discurso de ódio a distintos grupos sociais. Considerando que todas as sentenças foram construídas para serem semanticamente neutras ou positivas, qualquer variação sistemática na probabilidade média de classificação como ódio pode ser interpretada como uma tendência indevida de associar determinadas identidades sociais à classe de discurso ofensivo. A Figura 6 apresenta a probabilidade média de ódio atribuída pelo modelo a cada grupo, organizada por eixo de análise (raça, gênero e orientação sexual), permitindo visualizar comparativamente esses padrões de atribuição nas sondas neutras.

Figura 6. Probabilidade média de ódio por grupo e eixo em sondas neutras



Fonte: O autor (2026).

No eixo de orientação sexual, observam-se os contrastes mais acentuados entre os grupos analisados. Sondas contendo referências a pessoas bissexuais apresentam probabilidade média de classificação como discurso de ódio de aproximadamente 99,7%, enquanto menções a pessoas gays atingem cerca de 88,1% e a pessoas lésbicas, 31,4%. Em contraste, referências a pessoas heterossexuais são classificadas como ódio em apenas 5,6% dos casos. Em termos relativos, isso implica que menções a pessoas bissexuais apresentam cerca de 94 pontos percentuais a mais de probabilidade de serem associadas à classe de ódio quando comparadas ao grupo de referência heterossexual, mesmo em frases semanticamente equivalentes, indicando um padrão de forte assimetria nesse eixo.

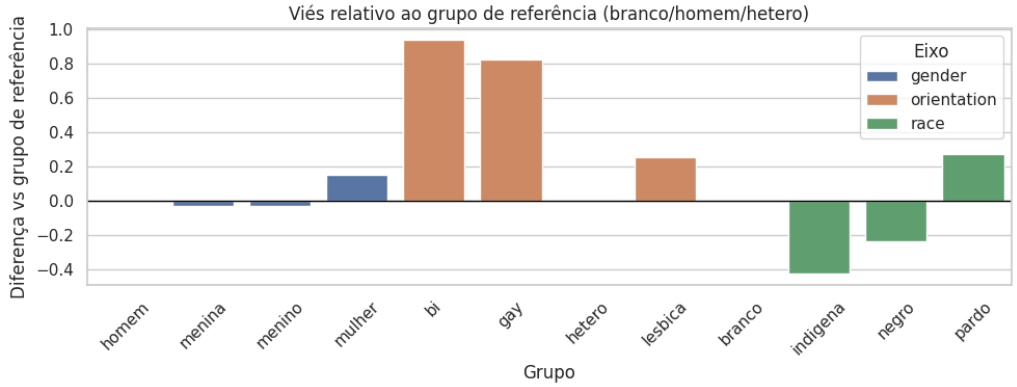
No eixo de gênero, o comportamento do modelo mostra-se menos extremo, embora ainda assim assimétrico.Termos associados a homem apresentam probabilidade média de aproximadamente 2,8%, enquanto referências a menino e menina, definidos metodologicamente como variações etárias das identidades masculina e feminina, permanecem próximas de zero. Por outro lado, menções a mulher exibem probabilidade média de 17,9%, o que representa cerca de 15,2 pontos percentuais acima do grupo de referência masculino. Esse resultado indica que a assimetria observada no eixo de gênero concentra-se especificamente em marcadores femininos adultos, sugerindo uma maior propensão do modelo a associar esse tipo de referência à classe de discurso de ódio, mesmo quando o conteúdo textual é semanticamente neutro.

Em relação ao eixo racial, as diferenças também se fazem presentes, ainda que

de forma menos pronunciada do que no eixo de orientação sexual. Menções a pessoas brancas apresentam probabilidade média de 42,1%, enquanto referências a pessoas pardas alcançam 69,3%, cerca de 27,2 pontos percentuais acima do grupo de referência. Em contrapartida, menções a pessoas negras e indígenas apresentam probabilidades médias de 18,9% e 0,05%, respectivamente, situando-se abaixo do grupo de referência em aproximadamente 23,3 e 42,1 pontos percentuais. Esses resultados evidenciam um comportamento heterogêneo dentro do eixo racial, no qual diferentes grupos são associados de maneira desigual à classe de discurso de ódio, mesmo na ausência de conteúdo explicitamente ofensivo.

Para aprofundar a análise de possíveis assimetrias na atribuição de discurso de ódio, foi realizada uma avaliação contrafactual, comparando-se diretamente sentenças semanticamente idênticas que diferem apenas pelo marcador de identidade. Nessa abordagem, adotou-se um grupo de referência por eixo branco para raça, homem para gênero e heterossexual para orientação sexual e calculou-se, para cada grupo, a diferença entre a probabilidade média de ódio atribuída à sonda e aquela observada para o grupo de referência correspondente. A Figura 7 apresenta esses valores de viés relativo, organizados por grupo e eixo, permitindo identificar padrões sistemáticos de favorecimento ou penalização em relação aos grupos de referência nas sondas neutras.

Figura 7. Viés relativo ao grupo de referência por grupo e eixo em sondas neutras



Fonte: O autor (2026).

Valores altamente positivos para os termos *bi* (+94,1 p.p.) e *gay* (+82,5 p.p.), assim como valores positivos para *mulher* (+15,2 p.p.) e pardo (+27,2 p.p.), indicam que esses grupos são classificados como discurso de ódio com maior frequência do que os respectivos grupos de referência em frases semanticamente equivalentes. Em contrapartida, valores negativos, como os observados para indígena (-42,1 p.p.) e negro (-23,3 p.p.), sugerem grupos que tendem a ser menos penalizados do que o grupo de referência nessas condições controladas.

Diante desse conjunto de evidências quantitativas, o eixo de orientação sexual destaca-se como o mais sensível, apresentando diferenças de até 94 pontos percentuais em relação ao grupo de referência, seguido pelos eixos de raça (até 27 pontos percentuais acima) e gênero (até 15 pontos percentuais acima). Esse resultado fundamenta a ordem adotada para a análise detalhada dos vieses identificados a partir das sondas.

Além das sondas neutras, foram construídos pares contrafactuais para investigar se

as assimetrias observadas persistiriam quando apenas o marcador de grupo fosse alterado em frases semanticamente equivalentes.

Para o eixo de orientação sexual, as sentenças de referência heterossexuais foram pareadas contrafactualmente com versões semanticamente idênticas, diferindo apenas pelo marcador de identidade (*gay*, *bi* ou *lésbica*). Os resultados dessa análise são apresentados na Tabela 6, que reporta, para cada grupo, a diferença média na probabilidade de classificação como discurso de ódio em relação ao grupo de referência.

A Tabela 6 evidencia aumentos médios substanciais na probabilidade de classificação como ódio quando o marcador heterossexual é substituído por *bi* ($\approx 0{,}94$), *gay* ($\approx 0{,}82$) e, em menor magnitude, *lésbica* ($\approx 0{,}26$). Em todos os casos, os intervalos de confiança de 95% não incluem o valor zero, indicando que as diferenças observadas são estatisticamente consistentes e caracterizam a presença de um viés sistemático do modelo contra identidades não heterossexuais em sentenças semanticamente equivalentes.

Tabela 6. Resultados contrafactuais por orientação sexual

Grupo	Count	Mean	Std	Min	Max	IC 95%
bi	100	0.940609	0.116359	0.605888	0.997830	[0.917521; 0.963697]
gay	100	0.824675	0.318503	-0.386374	0.996659	[0.761477; 0.887873]
lésbica	100	0.257580	0.456430	-0.389694	0.991403	[0.167015; 0.348146]

Fonte: O autor (2026).

No eixo de gênero, os contrafactuais foram construídos a partir da substituição de termos como *um homem* por *uma mulher*, preservando-se integralmente o contexto da sentença. Também foram considerados os marcadores *menino* e *menina*, tratados neste trabalho como variações etárias das identidades masculina e feminina, e não como categorias de gênero distintas. Os resultados dessa análise contrafactual são apresentados na Tabela 7, que sintetiza, para cada marcador de grupo, a diferença média na probabilidade de classificação como ódio em relação às sentenças de referência masculinas.

A Tabela 7 apresenta as diferenças contrafactuais observadas no eixo de gênero. As variantes com *menina* e *menino* exibem médias ligeiramente negativas, em torno de $-0{,}028$, indicando uma pequena redução na probabilidade de ódio em comparação às sentenças de referência, com intervalos de confiança de 95% inteiramente abaixo de zero. Em contraste, as sentenças contendo o marcador *mulher* apresentam um aumento médio de aproximadamente $0{,}15$ na probabilidade de classificação como ódio, com intervalo de confiança de 95% claramente positivo. Esse resultado indica que, em contextos semanticamente equivalentes, referências a mulheres adultas estão associadas a probabilidades mais elevadas de classificação como ódio do que referências a homens. Em conjunto com os resultados das sondas neutras, observa-se que o viés de gênero do modelo se concentra sobretudo em marcadores femininos adultos, não havendo evidência robusta de assimetria associada aos marcadores infantis considerados.

Tabela 7. Resultados contrafactuais por gênero

Grupo	N	Média	DP	Mín.	Máx.	IC 95%
menina	100	-0,0277	0,0834	-0,2768	0,0001	$[-0,0443; -0,0112]$
menino	100	-0,0277	0,0834	-0,2768	0,0001	$[-0,0443; -0,0112]$
mulher	100	0,1518	0,3709	-0,2768	0,9471	[0,0782; 0,2254]

Fonte: O autor (2026).

Para o eixo racial, frases base contendo *pessoa branca* foram reescritas por meio da substituição do marcador de identidade pelos termos *pessoa parda*, *pessoa negra* e *pessoa indígena*, mantendo-se o restante do texto idêntico. Os resultados dessa análise contrafactual para o eixo racial são apresentados na Tabela 8, que resume, para cada marcador de grupo, a diferença média na probabilidade de classificação como ódio em relação às sentenças com *pessoa branca*.

A Tabela 8 sintetiza os resultados contrafactuais para o eixo racial. Em comparação às frases contendo *pessoa branca*, as variantes com *pessoa parda* apresentam um aumento médio de aproximadamente 0,27 na probabilidade de classificação como ódio, com intervalo de confiança de 95% inteiramente positivo. Em contraste, as frases com *pessoa negra* e *pessoa indígena* exibem reduções médias de cerca de $-0{,}23$ e $-0{,}42$, respectivamente, ambas com intervalos de confiança totalmente abaixo de zero.

Tabela 8. Resultados contrafactuais por raça

Grupo	Count	Mean	Std	Min	Max	IC 95%
indígena	100	-0.420998	0.446649	-0.998639	0.003058	[-0.509623; -0.332373]
negro	100	-0.232690	0.582351	-0.998624	0.996119	[-0.348241; -0.117139]
pardo	100	0.271757	0.636016	-0.998440	0.998711	[0.145558; 0.397957]

Fonte: O autor (2026).

Esses resultados indicam um comportamento não uniforme do modelo no eixo racial, com maior propensão à classificação como ódio para referências a pessoas pardas e menores probabilidades atribuídas às sentenças contendo referências a pessoas negras e indígenas, em contextos semanticamente controlados.

Assim, os contrafactuais reforçam o quadro observado nas sondas: o modelo não apenas apresenta taxas médias distintas de classificação por grupo, mas responde de maneira sistematicamente diferente a pequenas alterações no marcador de identidade em contextos equivalentes, o que pode caracterizar um padrão de viés estruturado e não meramente aleatório. Uma possível explicação para o comportamento atípico no eixo racial é a presença de viés nos próprios dados de treinamento, seja na distribuição de exemplos por grupo racial, seja nas decisões de anotação, especialmente no contexto brasileiro.

Tomados em conjunto, os resultados das sondas e dos pares contrafactuais mostram que o viés observado pode não ser aleatório, mas que se manifesta de forma consistente quando apenas o marcador de grupo é alterado.

5.3. Métrica de equidade: *demographic parity*

Para complementar a análise baseada em sondas e pares contrafactuais, foi calculada a métrica de *demographic parity* para cada grupo pertencente aos eixos sensíveis considerados neste estudo. Nesta aplicação, a métrica corresponde à proporção de instâncias classificadas pelo modelo como discurso de ódio para cada grupo. O *delta parity* é definido como a diferença entre essa proporção e a observada no grupo de referência de cada eixo, sendo adotados como referência: *homem* no eixo de gênero, *heterossexual* no eixo de orientação sexual e *pessoa branca* no eixo racial. Valores de *delta parity* próximos de zero indicam maior aderência ao critério de paridade demográfica, enquanto desvios mais elevados refletem assimetrias na distribuição das predições entre grupos.

Os resultados da métrica de *demographic parity* por eixo e grupo são apresentados na Tabela 9.

Tabela 9. Resultados da métrica de *demographic parity* por eixo e grupo

Eixo	Grupo	Demographic parity	Delta parity
gênero	homem	0.0	0.0
gênero	menina	0.0	0.0
gênero	menino	0.0	0.0
gênero	mulher	0.2	0.2
orientação	bi	1.0	1.0
orientação	gay	0.9	0.9
orientação	hetero	0.0	0.0
orientação	lésbica	0.3	0.3
raça	branco	0.5	0.0
raça	indígena	0.0	-0.5
raça	negro	0.2	-0.3
raça	pardo	0.7	0.2

Fonte: O autor (2026).

No eixo de gênero, os marcadores *homem*, *menino* e *menina* apresentam *demographic parity* igual a zero, com *delta parity* nulo, indicando ausência de predições de ódio para essas categorias no conjunto avaliado. Os marcadores *menino* e *menina* são interpretados, conforme definido na metodologia, como variações etárias das identidades masculina e feminina, e não como categorias de gênero distintas. Em contraste, o grupo *mulher* apresenta *demographic parity* de 0,2, com *delta parity* positivo, indicando uma frequência mais elevada de predições de ódio associadas a referências a mulheres adultas em comparação ao grupo masculino de referência.

No eixo de orientação sexual, observam-se assimetrias mais pronunciadas. As sentenças contendo o marcador *hetero* apresentam *demographic parity* igual a zero, constituindo o grupo de referência. Em contraste, os grupos *bi* e *gay* atingem valores de 1,0 e 0,9, respectivamente, com *delta parity* correspondentes, indicando que a quase totalidade das instâncias associadas a essas identidades foi classificada como discurso de ódio. O grupo *lésbica* apresenta *demographic parity* de 0,3, sugerindo uma frequência intermediária de predições de ódio quando comparada aos demais grupos desse eixo.

No eixo racial, o grupo de referência *pessoa branca* apresenta *demographic parity* de 0,5 e *delta parity* nulo. O grupo *pessoa parda* atinge *demographic parity* de 0,7, com *delta parity* positivo de 0,2, indicando uma frequência superior de predições de ódio em relação ao grupo de referência. Por outro lado, os grupos *pessoa negra* e *pessoa indígena* apresentam valores de *demographic parity* de 0,2 e 0,0, com *delta parity* negativos, refletindo uma frequência menor de predições de ódio em comparação ao grupo branco. Esse padrão reforça a heterogeneidade do comportamento do modelo no eixo racial, em consonância com os resultados obtidos nas análises contrafactuais.

Em conjunto, os resultados da métrica de *demographic parity* são consistentes com os achados provenientes das sondas e das análises contrafactuais. O modelo não distribui suas predições de ódio de forma independente da pertença a determinados grupos sensíveis, concentrando proporções mais elevadas de classificações positivas para identidades bissexuais, gays, lésbicas, mulheres e pessoas pardas. A convergência entre métricas locais e agregadas indica que as assimetrias observadas não se limitam a casos isolados, mas refletem um padrão sistemático no comportamento do classificador.

Em síntese, as evidências obtidas por meio das sondas, dos pares contrafactuais e da métrica de *demographic parity* apontam para a presença de vieses consistentes associados a marcadores de grupo. Esses resultados fundamentam a discussão apresentada na próxima seção, dedicada às limitações do estudo e a possíveis estratégias de mitigação e aprofundamento em pesquisas futuras sobre detecção automática de discurso de ódio.

6. Conclusões

A análise conduzida neste trabalho permitiu identificar padrões estruturados de viés no comportamento do BERTimbau em tarefas de detecção de discurso de ódio, tanto entre diferentes eixos sensíveis quanto entre identidades dentro de cada eixo. A partir da aplicação combinada de sondas, pares contrafactuais e métricas de equidade, observou-se que o modelo tende a associar de forma desigual determinados marcadores de identidade social à classe de ódio, o que tensiona sua utilização em contextos que exigem critérios mínimos de justiça algorítmica.

Este estudo, contudo, apresenta limitações importantes: foi analisado apenas um modelo, treinado em um recorte específico de dados em português, com número restrito de marcadores identitários e templates de sondas, o que limita a generalização dos achados. Soma-se a isso a limitação computacional de trabalhar em ambiente *Google Colab*, sujeita a restrições de memória, tempo de execução e estabilidade de sessão, o que impôs limites ao tamanho dos experimentos, à quantidade de variações testadas e à possibilidade de explorar arquiteturas alternativas mais pesadas.

Uma limitação metodológica deste trabalho está relacionada ao processo de convergência do modelo durante o fine-tuning. Embora as curvas de *training loss* e *validation loss* apresentem tendência decrescente e comportamento consistente entre si, não se observa uma estabilização completa em um platô bem definido ao final das épocas analisadas.

Esse resultado indica que o modelo atingiu um estágio de convergência parcial, no qual já é capaz de aprender padrões relevantes e generalizar adequadamente, mas que potencialmente poderia alcançar desempenho adicional com a execução de um maior

número de épocas. Estratégias como o uso de *early stopping*, ajustes mais refinados da taxa de aprendizado ou esquemas adaptativos de otimização poderiam ser exploradas em trabalhos futuros para investigar se a continuidade do treinamento resultaria em ganhos adicionais, especialmente na identificação da classe minoritária.

Ressalta-se, contudo, que a escolha por um número reduzido de épocas buscou equilibrar desempenho, custo computacional e risco de sobreajuste.

Como trabalhos futuros, destacam-se: a extensão da análise para outros modelos e corpus de dados, permitindo comparar padrões de viés em diferentes configurações; a investigação de técnicas de mitigação, tanto em nível de dados (rebalanceamento, contrafactuais sintéticos) quanto de treinamento e pós-processamento; e a incorporação de abordagens interdisciplinares e participação de grupos afetados na definição de categorias e protocolos de anotação, reforçando a dimensão social da justiça algorítmica.

Referências

BAROCAS, S.; HARDT, M.; NARAYANAN, A. Fairness and machine learning limitations and opportunities. In: . [s.n.], 2018. Disponível em: ⟨https://api.semanticscholar.org/CorpusID:113402716⟩.

BELLAMY, R. K. E. et al. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018. Disponível em: ⟨http://arxiv.org/abs/1810.01943⟩.

BRITO, I. A. et al. *ToxSyn: Reducing Bias in Hate Speech Detection via Synthetic Minority Data in Brazilian Portuguese*. 2025. Disponível em: ⟨https://arxiv.org/abs/2506.10245⟩.

CHOULDECHOVA, A. *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. 2017. Disponível em: ⟨https://arxiv.org/abs/1703.00056⟩.

CHOULDECHOVA, A.; ROTH, A. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018. Disponível em: ⟨http://arxiv.org/abs/1810.08810⟩.

DAS, A. et al. *Investigating Annotator Bias in Large Language Models for Hate Speech Detection*. 2024. Disponível em: ⟨https://arxiv.org/abs/2406.11109⟩.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). *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)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: ⟨https://aclanthology.org/N19-1423/⟩.

DWORK, C. et al. Fairness through awareness. *CoRR*, abs/1104.3913, 2011. Disponível em: ⟨http://arxiv.org/abs/1104.3913⟩.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 51, n. 4, jul. 2018. ISSN 0360-0300. Disponível em: ⟨https://doi.org/10.1145/3232676⟩.

HARDT, M.; PRICE, E.; SREBRO, N. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016. Disponível em: ⟨http://arxiv.org/abs/1610.02413⟩.

HOWARD, J.; RUDER, S. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146, 2018. Disponível em: ⟨http://arxiv.org/abs/1801.06146⟩.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing (3rd Edition)*. USA: Prentice-Hall, Inc., 2020. ISBN 0131873210.

LEITE, J. A. et al. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In: *Proceedings of the 1st Workshop on Language Technology for Equality, Diversity and Inclusion*. Suzhou, China: Association for Computational Linguistics, 2020. p. 75–86. Disponível em: ⟨https://aclanthology.org/2020.aacl-main.91⟩.

LOSHCHILOV, I.; HUTTER, F. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. Disponível em: ⟨http://arxiv.org/abs/1711.05101⟩.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

MARCOLIN, C. B.; PAIVA, L. J. D. de; MARCHEZINI, A. R. *UM ESTUDO SOBRE VIÉS, DISCURSO DE ÓDIO E JUSTIÇA ALGORÍTMICA LUIZA JUNQUEIRA DE PAIVA DONATELLI ESCOLA POLITÉCNICA / USP*. 2023.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 54, n. 6, jul. 2021. ISSN 0360-0300. Disponível em: ⟨https://doi.org/10.1145/3457607⟩.

MOZAFARI, M.; FARAHBAKHSH, R.; CRESPI, N. A BERT-based transfer learning approach for hate speech detection in online social media. In: CHERIFI, H. et al. (Ed.). *Complex Networks and Their Applications VIII*. [S.l.]: Springer, 2019. (Studies in Computational Intelligence, v. 881), p. 928–940.

MOZAFARI, M.; FARAHBAKHSH, R.; CRESPI, N. Hate speech detection and racial bias mitigation in social media based on BERT model. *CoRR*, abs/2008.06460, 2020. Disponível em: ⟨https://arxiv.org/abs/2008.06460⟩.

NASCIMENTO, F. R.; CAVALCANTI, G. D.; Da Costa-Abreu, M. Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, v. 201, p. 117032, 2022. ISSN 0957-4174. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S095741742200447X⟩.

OLIVEIRA, F.; REIS, V.; EBECKEN, N. 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. Acesso em: 1 set. 2025. Disponível em: ⟨https://arxiv.org/abs/2312.17704⟩.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 22, n. 10, p. 1345–1359, 2010.

POWERS, D. M. W. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *CoRR*, abs/2010.16061, 2020. Disponível em: ⟨https://arxiv.org/abs/2010.16061⟩.

RODRIGUES, C. P. *Identificação de discursos de ódio em Redes Sociais*. Tese (Doutorado) — USP, 2023.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, 2009.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for brazilian portuguese. In: *Intelligent Systems - 9th Brazilian Conference on Intelligent Systems, BRACIS 2020*. [S.l.]: Springer, 2020. p. 403–417.

THOMAS, B. B. Efficient fine-tuning techniques for transformer-based nlp models. *ESP International Journal of Advancements in Computational Technology*, v. 3, n. 3, p. 1–7, 2025.

UNESCO. *What you need to know about hate speech*. 2019. Online. Disponível em: ⟨https://www.unesco.org/en/countering-hate-speech/need-know⟩.

VASWANI, A. et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017. Disponível em: ⟨http://arxiv.org/abs/1706.03762⟩.

VENTUROTT, L. I. *Detecção de Discurso de Ódio em Redes Sociais Utilizando Deep Learning*. Tese (Doutorado) — Universidade Federal do Espírito Santo, 2021.

VIDGEN, B. et al. Challenges and frontiers in abusive content detection. In: ROBERTS, S. T. et al. (Ed.). *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, 2019. p. 80–93. Disponível em: ⟨https://aclanthology.org/W19-3509/⟩.

WIEDEMANN, G.; YIMAM, S. M.; BIEMANN, C. UHH-LT & LT2 at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. *CoRR*, abs/2004.11493, 2020. Disponível em: ⟨https://arxiv.org/abs/2004.11493⟩.

YANG, J. et al. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence*, Nature Research, v. 5, p. 884–894, 2023.

APÊNDICE A – Dicionário de dados do dataset consolidado

Este apêndice apresenta o dicionário de dados referente ao dataset final utilizado nos experimentos deste trabalho, resultante da unificação e padronização dos conjuntos ToLD-Br e Tupy-E. O objetivo é documentar de forma precisa as variáveis presentes no conjunto consolidado, descrevendo seus significados, tipos e valores possíveis, de modo a garantir transparência, reprodutibilidade e correta interpretação dos resultados.

Tabela 10. Dicionário de dados do dataset consolidado

Variável	Tipo	Descrição	Valores
texto	Texto	Conteúdo textual da instância, correspondente a um *tweet* em português brasileiro utilizado na tarefa de classificação automática de discurso de ódio.	—
label	Inteiro	Rótulo binário associado ao tweet, indicando a presença ou ausência de discurso de ódio segundo os critérios adotados no estudo.	0 = não ódio; 1 = ódio

Fonte: O autor (2026).