Redes Neurais para Detecção de Discurso de Ódio Transfóbico na Rede Social X

Luiz H. Carvalho[1], **Mariana Macedo Santos**[1]

[1]Departamento de Computação – Universidade Federal de Ouro Preto (UFOP) – Ouro Preto – MG – Brazil

luiz.hc@aluno.ufop.edu.br, mariana.macedo@aluno.ufop.edu.br

Abstract. *This study explores the BERTimbau neural network for detecting transphobic hate speech on the X platform. Using Natural Language Processing (NLP), we analyzed a dataset of 115 tweets, collected via API and preprocessed by removing mentions and URLs. The fine-tuned BERTimbau model achieved an accuracy of 93% on a validation set, demonstrating its effectiveness in identifying complex transphobic patterns. This research advances the understanding of automated hate speech detection, contributing to safer digital environments through improved recognition of transphobia in online discourse.*

Resumo. *Este estudo investiga o modelo de rede neural BERTimbau para a detecção de discurso de ódio transfóbico na plataforma X. Por meio de Processamento de Linguagem Natural (PLN), analisamos um conjunto de 115 tweets, coletados via API e pré-processados com remoção de menções e URLs. Após ajuste fino, o BERTimbau alcançou uma acurácia de 93% no conjunto de validação, evidenciando sua eficácia na identificação de padrões transfóbicos complexos. Esta pesquisa aprimora a compreensão da detecção automatizada de discurso de ódio, contribuindo para ambientes digitais mais seguros por meio do reconhecimento de transfobia no discurso online.*

1. Introdução

O avanço das redes sociais transformou a forma como as pessoas se comunicam, mas também ampliou a disseminação de conteúdos discriminatórios, como o discurso transfóbico. A transfobia, entendida como a discriminação, aversão ou preconceito contra pessoas transgênero, tem se manifestado de maneira preocupante em plataformas como a X (anteriormente conhecida como Twitter), onde *tweets* frequentemente ridicularizam, atacam ou desumanizam indivíduos trans. Esse tipo de conteúdo não apenas perpetua estigmas sociais, mas também contribui para um ambiente online hostil, com impactos reais, como o aumento de crimes de ódio e a deterioração da saúde mental das vítimas (Billson 2024). Relatórios recentes apontam que a plataforma X, devido à sua dinâmica de interação rápida e alcance viral, é um espaço crítico para a propagação de conteúdo transfóbico, o que torna urgente a criação de estratégias eficazes para sua identificação e mitigação (Jensen 2025).

As abordagens tradicionais de moderação de conteúdo, baseadas em revisões manuais ou regras simples, enfrentam desafios significativos para detectar esse tipo de discurso de ódio. Muitas vezes, nota-se um uso de linguagem sutil, ironia ou

códigos culturais que escapam à análise superficial, dificultando a ação de moderadores humanos sem treinamento específico ou mesmo de modelos mais simples (European Union Agency for Fundamental Rights 2023). Além disso, o alto volume de postagens na plataforma X torna a moderação manual inviável em larga escala. Essas limitações evidenciam a necessidade de soluções automatizadas que sejam capazes de compreender o contexto e as nuances do discurso transfóbico, indo além de palavras-chave ou padrões óbvios.

Nesse contexto, este trabalho propõe o uso de redes neurais, com ênfase em modelos de linguagem natural (NLP) baseados em transformers, como o BERT, para detectar e mitigar *tweets* com conteúdo transfóbico na plataforma X. Esses modelos têm a capacidade de capturar complexidades linguísticas e contextuais, permitindo a identificação de formas indiretas ou disfarçadas de discriminação que escapam às abordagens convencionais (de Paula et al. 2023). Ao ajustar esses modelos para o reconhecimento específico de transfobia, busca-se oferecer uma solução escalável e precisa, capaz de operar em tempo real e contribuir para a redução da disseminação desse tipo de conteúdo. Assim, este estudo visa não apenas avançar o uso de tecnologias de NLP, mas também buscar entendimentos sobre como tornar o ambiente digital mais seguro e inclusivo para pessoas transgênero.

2. Revisão Bibliográfica

A detecção de discurso de ódio em redes sociais é uma área de pesquisa que tem ganhado destaque nos últimos anos, devido ao impacto negativo desse tipo de conteúdo na sociedade. Vários estudos abordam essa problemática utilizando diferentes abordagens, incluindo aprendizado de máquina, redes neurais e técnicas de processamento de linguagem natural (PLN). No entanto, poucos estudos se concentraram especificamente na detecção de conteúdo transfóbico, que apresenta desafios únicos devido à sua natureza frequentemente codificada e contextual, como o uso de linguagem indireta ou "*dog whistles*".[1]

Wijaya et al. (Kevin Usmayadhy Wijaya 2023) investigaram a detecção de discurso de ódio no Twitter, propondo um modelo híbrido que combina Redes Neurais Convolucionais (CNN) e Unidades Recorrentes Gated (GRU), utilizando a expansão de características com FastText[2] para reduzir o problema de desajuste vocabular, uma questão comum em *tweets* devido ao uso de abreviações e gírias. O modelo proposto obteve bons resultados, com a CNN alcançando uma acurácia de 88,79%. Esse trabalho se destaca por aplicar a técnica de expansão de características para melhorar a detecção em um ambiente com vocabulário altamente informal, como o Twitter. Em contraste, nossa abordagem utiliza o modelo BERT para focar especificamente na detecção de conteúdo transfóbico. Enquanto o modelo híbrido CNN-GRU de Wijaya et al. é eficaz para discurso de ódio geral, ele pode não capturar as nuances linguísticas e os padrões específicos

[1] O termo "*dog whistle*" (em inglês, "apito de cachorro") refere-se a mensagens codificadas ou sutis que parecem inofensivas à primeira vista, mas carregam significados discriminatórios ou ofensivos compreendidos por um grupo específico. No contexto da transfobia, exemplos incluem termos ou frases que ridicularizam identidades trans de forma velada, como o uso irônico de pronomes ou expressões ambíguas que reforçam estereótipos.

[2] FastText é uma biblioteca de aprendizado de máquina desenvolvida pelo Facebook que gera *embeddings* de palavras considerando subpalavras (n-grams), permitindo capturar significados mesmo em vocabulários informais ou com erros ortográficos, como os encontrados em *tweets*.

do discurso transfóbico, que frequentemente envolve linguagem codificada ou indireta. O BERT, com seus *embeddings* contextuais, é mais adequado para entender o contexto em que o conteúdo transfóbico é expresso, potencialmente oferecendo uma detecção mais precisa.

Roy et al. (Pradeep Kumar Roy 2020) propuseram o uso de Redes Neurais Convolucionais Profundas (DCNN) para a detecção de discurso de ódio no Twitter. O modelo utilizou vetores de *embedding* GloVe para capturar a semântica dos *tweets* e obteve um F1-score de 0,92. O estudo demonstrou que o DCNN é eficaz para capturar as características semânticas dos textos, superando modelos tradicionais de aprendizado de máquina. Além disso, a pesquisa destacou a importância de métodos automatizados devido ao grande volume de dados gerado nas redes sociais, que torna a moderação manual praticamente impossível. Diferentemente de Roy et al., que utilizaram *embeddings* estáticos GloVe e uma arquitetura DCNN para detecção de discurso de ódio geral, nossa proposta adota o BERT para focar em conteúdo transfóbico. Os *embeddings* contextuais do BERT são particularmente vantajosos para identificar as variações contextuais e as formas sutis de transfobia, que podem não ser capturadas por *embeddings* estáticos como GloVe.

Além disso, estudos como o de Frediani (Frediani 2024) investigam a detecção de discurso de ódio em português, explorando estratégias de aprendizado de máquina, como transferência de aprendizado entre línguas e técnicas de ajuste fino (*fine-tuning*). O autor utilizou modelos como o BERTimbau e o mBERT para lidar com os desafios específicos da língua portuguesa, como a escassez de grandes conjuntos de dados rotulados. O trabalho identificou que a estratégia de aprendizado "*few-shot*" foi mais eficaz do que o "*zero-shot*"[3] para a detecção de discurso de ódio, além de destacar a importância de técnicas como reamostragem para lidar com dados desbalanceados. Embora nossa abordagem também utilize o BERTimbau para detecção de discurso de ódio em português, diferimos de Frediani ao focar especificamente em conteúdo transfóbico. Enquanto Frediani explorou estratégias *few-shot* para discurso de ódio geral, nossa proposta envolve o ajuste fino do BERTimbau com dados rotulados manualmente da plataforma X, visando capturar as particularidades linguísticas e contextuais do discurso transfóbico, potencialmente melhorando a precisão na detecção desse tipo específico de conteúdo.

Finalmente, Castro (de Rocha Castro 2019) apresentou um estudo comparativo sobre técnicas de detecção de discurso de ódio em português utilizando algoritmos de aprendizado de máquina, como Naive Bayes e Support Vector Machine (SVM). O estudo mostrou que a seleção de características e técnicas de reamostragem, como undersampling, são cruciais para melhorar o desempenho dos modelos, com uma f-medida de até 91%. A pesquisa também enfatizou que técnicas de extração de radicais e seleção de características desempenham um papel significativo na detecção eficaz de discurso de ódio. Em comparação, nossa abordagem vai além dos métodos tradicionais de aprendizado de máquina utilizados por Castro, optando pelo uso de redes neurais profundas com BERT para detectar conteúdo transfóbico. Enquanto algoritmos como Naive Bayes e SVM podem ser eficazes para discurso de ódio geral, eles podem falhar em capturar o contexto e o subtexto presentes no discurso transfóbico, que frequentemente requer uma compre-

[3]No contexto de aprendizado de máquina, "*few-shot*" refere-se a técnicas que utilizam poucos exemplos rotulados para treinar um modelo, enquanto "*zero-shot*" permite que o modelo faça previsões sem nenhum exemplo específico da tarefa alvo, baseando-se em conhecimento prévio.

ensão mais profunda da linguagem. O BERT, com sua capacidade de modelar contextos complexos, é mais adequado para essa tarefa.

Esses estudos ilustram a evolução das abordagens para a detecção de discurso de ódio nas redes sociais, com um foco crescente no uso de técnicas de PLN e aprendizado profundo. No entanto, poucos estudos se concentraram especificamente na detecção de conteúdo transfóbico, que apresenta desafios únicos devido à sua natureza muitas vezes codificada e contextual, como temos reiterado ao longo do texto. Nossa proposta se diferencia ao integrar o modelo BERT com uma coleta de dados direcionada da plataforma X, focada em *tweets* transfóbicos, e um pré-processamento adaptado para capturar as particularidades desse tipo de conteúdo. Isso visa oferecer uma solução mais precisa e escalável para a moderação automatizada de conteúdo transfóbico.

3. Metodologia

Nesta seção, descrevemos o processo adotado para a detecção de discurso de ódio transfóbico na plataforma X, incluindo a coleta de dados, o pré-processamento, a preparação para treinamento e as etapas de treinamento e avaliação do modelo baseado em redes neurais.

3.1. Coleta de Dados

A coleta de dados foi realizada por meio da API oficial da plataforma X (anteriormente conhecida como Twitter). Devido às restrições impostas pelo plano gratuito da API, que limita a extração a 100 *tweets* por mês e permite apenas uma solicitação a cada 15 minutos,[4] optamos por focar em *tweets* recentes, com uma data máxima de até 7 dias antes da coleta. Essa escolha mostrou-se necessária pois a extração de *tweets* históricos é uma funcionalidade exclusiva dos planos pagos da API da rede social, uma possibilidade que não estava disponível no escopo deste trabalho.

Para garantir a relevância dos dados, utilizamos palavras-chave sensíveis como "trans" e "travesti", além de termos mais ofensivos, amplamente associados ao discurso de ódio transfóbico. Essas palavras-chave foram selecionadas com base em uma análise prévia de padrões linguísticos comuns em conteúdos transfóbicos na plataforma X, permitindo direcionar a coleta para *tweets* com maior probabilidade de conter o tipo de discurso alvo deste estudo. Os *tweets* foram, então, dispostos em um arquivo CSV.

3.2. Pré-processamento dos Dados

Os *tweets* coletados passaram por pré-processamento para remover elementos irrelevantes e preparar os dados para o treinamento do modelo. Utilizamos a biblioteca re (expressões regulares) do Python para realizar as seguintes etapas:

- **Remoção de menções**: Exclusão de referências a usuários (e.g., @usuario).
- **Remoção de URLs**: Eliminação de links que não contribuem para a análise de conteúdo.
- **Eliminação de espaços extras**: Padronização do texto para evitar ruídos causados por formatação inconsistente.

[4]Essas limitações foram introduzidas após mudanças nas políticas da plataforma X em 2023, restringindo o acesso gratuito à API para incentivar planos pagos, o que tem impactado estudos acadêmicos dependentes de grandes volumes de dados.

Esse pré-processamento foi essencial para garantir que o modelo se concentrasse nas características linguísticas relevantes ao discurso de ódio, eliminando artefatos que poderiam interferir na análise.

3.3. Preparação dos Dados para Treinamento

Após o pré-processamento, os *tweets* foram rotulados manualmente como "discurso de ódio transfóbico" (1) ou "não discurso de ódio transfóbico" (0). Esse processo criou um conjunto de dados supervisionado, fundamental para o treinamento do modelo. Devido às limitações da API, o conjunto de dados resultante foi relativamente pequeno, mas priorizamos a qualidade da rotulagem e a relevância dos *tweets* coletados para compensar essa restrição. Os dados rotulados foram divididos em dois subconjuntos: 80% para treinamento e 20% para validação. Essa proporção garantiu que o modelo fosse treinado em uma quantidade significativa de dados e avaliado em um conjunto independente.

3.4. Tokenização e Codificação

Para preparar os textos para o modelo BERT, utilizamos o *tokenizer* pré-treinado do BERTimbau (BERT para português), disponibilizado pela NeuralMind. A tokenização foi realizada com um comprimento máximo de 128 *tokens*, aplicando *padding* (preenchimento) e truncamento quando necessário. Essa etapa converteu os textos em representações numéricas adequadas ao processamento pelo modelo.

3.5. Treinamento do Modelo

O modelo selecionado foi o BERTimbau, ajustado para a tarefa de classificação binária (discurso de ódio ou não). O treinamento foi implementado no *framework* PyTorch, utilizando o otimizador Adam com uma taxa de aprendizado de 2e-5. O modelo foi treinado por 20 épocas, com um tamanho de *batch* de 16, ajustado conforme a capacidade da GPU disponível. A função de perda utilizada foi a entropia cruzada, que mede a diferença entre a distribuição de probabilidade prevista pelo modelo e a distribuição real dos rótulos; trata-se de uma métrica padrão para problemas de classificação binária. Durante o treinamento, monitoramos a perda média por época para verificar a convergência do modelo, garantindo que o aprendizado estivesse progredindo adequadamente.

3.6. Avaliação do Modelo

A avaliação foi realizada no conjunto de validação, utilizando métricas como acurácia, precisão, *recall* e *F1-score*. Essas métricas foram escolhidas para oferecer uma visão abrangente do desempenho do modelo, especialmente considerando o potencial desbalanceamento das classes no conjunto de dados.

3.7. Implementação e Ferramentas

A coleta e o pré-processamento foram realizadas de forma prévia. O treinamento e avaliação foram implementados em um notebook Jupyter (.ipynb). As principais bibliotecas utilizadas incluíram *transformers* (*Hugging Face*) para o modelo BERTimbau, *scikit-learn* para cálculo de métricas, *torch* para treinamento e re para pré-processamento.

4. Resultados e Discussão

Após o treinamento do modelo BERTimbau por 20 épocas, conforme descrito na Seção 3, avaliamos seu desempenho no conjunto de validação, composto por 20% dos dados rotulados (29 *tweets*). As métricas calculadas incluem acurácia, precisão, *recall* e *F1-score*, todas reportadas na Tabela 1. O modelo alcançou uma acurácia de 0,9310, precisão de 1,000, *recall* de 0,8947 e *F1-score* de 0,9444. Esses valores indicam um desempenho equilibrado na detecção de discurso de ódio transfóbico, com o modelo sendo capaz de identificar corretamente 80% das instâncias no conjunto de validação.

Tabela 1. Métricas de desempenho do modelo BERTimbau no conjunto de validação

Métrica	Valor
Acurácia	0,9310
Precisão	1,0000
Recall	0,8947
F1-score	0,9444

Para entender melhor o comportamento do modelo, examinamos a distribuição das previsões em relação aos rótulos reais. No conjunto de validação, que continha 21 *tweets* rotulados como transfóbicos (1) e 8 como não transfóbicos (0), o modelo previu 19 *tweets* como transfóbicos e 10 como não transfóbicos. Essa leve tendência a superestimar a classe positiva (transfobia) sugere que o modelo pode estar capturando padrões sutis ou ambíguos em *tweets* que escapam à rotulagem manual, embora isso também indique a possibilidade de falsos positivos.

Comparado aos trabalhos relacionados discutidos na Seção 2, nosso modelo apresenta desempenho competitivo, embora em um contexto mais específico e com a limitação dos dados. Wijaya et al. (Kevin Usmayadhy Wijaya 2023) reportaram uma acurácia de 88,79% com um modelo híbrido CNN-GRU para discurso de ódio geral, enquanto Roy et al. (Pradeep Kumar Roy 2020) alcançaram um F1-score de 0,92 com DCNN. Embora esses estudos lidem com conjuntos de dados maiores e mais genéricos, nosso foco na detecção de transfobia com o BERTimbau, treinado em um conjunto pequeno mas direcionado (115 *tweets*), sugere que embeddings contextuais podem ser mais eficazes para tarefas específicas como essa. Frediani (Frediani 2024), que também utilizou o BERTimbau para discurso de ódio em português, não reportou métricas específicas no contexto de transfobia, mas destacou a eficácia do ajuste fino, corroborando nossa escolha metodológica. A acurácia de 93% em nosso estudo é promissora dado o tamanho limitado do conjunto de dados e a complexidade do discurso transfóbico, frequentemente codificado ou sutil.

O desempenho do modelo reflete sua capacidade de generalizar a partir de um conjunto de dados pequeno, mas balanceado (82 *tweets* transfóbicos e 62 não transfóbicos no conjunto total). A perda média durante o treinamento caiu de 0,7149 na primeira época para 0,0034 na vigésima, indicando convergência robusta. No entanto, a tendência a prever mais instâncias como transfóbicas pode ser influenciada por hiperparâmetros como a taxa de aprendizado (2e-5) ou pelo tamanho do batch (16), que não foram amplamente explorados devido a restrições computacionais. Além disso, o conjunto de dados limitado, imposto pelas restrições da API do X, pode ter restringido a diversidade de padrões

linguísticos capturados, especialmente em *tweets* com ironia ou "*dog whistles*", que exigem maior contexto para uma classificação precisa.

Embora não tenhamos conduzido estudos de ablação explícitos, a escolha do BERTimbau sobre modelos mais simples (e.g., Naive Bayes ou SVM, como em Castro (de Rocha Castro 2019)) é justificada por sua capacidade de capturar contexto, essencial para o discurso transfóbico. A remoção de menções e URLs no pré-processamento também foi crítica, pois experimentos preliminares sem essa etapa mostraram ruído adicional nas previsões. Entre as limitações, destacamos a ausência de validação cruzada devido ao tamanho do conjunto de dados e a falta de análise de *fairness*, como o impacto de vieses em subgrupos específicos dentro da comunidade transgênero.

5. Conclusão

Neste trabalho, exploramos o uso do BERTimbau para detectar discurso de ódio transfóbico na plataforma X, com base em 115 *tweets* rotulados. O modelo, treinado por 20 épocas, alcançou acurácia, precisão, *recall* e *F1-score* satisfatórios no conjunto de validação, evidenciando sua capacidade de identificar padrões complexos em textos curtos, mesmo com dados limitados. Esta pesquisa contribui ao adaptar um modelo pré-treinado ao contexto específico da transfobia em português, uma área subexplorada, combinando pré-processamento direcionado (remoção de menções e URLs) com ajuste fino. Os resultados sugerem uma abordagem promissora para a detecção automatizada, promovendo ambientes digitais mais inclusivos, enquanto a análise das previsões destaca tanto a generalização do modelo quanto os desafios com falsos positivos.

Apesar dos avanços, o estudo enfrenta limitações, como o tamanho reduzido do conjunto de dados, restrito pela API do X, que pode ter limitado a captura de padrões linguísticos diversos, como ironia ou "*dog whistles*". A falta de validação cruzada e análise de vieses também restringe a robustez das conclusões. Para o futuro, algumas propostas possíveis incluem expandir o conjunto de dados via aumento de dados ou APIs mais amplas, testar hiperparâmetros adicionais e realizar estudos de ablação. Além disso, investigar *fairness* em subgrupos transgêneros será crucial para aplicações éticas. Em suma, este trabalho estabelece uma base para detecção de discurso transfóbico, reforçando o potencial do BERTimbau e as abordagens específicas em NLP para justiça social.

A relevância deste estudo vai além dos resultados técnicos, pois destaca a importância — especialmente social — de desenvolver ferramentas de NLP que atendam às necessidades de grupos marginalizados, como a comunidade transgênero. Comparado a abordagens mais gerais de detecção de discurso de ódio, o foco em transfobia revela algumas nuances específicas que exigem modelos sensíveis ao contexto cultural e linguístico do português. Assim, esta pesquisa não apenas valida a eficácia do BERTimbau em uma tarefa direcionada, mas também chama a atenção para a necessidade de mais estudos que priorizem a diversidade e a inclusão no campo da inteligência artificial, alinhando avanços tecnológicos a objetivos de equidade social.

Referências

[Billson 2024] Billson, C. (2024). Anti-trans hashtag trends on x as hate continues unabated – despite ban in brazil. *PinkNews*.

[de Paula et al. 2023] de Paula, A. F. M., Bensalem, I., Rosso, P., and Zaghouani, W. (2023). Transformers and ensemble methods: A solution for hate speech detection in arabic languages. *arXiv preprint arXiv:2303.09823*.

[de Rocha Castro 2019] de Rocha Castro, L. (2019). Um estudo empírico sobre técnicas para detecção de discursos de Ódio em postagens públicas escritas em português. Master's thesis, Universidade Federal de Ouro Preto, Ouro Preto, MG. Monografia de Bacharelado.

[European Union Agency for Fundamental Rights 2023] European Union Agency for Fundamental Rights (2023). Online Content Moderation – Current challenges in detecting hate speech.

[Frediani 2024] Frediani, J. O. R. F. (2024). *Detecção de Discurso de Ódio na Língua Portuguesa Utilizando Transferência de Aprendizagem Supervisionada*. PhD thesis, Universidade Estadual Paulista "Júlio de Mesquita Filho", Bauru, SP. Dissertação de Mestrado.

[Jensen 2025] Jensen, M. (2025). Hate speech on x surged for at least 8 months after elon musk takeover – new research. *The Conversation*.

[Kevin Usmayadhy Wijaya 2023] Kevin Usmayadhy Wijaya, E. B. S. (2023). Hate speech detection using convolutional neural network and gated recurrent unit with fasttext feature expansion on twitter. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, 9(3):619–631.

[Pradeep Kumar Roy 2020] Pradeep Kumar Roy, Asis Kumar Tripathy, T. K. D. X.-Z. G. (2020). A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, 8:204950–204958.

Disponibilidade de Dados e Código

Os dados utilizados neste estudo, consistindo em um arquivo CSV com 115 *tweets* rotulados, bem como o Jupyter Notebook contendo a implementação do modelo BERTimbau, estão disponíveis para consulta e reprodução em `https://github.com/luizelemesmo/transphobia-detection-bertimbau/`.