INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PICOS
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

MARIA EMÍLIA CARVALHO LEITE

DETECÇÃO DE DISCURSO TÓXICO EM REDES SOCIAIS BRASILEIRAS: UM ESTUDO COMPARATIVO ENTRE MODELOS CLÁSSICOS DE APRENDIZADO DE MÁQUINA E MODELOS DE LINGUAGEM PRÉ-TREINADOS

PICOS, PIAUÍ
Agosto de 2026

MARIA EMÍLIA CARVALHO LEITE

DETECÇÃO DE DISCURSO TÓXICO EM REDES SOCIAIS BRASILEIRAS: UM ESTUDO COMPARATIVO ENTRE MODELOS CLÁSSICOS DE APRENDIZADO DE MÁQUINA E MODELOS DE LINGUAGEM PRÉ-TREINADOS

Trabalho de Conclusão de Curso (Artigo Científico) apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

Orientador: Prof. Dr. Rogério Figueiredo de Sousa

PICOS, PIAUÍ
Agosto de 2026

MARIA EMÍLIA CARVALHO LEITE

DETECÇÃO DE DISCURSO TÓXICO EM REDES SOCIAIS BRASILEIRAS: UM ESTUDO COMPARATIVO ENTRE MODELOS CLÁSSICOS DE APRENDIZADO DE MÁQUINA E MODELOS DE LINGUAGEM PRÉ-TREINADOS

Trabalho de Conclusão de Curso (Artigo Científico) apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

Aprovado em 12/08/2026.

BANCA EXAMINADORA:

Prof. Dr. Rogério Figueiredo de Sousa(Orientador)
Instituto Federal do Piauí (IFPI)

Prof. M^{e}. Israel Cardoso Araujo
Instituto Federal do Piauí (IFPI)

Prof. M^{e}. Anthony Irlan Marques Luz
Instituto Federal do Piauí (IFPI)

PICOS, PIAUÍ
Agosto de 2026

DETECÇÃO DE DISCURSO TÓXICO EM REDES SOCIAIS BRASILEIRAS: UM ESTUDO COMPARATIVO ENTRE MODELOS CLÁSSICOS DE APRENDIZADO DE MÁQUINA E MODELOS DE LINGUAGEM PRÉ-TREINADOS

RESUMO

A disseminação de comentários tóxicos em redes sociais representa um desafio para a moderação de conteúdo digital, especialmente no português brasileiro, idioma marcado pelo uso expressivo de palavrões e gírias que tornam ambígua a distinção automática entre ofensa e informalidade comunicativa. Este trabalho apresenta um estudo comparativo entre quatro abordagens de classificação automática de toxicidade, incluindo TF-IDF com Regressão Logística, TF-IDF com Máquina de Vetores de Suporte (SVM), BERTimbau e XLM-RoBERTa, os dois últimos submetidos a ajuste fino (*fine-tuning*) sobre o ToLD-BR, conjunto de dados composto por 21.000 comentários do Twitter anotados quanto à toxicidade. Os modelos foram avaliados pelas métricas de Acurácia, F1-Macro, Precisão e Recall. Os resultados indicaram que os modelos *Transformer* superaram os métodos clássicos em todas as métricas avaliadas, com o BERTimbau alcançando F1-Macro de 0,77 e Acurácia de 0,77, com ganho de 3,37 pontos percentuais sobre o melhor *baseline*. O BERTimbau, modelo monolíngue treinado exclusivamente em português, superou o XLM-RoBERTa multilíngue, sugerindo que a especialização no idioma é fator relevante para a tarefa. A análise qualitativa dos erros revelou que o principal desafio consiste na ambiguidade entre uso ofensivo e uso expressivo de linguagem coloquial no português brasileiro.

Palavras-chave: detecção de toxicidade; processamento de linguagem natural; *transformers*; BERTimbau; português brasileiro.

ABSTRACT

The dissemination of toxic comments on social media platforms represents a significant challenge for digital content moderation, particularly in Brazilian Portuguese, a language characterized by the expressive use of profanity and slang that creates ambiguity in the automatic distinction between genuine offense and communicative informality. This paper presents a comparative study of four automatic toxicity classification approaches, including TF-IDF with Logistic Regression, TF-IDF with Support Vector Machine (SVM), BERTimbau and XLM-RoBERTa, the latter two fine-tuned on the ToLD-BR dataset, comprising 21,000 Twitter comments annotated for toxicity. The models were evaluated using Accuracy, Macro F1-score, Precision and Recall. Results showed that Transformer models outperformed classical methods across all evaluated metrics, with BERTimbau achieving Macro F1-score of 0.77 and Accuracy of 0.77, representing a gain of 3.37 percentage points over the best baseline. BERTimbau, a monolingual model trained exclusively on Brazilian Portuguese, outperformed the multilingual XLM-RoBERTa, suggesting that language specialization is a relevant factor for this task. Qualitative error analysis revealed that the main challenge lies in the ambiguity between offensive and expressive colloquial language use in Brazilian Portuguese. These findings reinforce the importance of language-specific models for automated content moderation.

Keywords: toxicity detection; natural language processing; transformers; BERTimbau; Brazilian Portuguese.

Data de Aprovação: 12/08/2026 (Data de Apresentação do Artigo)

1 INTRODUÇÃO

A expansão das redes sociais transformou os padrões de comunicação e ampliou a produção diária de conteúdo por usuários. Plataformas como Twitter/X, Instagram e YouTube reúnem debates políticos, notícias, entretenimento e conversas pessoais, permitindo que mensagens alcancem grandes públicos em pouco tempo. Essa dinâmica tornou a comunicação mais rápida e participativa, mas também aumentou a quantidade de textos que precisam ser organizados, analisados e moderados. Empresas, instituições públicas e comunidades virtuais passaram a depender desses espaços para manter contato direto com seus públicos.

No Brasil, esse cenário possui dimensão expressiva. Segundo o relatório *Digital 2024* (WE ARE SOCIAL; MELTWATER, 2024), o país conta com mais de 144 milhões de usuários ativos em redes sociais e apresenta elevado nível de participação nessas plataformas. Esse volume torna os ambientes digitais espaços importantes para a circulação de opiniões e para a interação entre pessoas de diferentes regiões e grupos sociais.

O crescimento dessas plataformas também ampliou a circulação de comentários ofensivos, ataques pessoais, discurso de ódio e outras formas de conteúdo prejudicial. A quantidade de publicações ultrapassa a capacidade de revisão humana e exige mecanismos automáticos de apoio à moderação (SCHMIDT; WIEGAND, 2017). No português brasileiro, o problema é agravado pelo uso de palavrões, gírias, ironia e humor. Essas características dificultam a distinção entre uma ofensa real e uma expressão informal usada sem intenção de agredir.

O Processamento de Linguagem Natural tem sido aplicado a esse problema por meio de diferentes técnicas. Métodos baseados em TF-IDF, Regressão Logística e Máquinas de Vetores de Suporte foram referências na classificação de textos (JOACHIMS, 1998). Posteriormente, a arquitetura *Transformer* (VASWANI *et al.*, 2017) e o BERT (DEVLIN *et al.*, 2019) permitiram criar modelos capazes de considerar o contexto das palavras. Para o português, o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) oferece uma alternativa monolíngue, enquanto o XLM-RoBERTa (CONNEAU *et al.*, 2020) atende a vários idiomas. No ToLD-BR, o estudo original concentrou-se na criação do conjunto e na avaliação de métodos AutoML e mBERT (LEITE *et al.*, 2020b), enquanto trabalhos posteriores investigaram grandes modelos de linguagem por meio de instruções (OLIVEIRA *et al.*, 2024). Permanece, portanto, uma lacuna específica na comparação controlada

entre classificadores clássicos, um modelo monolíngue e um modelo multilíngue, todos avaliados na mesma partição e acompanhados por análises de custo de inferência e padrões de erro.

Este trabalho compara TF-IDF com Regressão Logística, TF-IDF com SVM, BERTimbau e XLM-RoBERTa na classificação de comentários tóxicos. Os modelos contextuais receberam ajuste fino no ToLD-BR (LEITE *et al.*, 2020b), conjunto formado por 21.000 comentários do Twitter. A pesquisa verifica se os modelos pré-treinados superam os métodos clássicos, se a especialização no português oferece vantagem sobre uma abordagem multilíngue e quais erros permanecem após o treinamento. As contribuições incluem a comparação dos quatro modelos sob a mesma divisão dos dados, a análise do custo de inferência e o estudo qualitativo de falsos positivos e falsos negativos.

O restante do artigo está organizado da seguinte forma. A seção de Referencial Teórico reúne os conceitos e modelos necessários para compreender o problema. Em seguida, a seção de Trabalhos Relacionados compara pesquisas anteriores com este estudo. A Metodologia detalha o conjunto de dados, o pré-processamento, os modelos e o protocolo experimental. A seção de Resultados e Discussão apresenta as métricas, as matrizes de confusão, os padrões de erro, as ameaças à validade e os aspectos éticos. Por fim, a Conclusão reúne os principais achados e as propostas de trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 TOXICIDADE EM REDES SOCIAIS

O conceito de toxicidade no contexto das redes sociais abrange um conjunto amplo de manifestações de linguagem prejudicial ao ambiente digital, incluindo discurso de ódio (*hate speech*), assédio, ameaças, conteúdo obsceno e ataques a grupos minoritários (FORTUNA; NUNES, 2018). Pesquisas brasileiras têm evidenciado que o fenômeno assume características próprias no idioma, sobretudo pela fluidez entre o uso ofensivo e o uso expressivo de palavrões e gírias na comunicação informal (LEITE *et al.*, 2020b).

Fortuna & Nunes (2018) realizaram uma revisão sistemática sobre detecção de discurso de ódio e apontam que a diversidade de definições adotadas na literatura dificulta a comparação entre sistemas. As principais categorias de conteúdo tóxico tratadas nos estudos incluem racismo, misoginia, homofobia e insultos gerais. No Brasil, Leite *et al.* (2020b) identificaram seis categorias de toxicidade no ToLD-BR, construído a partir de comentários do Twitter, abrangendo desde ofensas diretas até xenofobia e conteúdo obsceno.

A moderação automática de conteúdo tóxico ganhou relevância institucional no Brasil com o Marco Civil da Internet (Lei 12.965/2014) e com as discussões em torno da regulação das plataformas digitais. Nesse contexto, soluções baseadas em aprendizado

de máquina surgem como instrumentos complementares à moderação humana, com potencial para operação em escala (SCHMIDT; WIEGAND, 2017).

2.2 DELIMITAÇÃO CONCEITUAL E DESAFIOS DE ANOTAÇÃO

Toxicidade, linguagem ofensiva e discurso de ódio são conceitos relacionados, mas não equivalentes. A toxicidade funciona como categoria abrangente para mensagens que prejudicam a interação, enquanto o discurso de ódio costuma envolver ataques dirigidos a grupos definidos por características identitárias. A linguagem ofensiva também pode ocorrer sem referência a um grupo social, como em insultos individuais ou expressões obscenas. A ausência de uma taxonomia única faz com que diferentes conjuntos de dados adotem critérios e níveis de granularidade próprios (FORTUNA; NUNES, 2018; VARGAS *et al.*, 2022).

Neste trabalho, a tarefa binária segue a operacionalização do ToLD-BR. Um texto é considerado tóxico quando recebe anotação positiva em pelo menos uma das categorias originais. Essa transformação facilita a comparação entre modelos, porém reúne fenômenos linguísticos distintos sob uma única classe. Insultos explícitos, misoginia, racismo e obscenidade passam a contribuir para o mesmo rótulo, apesar de apresentarem frequências, estruturas e graus de dependência contextual diferentes. O estudo que introduziu o ToLD-BR relata maior dificuldade nas categorias menos frequentes e destaca a necessidade de modelos sensíveis às diferentes formas de toxicidade (LEITE *et al.*, 2020b).

A anotação depende da interpretação humana. Ironia, palavrões usados como intensificadores e referências a acontecimentos externos podem produzir discordância entre avaliadores. Por esse motivo, o rótulo deve ser entendido como uma decisão construída segundo as regras do conjunto de dados, e não como uma propriedade inteiramente objetiva da frase. Essa distinção é importante na leitura dos resultados. Uma previsão divergente do rótulo constitui erro para fins de avaliação, mas pode revelar ambiguidade real no uso da linguagem.

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL E REPRESENTAÇÃO DE TEXTO

O Processamento de Linguagem Natural é um subcampo da inteligência artificial dedicado ao desenvolvimento de métodos computacionais capazes de compreender, interpretar e gerar texto em linguagem humana (JURAFSKY; MARTIN, 2023). Na tarefa de classificação de texto, o primeiro passo consiste em transformar sequências de palavras em representações numéricas processáveis por algoritmos de aprendizado de máquina.

A representação TF-IDF (*Term Frequency-Inverse Document Frequency*) atribui pesos a cada termo em função de sua frequência local e de sua raridade no *corpus*

(SALTON; BUCKLEY, 1988). Apesar de sua simplicidade, tem sido eficaz em tarefas de classificação quando combinada a classificadores lineares. Joachims (1998) demonstraram que o SVM com representações TF-IDF produz resultados competitivos em categorização de textos, estabelecendo esse par como um *baseline* robusto para a área.

2.4 MODELOS *TRANSFORMER* E BERT

Em 2017, Vaswani *et al.* (2017) introduziram a arquitetura *Transformer*, baseada exclusivamente no mecanismo de atenção (*self-attention*), eliminando a necessidade de redes recorrentes para modelagem de sequências. Essa arquitetura possibilitou o treinamento em larga escala e deu origem a modelos de linguagem pré-treinados que transformaram o estado da arte em NLP.

O BERT (*Bidirectional Encoder Representations from Transformers*), proposto por Devlin *et al.* (2019), introduziu o pré-treinamento bidirecional sobre grandes *corpora* de texto, permitindo que o modelo capturasse dependências contextuais em ambas as direções da sentença. Após o pré-treinamento, o modelo pode ser ajustado (*fine-tuned*) para tarefas específicas com relativamente poucos dados rotulados.

Para o português, o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) é o principal modelo monolíngue disponível, pré-treinado em *corpus* de aproximadamente 2,68 bilhões de palavras em português brasileiro provenientes da BrWaC (*Brazilian Web as Corpus*) e da *Wikipedia* em português. Souza, Nogueira & Lotufo (2020) demonstraram que o BERTimbau supera modelos multilíngues em tarefas de reconhecimento de entidades nomeadas e análise de sentimento no idioma. O XLM-RoBERTa (CONNEAU *et al.*, 2020), por sua vez, é um modelo multilíngue treinado em 2,5 terabytes de texto em 100 idiomas, incluindo o português, relevante para comparações com modelos monolíngues.

3 TRABALHOS RELACIONADOS

A detecção automática de linguagem tóxica e ofensiva em português brasileiro tem recebido atenção crescente na comunidade de Processamento de Linguagem Natural. Os trabalhos existentes variam quanto ao tipo de *corpus* utilizado, às plataformas de coleta, aos modelos avaliados e às métricas reportadas. Essas diferenças dificultam comparações diretas e permitem identificar tendências relevantes para o presente estudo.

Leite *et al.* (2020b) propuseram o ToLD-BR, conjunto de dados utilizado neste trabalho, composto por 21.000 comentários do Twitter em português brasileiro anotados segundo seis categorias de toxicidade. Os autores avaliaram *baselines* clássicos sobre o *corpus* e modelos multilíngues, reportando desempenho limitado dos métodos tradicionais frente à diversidade de categorias de toxicidade no idioma. Desde então, o

ToLD-BR serve de referência para estudos posteriores de detecção de toxicidade em português. O objetivo daquele trabalho foi introduzir o recurso, descrever sua anotação e estabelecer resultados iniciais com AutoML e mBERT. O presente estudo parte do *corpus* já publicado e realiza outro recorte experimental, comparando Regressão Logística, SVM, BERTimbau e XLM-RoBERTa sob a mesma partição, além de examinar tempo de inferência e padrões de erro.

Vargas *et al.* (2022) apresentaram o HateBR, primeiro *corpus* de grande escala com anotação especializada para detecção de linguagem ofensiva e discurso de ódio em português brasileiro, composto por 7.000 comentários coletados do Instagram de políticos. A anotação do *corpus* abrange classificação binária (ofensivo/não-ofensivo), nível de ofensividade e nove grupos de discurso de ódio. Os experimentos com modelos BERT monolíngues alcançaram F1-Macro de 0,76 na classificação binária, superando o estado da arte para o português à época.

Trajano, Bordini & Vieira (2024) introduziram o OLID-BR (*Offensive Language Identification Dataset for Brazilian Portuguese*), *corpus* com 7.943 comentários provenientes do YouTube e Twitter, anotado em esquema hierárquico de três camadas compatível com *datasets* em outros idiomas, o que permite o desenvolvimento de modelos multilíngues. Os experimentos com modelos BERT monolíngues atingiram F1-Macro de 0,76 na tarefa binária de identificação de linguagem ofensiva. O OLID-BR é o trabalho relacionado mais próximo ao presente em termos de tarefas avaliadas e modelos utilizados.

Uma linha mais recente de pesquisa passou a comparar modelos treinados especificamente para classificação com grandes modelos de linguagem usados por meio de instruções. Oliveira *et al.* (2024) avaliaram modelos monolíngues e multilíngues no ToLD-BR e observaram que a especialização em português favorece a interpretação de linguagem coloquial. Os autores também registraram resultados competitivos com estratégias de poucos exemplos, embora o custo computacional e a dependência de serviços externos sejam fatores relevantes. O presente estudo mantém o foco em modelos codificadores ajustados diretamente para a tarefa e em classificadores clássicos. Assim, difere de Oliveira *et al.* (2024) tanto pela família de modelos quanto pelo protocolo: os quatro classificadores são avaliados na mesma partição, com métricas, tempo de inferência, matrizes de confusão e análise dos erros.

Outra questão recente envolve a transferência entre plataformas. Moreira *et al.* (2026) analisaram seis conjuntos de toxicidade em português e mostraram que diferenças lexicais, estratégias de coleta e práticas de anotação influenciam a generalização entre domínios. Conjuntos com maior diversidade de vocabulário tóxico apresentaram melhor transferência, enquanto recursos concentrados em palavrões sofreram quedas mais intensas fora do domínio de treinamento. Esse resultado reforça a necessidade de interpretar o desempenho no ToLD-BR como evidência específica para o domínio

analisado.

Em um trabalho experimental com dados reais, Zampieri *et al.* (2019) criaram o OLID a partir de mais de 14 mil mensagens do Twitter em inglês. A anotação foi organizada em três níveis, destinados a identificar a presença de linguagem ofensiva, o tipo da ofensa e o alvo atingido. Os autores também compararam diferentes modelos de aprendizado de máquina sobre o conjunto. Essa estrutura hierárquica serviu de base para adaptações em outros idiomas, entre elas o OLID-BR, e mostra como uma classificação detalhada pode fornecer mais informações do que um único rótulo binário.

O Quadro 1 resume os principais trabalhos relacionados e estabelece uma comparação com o presente estudo quanto ao *corpus*, à plataforma, à abordagem, ao resultado reportado e às diferenças de escopo.

Quadro 1 – Comparativo dos principais trabalhos relacionados

Trabalho	***Corpus*/ idioma**	**Plataforma**	**Abordagem**	**Resultado**	**Diferença para este trabalho**
Zampieri *et al.* (2019)	OLID (14k), inglês	Twitter	Modelos supervisionados	Três tarefas hierárquicas	Outro idioma e classificação em três níveis.
Leite *et al.* (2020b)	ToLD-BR (21k), português	Twitter	*Baselines* e mBERT	F1-Macro 0,72	Introduz o conjunto; não compara BERTimbau e XLM-RoBERTa.
Vargas *et al.* (2022)	HateBR (7k), português	Instagram	BERTimbau	F1-Macro 0,76	Outro conjunto, outra plataforma e anotação em três níveis.
Trajano, Bordini & Vieira (2024)	OLID-BR (7,9k), português	Twitter/ YouTube	BERTimbau	F1-Macro 0,76	Tarefa hierárquica e conjunto diferente do ToLD-BR.
Oliveira *et al.* (2024)	ToLD-BR, português	Twitter	LLMs por instruções	Varia por modelo	Avalia LLMs; este trabalho usa ajuste fino e métodos clássicos.
Moreira *et al.* (2026)	Seis conjuntos, português	Múltiplas	Análise de conjuntos	Varia entre bases	Analisa repositórios; este trabalho compara modelos em uma base controlada.
Este trabalho	**ToLD-BR (21k), português**	**Twitter**	**TF-IDF e *Transformers***	**F1-Macro 0,77**	**Compara custo, métricas e erros sob a mesma partição.**

Fonte: Elaborado pela autora (2026).

O presente estudo confirma achados anteriores sobre a superioridade dos modelos *Transformer* em relação aos clássicos e sobre a vantagem do BERTimbau frente a modelos multilíngues no português brasileiro. Sua contribuição reúne uma comparação sistemática entre quatro abordagens sobre o ToLD-BR, a inclusão do XLM-RoBERTa como segundo modelo *Transformer* e uma análise qualitativa dos padrões de erro associados ao uso expressivo de palavrões no português coloquial.

4 METODOLOGIA

4.1 TIPO DE PESQUISA

Esta pesquisa é de natureza aplicada e quantitativa, com abordagem experimental. O objetivo é avaliar e comparar o desempenho de diferentes abordagens computacionais para classificação de texto, por meio de experimentos controlados e reprodutíveis, com métricas objetivas de avaliação, conforme a caracterização de pesquisas quantitativas apresentada por Creswell (2014).

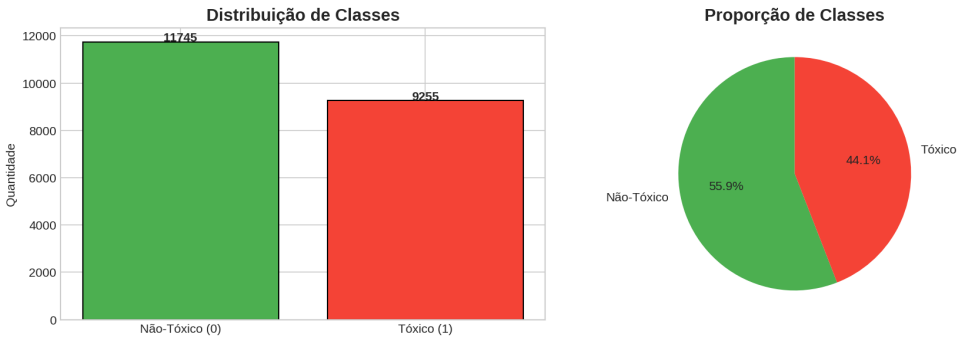
4.2 *CORPUS*

O conjunto de dados utilizado foi o ToLD-BR (*Toxic Language Detection in Brazilian Portuguese*), proposto por Leite *et al.* (2020b). O ToLD-BR é composto por comentários coletados do Twitter em língua portuguesa e anotados por múltiplos anotadores quanto à presença de seis categorias de linguagem tóxica, distribuídas entre insulto, homofobia, obscenidade, racismo, misoginia e xenofobia.

Os arquivos originais, a descrição das colunas, as licenças e os materiais de reprodução podem ser consultados no repositório oficial do ToLD-BR (LEITE *et al.*, 2020a). Neste estudo, os rótulos das seis categorias foram agregados para formar a tarefa binária descrita a seguir.

A classificação binária dos experimentos considerou tóxico (classe 1) todo comentário com anotação positiva em ao menos uma categoria. Os demais foram classificados como não-tóxicos (classe 0). Após esse processo e a remoção de entradas nulas, o conjunto final resultou em 21.000 comentários, dos quais 11.745 (55,9%) foram classificados como não-tóxicos e 9.255 (44,1%) como tóxicos, conforme ilustrado na Figura 1.

Figura 1 – Distribuição de classes no conjunto de dados ToLD-BR



Fonte: Elaborado pela autora (2026).

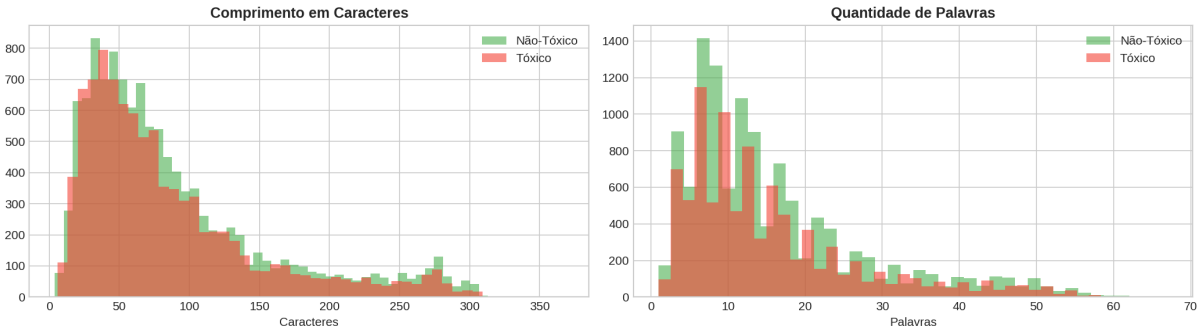
A Figura 1 mostra um desequilíbrio moderado entre as classes. Os comentários não tóxicos representam 55,9% do conjunto, enquanto os tóxicos correspondem a 44,1%. A diferença não impede o treinamento dos modelos, mas justifica o uso de F1-Macro

como métrica principal, pois essa métrica atribui o mesmo peso ao desempenho obtido em cada classe.

4.3 ANÁLISE EXPLORATÓRIA DOS TEXTOS

A análise exploratória também considerou o tamanho dos comentários. A Figura 2 apresenta as distribuições do número de caracteres e de palavras nas duas classes.

Figura 2 – Distribuição do comprimento dos comentários por classe



Fonte: Elaborado pela autora (2026).

As distribuições possuem forte sobreposição e concentram a maior parte dos comentários nas faixas de menor comprimento. As duas classes também apresentam uma cauda formada por textos mais longos. Esse comportamento indica que o tamanho do comentário, isoladamente, oferece pouca capacidade de separação entre conteúdo tóxico e não tóxico.

A frequência visual das palavras foi examinada separadamente para cada classe. A Figura 3 apresenta as nuvens de palavras dos comentários não tóxicos e tóxicos.

Figura 3 – Nuvens de palavras dos comentários não tóxicos e tóxicos

Fonte: Elaborado pela autora (2026).

As duas nuvens compartilham termos frequentes associados à linguagem informal das redes sociais. Os comentários tóxicos dão maior destaque visual a palavrões e formas de tratamento potencialmente ofensivas. Parte desse vocabulário também

aparece na classe não tóxica, o que reforça a dificuldade de classificar os textos apenas pela presença de palavras específicas. As nuvens têm função descritiva e não estabelecem, por si mesmas, associação estatística entre um termo e a toxicidade.

4.4 PRÉ-PROCESSAMENTO

A normalização converteu os textos para letras minúsculas e removeu *links*, menções de usuários, *hashtags*, dígitos numéricos, caracteres especiais e espaços redundantes. Apenas as letras do alfabeto português foram preservadas. Essa normalização foi aplicada de forma idêntica aos textos usados pelos classificadores clássicos, enquanto os modelos *Transformer* utilizaram seus próprios tokenizadores após a limpeza.

4.5 MODELOS AVALIADOS

O estudo avaliou quatro modelos organizados em dois grupos. Os *baselines* clássicos utilizam vetorização TF-IDF com unigramas e bigramas, limitada a 50.000 termos. O primeiro combina TF-IDF com Regressão Logística, regularização L2 e C = 1,0. O segundo combina TF-IDF com *LinearSVC*. Essas abordagens foram escolhidas por constituírem referências lineares fortes e de baixo custo para textos representados em espaços esparsos de alta dimensionalidade (JOACHIMS, 1998). A Regressão Logística fornece uma referência probabilística, enquanto o SVM representa a classificação por máxima margem. O limite de 50.000 termos controla o consumo de memória e foi mantido igual nos dois modelos; C = 1,0 corresponde à regularização padrão usada como configuração de referência, sem ajuste pelo conjunto de teste.

Os modelos *Transformer* foram ajustados com a biblioteca *Transformers* no Google Colaboratory com GPU NVIDIA T4, por 3 épocas, taxa de aprendizado de 2×10^{-5}, *batch size* de 16 e comprimento máximo de 128 *tokens*. Foram utilizados o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), versão `neuralmind/bert-base-portuguese-cased`, e o XLM-RoBERTa (CONNEAU *et al.*, 2020), versão `FacebookAI/xlm-roberta-base`. O BERTimbau foi selecionado como representante monolíngue do português, e o XLM-RoBERTa como alternativa multilíngue. Três épocas, *batch size* 16 e taxa de 2×10^{-5} pertencem às faixas avaliadas nos experimentos de ajuste fino do BERT (DEVLIN *et al.*, 2019). O limite de 128 *tokens* foi adotado porque a análise exploratória mostrou predominância de comentários curtos, reduzindo preenchimento e custo computacional para a maioria dos exemplos.

4.6 MÉTRICAS DE AVALIAÇÃO

Os modelos foram avaliados por Acurácia, F1-Macro (métrica principal, por ser mais adequada a conjuntos com algum desequilíbrio entre classes), F1 da classe tóxica,

Precisão Macro, Recall Macro e tempo de inferência em segundos como indicador de viabilidade computacional.

4.7 PROTOCOLO EXPERIMENTAL E REPRODUTIBILIDADE

Todos os modelos foram comparados sobre a mesma divisão estratificada, formada por 80% dos dados para treinamento e 20% para teste. O conjunto de teste contém 4.200 comentários, dos quais 2.349 pertencem à classe não tóxica e 1.851 à classe tóxica. A preservação aproximada da proporção original reduz o risco de uma avaliação favorecida por alterações casuais na distribuição das classes. A semente aleatória 42 foi mantida durante a preparação dos dados para permitir a reconstrução da partição.

Não foi reservado um conjunto de validação separado. Como não houve busca de hiperparâmetros nem parada antecipada, as configurações foram definidas antes da avaliação final, a partir da literatura e das restrições do ambiente computacional. A partição de teste foi mantida para o cálculo final das métricas. Essa escolha torna o protocolo mais simples, mas impede estimar como diferentes configurações se comportariam em um conjunto de validação e é considerada uma limitação do estudo.

O repositório do ToLD-BR também oferece as divisões de treinamento, desenvolvimento e teste usadas no artigo original (LEITE *et al.*, 2020a). Elas não foram adotadas aqui porque, após a agregação binária e a remoção das entradas nulas, foi criada uma única divisão estratificada comum aos quatro modelos. Essa decisão favorece a comparação interna deste estudo, mas reduz a comparabilidade direta com os resultados produzidos nas partições originais.

Os classificadores clássicos e os modelos contextuais receberam os mesmos exemplos de treinamento e foram avaliados nos mesmos comentários. Essa decisão controla uma fonte importante de variação, pois impede que diferenças de desempenho sejam atribuídas a conjuntos de teste distintos. Para os modelos *Transformer*, o texto foi limitado a 128 *tokens*, e o ajuste fino utilizou três épocas, *batch size* 16 e taxa de aprendizado de 2×10^{-5}. Os parâmetros foram mantidos constantes entre BERTimbau e XLM-RoBERTa.

O tempo de inferência corresponde ao processamento completo do conjunto de teste no ambiente utilizado. Essa medida oferece uma indicação prática do custo relativo, mas não constitui uma comparação universal de eficiência. Versões de bibliotecas, aquecimento da GPU, transferência de dados e otimizações de execução podem modificar os valores absolutos. A conclusão sobre eficiência concentra-se na diferença de ordem de grandeza entre os métodos clássicos e os modelos contextuais.

O protocolo privilegia uma comparação controlada e compreensível, adequada ao escopo do estudo. Não foram executadas buscas extensas de hiperparâmetros nem repetições com várias sementes. Os valores reportados representam uma execução com configuração fixa. Essa escolha reduz o custo experimental, mas limita a estimativa

da variabilidade dos resultados e será retomada na discussão das ameaças à validade.

O notebook com o código dos experimentos e as configurações utilizadas está disponível no Google Colaboratory (LEITE, 2026). O *corpus*, as licenças e os materiais de reprodução do ToLD-BR estão disponíveis no repositório oficial dos autores (LEITE *et al.*, 2020a).

5 RESULTADOS E DISCUSSÃO

5.1 RESULTADOS QUANTITATIVOS

A Tabela 1 apresenta o desempenho comparativo dos quatro modelos avaliados sobre o conjunto de teste.

Tabela 1 – Comparativo de desempenho dos modelos avaliados

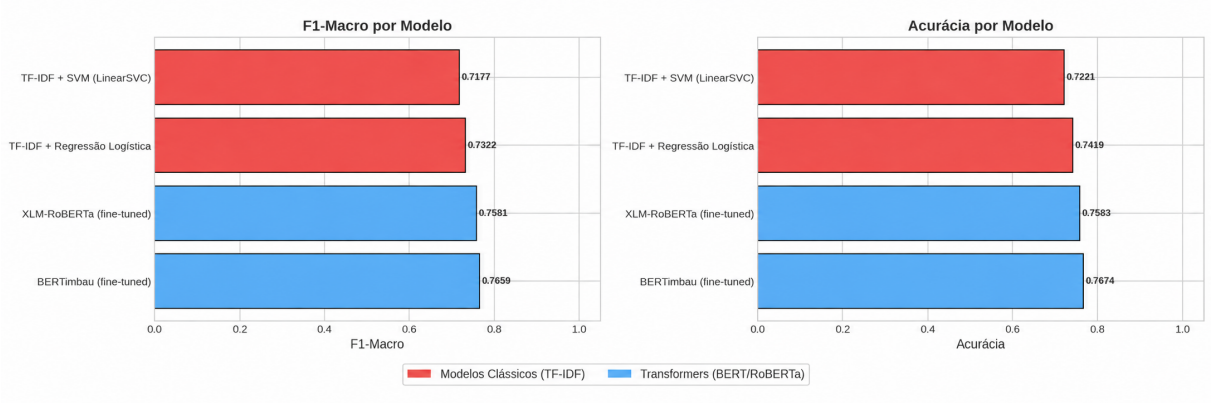
Modelo	**Acurácia**	**F1-Macro**	**F1-Tóxico**	**Precisão**	**Recall**	**Tempo (s)**
BERTimbau	0,7674	0,7659	0,7473	0,7654	0,7688	6,143
XLM-RoBERTa	0,7583	0,7581	0,7512	0,7629	0,7657	8,180
TF-IDF + LR	0,7419	0,7322	0,6812	0,7431	0,7296	0,001
TF-IDF + SVM	0,7221	0,7177	0,6824	0,7181	0,7174	0,001

Fonte: Elaborado pela autora (2026).

O BERTimbau obteve o melhor desempenho geral, com F1-Macro de 0,7659 e Acurácia de 0,7674, confirmando a hipótese de que modelos *Transformer* superam métodos baseados em TF-IDF para a tarefa. O ganho sobre o melhor *baseline* foi de 3,37 pontos percentuais em F1-Macro. O fato de o BERTimbau ter superado o XLM-RoBERTa em 0,78 ponto percentual sugere que a especialização no idioma é fator relevante, resultado consistente com Souza, Nogueira & Lotufo (2020) e Oliveira *et al.* (2024).

Os resultados por classe revelam diferenças entre os modelos *Transformer*. O XLM-RoBERTa obteve F1 da classe tóxica de 0,7512, superior ao BERTimbau (0,7473), indicando maior sensibilidade à classe positiva. Enquanto o BERTimbau apresenta menor taxa de falsos positivos (24,3%), o XLM-RoBERTa detecta mais tóxicos ao custo de mais falsos positivos (29,6%). A escolha entre os modelos depende do contexto de uso. Sistemas de moderação automática devem privilegiar a Precisão. Já os sistemas de triagem para revisão humana podem tolerar mais falsos positivos. A Figura 4 ilustra o comparativo entre os modelos.

Figura 4 – Comparativo de F1-Macro e Acurácia entre os modelos avaliados



Fonte: Elaborado pela autora (2026).

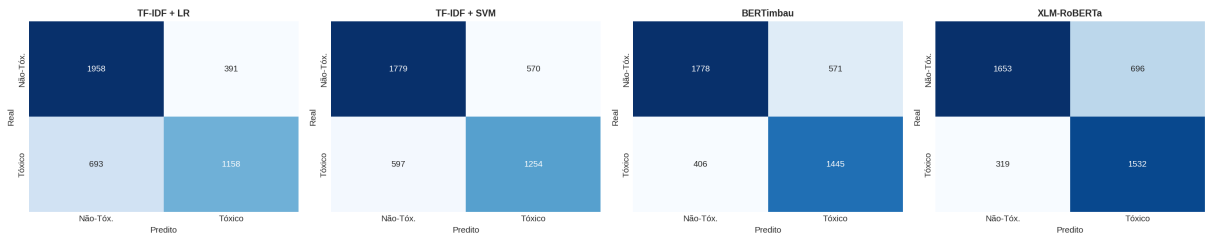
A representação gráfica evidencia a vantagem dos modelos pré-treinados nas duas métricas. BERTimbau e XLM-RoBERTa ocupam as primeiras posições, enquanto os métodos baseados em TF-IDF apresentam valores inferiores. A diferença entre os dois modelos *Transformer* é pequena, embora o BERTimbau mantenha o melhor resultado em F1-Macro e Acurácia.

Também houve diferença expressiva no tempo de inferência. Os modelos clássicos classificaram o conjunto de teste em 0,001 segundo. O BERTimbau levou 6,14 segundos e o XLM-RoBERTa, 8,18 segundos. Essa diferença de quatro a cinco ordens de grandeza é relevante para ambientes com requisitos de latência, como plataformas de moderação em tempo real.

5.2 ANÁLISE DAS MATRIZES DE CONFUSÃO

A Figura 5 apresenta as matrizes de confusão dos quatro modelos sobre o conjunto de teste.

Figura 5 – Matrizes de confusão dos quatro modelos avaliados



Fonte: Elaborado pela autora (2026).

O TF-IDF com Regressão Logística apresentou a menor taxa de falsos positivos (16,6%) e a maior taxa de falsos negativos (37,4%), indicando tendência a subestimar a toxicidade. O SVM apresentou comportamento similar, com 32,3% de falsos negativos.

O BERTimbau reduziu essa taxa para 21,9%, e o XLM-RoBERTa alcançou a menor taxa de falsos negativos entre todos os modelos (17,2%), ao custo da maior taxa de falsos positivos (29,6%). Os resultados indicam que os *Transformers* melhoram substancialmente a detecção de toxicidade, com redução relevante de falsos negativos em relação aos métodos clássicos.

5.3 ANÁLISE QUALITATIVA DOS ERROS

A análise de uma amostra dos 977 comentários classificados incorretamente pelo BERTimbau revelou dois padrões predominantes. O primeiro corresponde aos falsos positivos, nos quais o modelo atribui toxicidade a comentários rotulados como não tóxicos. A maioria desses casos envolve palavrões usados como intensificadores emocionais, sem intenção ofensiva, como em “cheguei amanhã tem botafogo porra” ou “meu casal finalmente beijou caralho esse momento eh todo meu”. O mesmo vocabulário ocorre em ofensas reais, embora exerça função pragmática distinta nesses exemplos.

O segundo padrão corresponde aos falsos negativos, nos quais o modelo não detecta toxicidade real. Nesses casos, a ofensa se manifesta de forma implícita, por meio de zombaria ou referência contextual que demanda conhecimento externo, como em “providencia gorda” ou “comparação mais ridícula ahahah”. Esses padrões são consistentes com os desafios identificados por Fortuna & Nunes (2018) e Leite *et al.* (2020b), que apontam a ambiguidade pragmática e a toxicidade implícita como os principais obstáculos para sistemas automáticos de detecção no português brasileiro.

5.4 DISCUSSÃO INTEGRADA DOS RESULTADOS

Os resultados permitem responder às questões apresentadas na introdução. Em relação à primeira, os dois modelos *Transformer* superaram as abordagens clássicas em F1-Macro e Acurácia. A vantagem ocorreu sob a mesma partição de teste e permaneceu visível nas matrizes de confusão, sobretudo pela redução de falsos negativos. Esse comportamento sugere que representações contextuais capturam relações que a frequência ponderada de termos não consegue descrever. O ganho absoluto de 3,37 pontos percentuais do BERTimbau sobre a Regressão Logística é relevante, mas deve ser considerado em conjunto com o aumento do custo computacional.

Quanto à segunda questão, o BERTimbau obteve o melhor F1-Macro e a melhor Acurácia, enquanto o XLM-RoBERTa apresentou maior F1 para a classe tóxica. A diferença de 0,78 ponto percentual entre os modelos é pequena e não sustenta uma afirmação universal de superioridade. O resultado oferece evidência favorável à especialização linguística nesta configuração experimental, em consonância com trabalhos que relatam benefícios de modelos monolíngues para o português (SOUZA; NOGUEIRA;

LOTUFO, 2020; OLIVEIRA *et al.*, 2024). Repetições com outras sementes, partições e conjuntos de dados são necessárias para avaliar a estabilidade dessa vantagem.

A terceira questão é respondida pela análise qualitativa. Os erros não se concentram apenas em palavras raras ou construções gramaticais incomuns. Muitos decorrem da função assumida por expressões frequentes em uma conversa. Um palavrão pode atuar como ataque, marca de surpresa, intensificador ou sinal de proximidade entre interlocutores. Em sentido inverso, uma ofensa pode dispensar vocabulário explicitamente agressivo e depender de conhecimento compartilhado. As nuvens de palavras e as matrizes de confusão convergem com essa leitura, pois mostram sobreposição lexical entre as classes e persistência de erros mesmo nos modelos contextuais.

Do ponto de vista operacional, a escolha do classificador depende do custo de cada tipo de erro. Uma ferramenta que apenas prioriza mensagens para revisão humana pode favorecer maior Recall e aceitar uma quantidade superior de falsos positivos. Um sistema que oculta conteúdo automaticamente exige maior Precisão, pois uma decisão incorreta afeta diretamente a participação do usuário. Os resultados não indicam um modelo ideal para todos os cenários. Eles fornecem subsídios para selecionar uma configuração compatível com o risco e a capacidade de revisão de cada aplicação.

5.5 AMEAÇAS À VALIDADE

O estudo reuniu as seis categorias do ToLD-BR em uma classificação binária. Isso facilita a comparação dos modelos, mas esconde diferenças entre insulto, obscenidade, racismo, misoginia, xenofobia e homofobia. Um modelo pode apresentar bom resultado geral e ainda falhar nas categorias com menos exemplos.

Os experimentos utilizaram uma única divisão dos dados e a semente 42. Como o treinamento dos modelos contém etapas aleatórias, novas execuções podem produzir pequenas diferenças. Também não houve um conjunto de validação separado nem uma busca ampla de hiperparâmetros. Por essa razão, a pequena vantagem do BERTimbau sobre o XLM-RoBERTa deve ser vista como um resultado desta configuração, e não como uma superioridade definitiva.

Outra limitação está na origem dos dados. O ToLD-BR contém comentários do Twitter coletados em 2019. Outras plataformas possuem públicos, formatos e vocabulários diferentes, e a linguagem das redes sociais muda com o tempo. Estudos com conjuntos em português confirmam que essas diferenças afetam a transferência dos modelos (MOREIRA *et al.*, 2026). Além disso, não foram calculados intervalos de confiança ou testes estatísticos. Os resultados mostram tendências importantes, mas precisam ser confirmados em novas bases e execuções.

5.6 CONSIDERAÇÕES ÉTICAS E USO RESPONSÁVEL

Erros de classificação podem gerar consequências diferentes. Um falso positivo pode ocultar uma fala legítima, inclusive linguagem informal ou usada por grupos minoritários. Um falso negativo pode manter ataques visíveis e ampliar o dano às pessoas atingidas. Por isso, a escolha do modelo e do limite de decisão deve considerar o modo como o sistema será utilizado.

Os modelos devem funcionar como apoio à moderação, com revisão humana e opção de contestação para o usuário. Avaliações periódicas ajudam a identificar perdas de desempenho e possíveis diferenças entre grupos. O conjunto usado neste trabalho já foi publicado e anonimizado, mas exemplos ofensivos ainda precisam ser apresentados com cuidado. Identificadores devem permanecer ocultos, e os rótulos não devem ser tratados como julgamentos definitivos sobre pessoas ou mensagens.

6 CONCLUSÃO

Este trabalho apresentou um estudo comparativo entre quatro abordagens de classificação automática de comentários tóxicos em português brasileiro, avaliando métodos clássicos baseados em TF-IDF e modelos de linguagem pré-treinados baseados na arquitetura *Transformer*. Os experimentos foram conduzidos sobre o conjunto ToLD-BR, composto por 21.000 comentários do Twitter, utilizando Acurácia, F1-Macro, Precisão, Recall e tempo de inferência como critérios de avaliação.

Os resultados confirmaram a hipótese central do trabalho. Os modelos *Transformer* superaram os *baselines* clássicos em todas as métricas de qualidade. O BERTimbau alcançou o melhor desempenho geral (F1-Macro = 0,7659; Acurácia = 0,7674), com ganho de 3,37 pontos percentuais sobre o melhor modelo clássico. O fato de o BERTimbau ter superado o XLM-RoBERTa sugere que a especialização em língua portuguesa é um fator relevante para essa tarefa, mesmo quando comparado a modelos multilíngues treinados em volumes de dados substancialmente maiores.

A análise qualitativa dos erros identificou a ambiguidade pragmática do português coloquial como o principal desafio ainda não resolvido. O uso expressivo de palavrões como intensificadores emocionais gera falsos positivos em quantidade significativa. Ofensas implícitas e contextualizadas, por sua vez, resultam em falsos negativos. Esses achados indicam a necessidade de abordagens capazes de representar o léxico e a função pragmática das expressões no contexto da comunicação digital brasileira.

Uma limitação deste trabalho decorre do uso exclusivo do *dataset* ToLD-BR, composto por comentários do Twitter. Essa escolha pode limitar a generalização dos resultados para outras plataformas com padrões discursivos distintos. Além disso, os modelos *Transformer* avaliados não foram submetidos a técnicas de otimização de hiperparâmetros, como busca em grade ou otimização bayesiana, o que pode ter deixado espaço para melhorias de desempenho.

6.1 TRABALHOS FUTUROS

Pesquisas futuras podem avaliar modelos mais recentes da família BERT para o português, como o BERTimbau-large, e modelos derivados de LLaMA adaptados ao idioma. Outra direção envolve a coleta e a anotação de comentários do Instagram, YouTube e outras plataformas digitais brasileiras, com o objetivo de ampliar a diversidade discursiva dos conjuntos de dados. Também são relevantes o desenvolvimento de técnicas de desambiguação pragmática e o estudo de arquiteturas distribuídas para inferência de modelos *Transformer* em tempo real em sistemas de moderação de grande escala.

REFERÊNCIAS

CONNEAU, A. *et al.* Unsupervised cross-lingual representation learning at scale. In: **Proceedings of the 58th Annual Meeting of the ACL**. Online: Association for Computational Linguistics, 2020. p. 8440–8451.

CRESWELL, J. W. **Research Design: Qualitative, Quantitative, and Mixed Methods Approaches**. 4. ed. Thousand Oaks: SAGE Publications, 2014.

DEVLIN, J. *et al.* Bert: pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of NAACL-HLT**. Minneapolis: Association for Computational Linguistics, 2019. p. 4171–4186.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. **ACM Computing Surveys**, v. 51, n. 4, p. 85:1–85:30, 2018.

JOACHIMS, T. Text categorization with support vector machines: learning with many relevant features. **Lecture Notes in Computer Science**, v. 1398, p. 137–142, 1998.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3. ed. Upper Saddle River: Prentice Hall, 2023.

LEITE, J. A. *et al.* **ToLD-Br: Toxic Language Dataset for Brazilian Portuguese**. 2020. Repositório GitHub. Disponível em: <https://github.com/joaoaleite/ToLD-Br>. Acesso em: 21 ago. 2026.

LEITE, J. A. *et al.* Toxic language detection in social media for Brazilian Portuguese: new dataset and multilingual analysis. In: **Proceedings of the 1st Conference of the Asia-Pacific Chapter of the ACL and the 10th International Joint Conference on NLP**. Suzhou, China: Association for Computational Linguistics, 2020. p. 914–924.

LEITE, M. E. C. **Código e experimentos para detecção de discurso tóxico em redes sociais brasileiras**. 2026. Notebook no Google Colaboratory. Disponível em: <https://colab.research.google.com/drive/1Td25LbQqUPXyMRmivVHg8wZx1CpGGdcG?usp=sharing>. Acesso em: 21 ago. 2026.

MOREIRA, L. S. *et al.* Anatomy of data repositories for the analysis and detection of toxicity in Portuguese. In: **Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) – Vol. 1**. Salvador, Brazil: Association for Computational Linguistics, 2026. p. 456–466.

OLIVEIRA, A. da S. *et al.* Toxic speech detection in Portuguese: A comparative study of large language models. In: **Proceedings of the 16th International Conference on Computational Processing of Portuguese – Vol. 1**. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. p. 108–116.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing and Management**, v. 24, n. 5, p. 513–523, 1988.

SCHMIDT, A.; WIEGAND, M. A survey on hate speech detection using natural language processing. **Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media**, p. 1–10, 2017.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: **9th Brazilian Conference on Intelligent Systems (BRACIS)**. Rio de Janeiro: Springer, 2020.

TRAJANO, D.; BORDINI, R. H.; VIEIRA, R. OLID-BR: offensive language identification dataset for Brazilian Portuguese. **Language Resources and Evaluation**, Springer, v. 58, n. 4, p. 1263–1289, 2024.

VARGAS, F. *et al.* HateBR: a large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)**. Marseille, France: European Language Resources Association, 2022. p. 7174–7183.

VASWANI, A. *et al.* Attention is all you need. In: **Advances in Neural Information Processing Systems**. Long Beach: Curran Associates, 2017. v. 30.

WE ARE SOCIAL; MELTWATER. **Digital 2024: Global Overview Report**. 2024. WE ARE SOCIAL. Disponível em: <https://wearesocial.com/br/blog/2024/01/digital-2024/>. Acesso em: 10 mar. 2025.

ZAMPIERI, M. *et al.* Predicting the type and target of offensive posts in social media. In: **Proceedings of NAACL-HLT**. Minneapolis: ACL, 2019. p. 1415–1420.