O que torna uma frase tóxica? Uma análise crítica de modelos especialistas em detecção de toxicidade

Gabriel Melo
gabrielmelo96@ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil

Flavio Figueredo
flaviovdf@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil

ABSTRACT

This study examines the performance of *specialized* Machine Learning models in the task of online toxicity detection, under the hypothesis that these systems disproportionately focus on isolated lexical items. We conduct a comparative analysis of the Detoxify and Perspective API models using the HateXplain (English) and ToLDE-Br (Portuguese) datasets. To assess model behavior, we employ the SHAP explainability framework, which enables the interpretation of feature importance in individual predictions. Our findings reveal a misalignment between model outputs and the nuanced, evolving nature of language on social media platforms. Furthermore, the results demonstrate an overreliance on negatively connoted keywords, which compromises the models' classification accuracy and raises concerns regarding their robustness and fairness in real-world applications.

KEYWORDS

Toxicity Detection, Natural Language Processing, Explainability, Online Content Moderation

1 INTRODUÇÃO

> **Aviso Ético:** Esta pesquisa envolve a análise de textos que contêm linguagem tóxica, incluindo ofensas e discurso de ódio. Os autores repudiam veementemente o teor dessas mensagens, cujo uso nesta pesquisa é estritamente acadêmico.

Plataformas de mídia social como Twitter e Reddit, ao permitirem interações massivas, criam um terreno fértil não apenas para a troca de conhecimento, mas também para a disseminação de conteúdo tóxico [5, 28]. A toxicidade online, definida aqui como qualquer texto rude, desrespeitoso ou irracional que provavelmente afaste um participante de uma discussão [36], representa um desafio significativo para a moderação de conteúdo.

Para combater esse problema em escala, a abordagem mais avançada atualmente baseia-se em modelos de Redes Neurais Profundas (RNP), com destaque para as *Large Language Models* (LLMs) — como ChatGPT, Claude e Gemini — que representam o estado da arte na detecção de linguagem tóxica e outras tarefas complexas de Processamento de Linguagem Natural (PLN). Contudo, a própria natureza generalista dessas arquiteturas, que lhes confere vasta versatilidade, abre espaço para a busca por um desempenho ainda mais otimizado em tarefas de nicho. Nesse cenário, modelos especializados, desenvolvidos exclusivamente para a tarefa de detecção de toxicidade, continuam sendo relevantes, pois permitem uma otimização direcionada da arquitetura e do treinamento em dados anotados com critérios bem definidos [26]. No entanto, a eficácia desses sistemas — sejam generalistas ou especializados — ainda é frequentemente questionada, uma vez que a toxicidade é uma categoria subjetiva, com normas que variam entre comunidades e indivíduos [7]. Muitos desses modelos operam como "caixas-pretas", o que levanta preocupações quanto à robustez e transparência de seus julgamentos [29]. Estudos mostram que até mesmo modelos especialistas amplamente utilizados, como o Perspective API, são suscetíveis a manipulações simples. Por exemplo, a frase *"They are stupid and ignorant"* foi classificada com 91% de toxicidade, mas ao ser alterada para *"They are st.upid and ig.norant"*, a pontuação caiu para 11% [16].

Considerando essa fragilidade, a hipótese que guia nossa pesquisa é: **Modelos especializados em detecção de toxicidade que utilizam RNP e PLN apresentam falhas ao atribuir atenção desproporcional a palavras isoladas.**

Para investigar essa hipótese, este trabalho emprega técnicas de explicabilidade da área de *eXplainable Artificial Intelligence* (XAI) [14] para analisar o comportamento de dois modelos amplamente difundidos: **Perspective API** [23] e **Detoxify** [15]. Nossa análise é conduzida em dois conjuntos de dados oriundos de redes sociais com anotações da causa da toxicidade em nível de palavra: o *HateXplain* [26], para a língua inglesa, e o *ToLDE-Br*, um novo corpus de explicações para o português brasileiro desenvolvido nesta pesquisa a partir da base ToLD-Br [24].

O objetivo deste artigo é: (i) avaliar o desempenho dos modelos Detoxify e Perspective API na detecção de toxicidade nas bases HateXplain (inglês) [26] e ToLDE-Br (português), por meio de métricas de Acurácia, Precisão, Revocação e AUC; (ii) quantificar a qualidade das justificativas geradas pelo SHAP, comparando-as com anotações humanas e um baseline aleatório para mensurar alinhamento e lacunas explicativas; e (iii) investigar padrões de erro em nível de palavra usando categorias LIWC [31], identificando classes semânticas com maiores taxas de falsos positivos e falsos negativos.

Para alcançar esses objetivos, o restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta o referencial teórico. A Seção 3 detalha os materiais e métodos utilizados. A Seção 4 apresenta e discute os resultados. Por fim, a Seção 5 conclui o trabalho.

2 TRABALHOS RELACIONADOS

A detecção de toxicidade online e a busca por interpretabilidade em modelos de Aprendizado de Máquina (AM) são áreas de pesquisa

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

consolidadas e vastas. Nesta seção, posicionamos nosso trabalho em relação a outros dessa temática.

2.1 O Desafio da Detecção de Toxicidade

A linguagem abusiva é abordada na literatura sob diferentes categorias, como discurso de ódio, cyberbullying, agressão e incivilidade [8, 11, 22]. Neste trabalho, adotamos o termo *toxicidade* para descrever expressões ofensivas e discriminatórias dirigidas a grupos minoritários, incluindo conteúdo sexista, racista, homofóbico e xenofóbico.

A detecção de toxicidade é geralmente formulada como uma tarefa de classificação supervisionada binária ou multicategoria, onde o objetivo é estimar a probabilidade de um texto ser percebido como tóxico. No entanto, essa tarefa é sensível a variações culturais, demográficas e linguísticas [4, 20], o que compromete a consistência das anotações [21, 30].

Desde abordagens baseadas em regras, como o sistema *Smokey* [34], até modelos com arquiteturas modernas baseadas em *Transformers* [35, 37], a detecção de toxicidade evoluiu significativamente. Estudos recentes também exploram os gatilhos conversacionais da toxicidade [1, 2] e a relação entre tópicos sensíveis e comentários tóxicos em plataformas como o YouTube [3].

A dinâmica de engajamento dos usuários em plataformas com moderação limitada, como o Telegram, impõe desafios complexos aos modelos de detecção de toxicidade. Um estudo de Cavalini et al. [6] sobre comunidades extremistas brasileiras ilustra essa questão: foi observado que o conteúdo altamente tóxico é o que alcança maior público e se torna viral, superando o alcance de mensagens menos ofensivas. Este padrão representa um desafio crítico, pois adiciona uma camada de complexidade à tarefa de monitoramento, exigindo que o conteúdo com maior potencial de dano seja priorizado. Consequentemente, a detecção de toxicidade nesses ambientes deve evoluir, focando não apenas na identificação da linguagem agressiva, mas também na previsão do seu risco de disseminação em larga escala.

2.2 Explicabilidade combinada a modelos de Detecção de Toxicidade

Diversos estudos têm aplicado técnicas de explicabilidade a modelos de detecção de toxicidade, com o objetivo de compreender seus vieses e propor melhorias. Neste contexto, destacam-se investigações que abordam a vulnerabilidade dos modelos a ataques adversariais [12], a inconsistência na definição de categorias entre diferentes conjuntos de dados [9], e a importância da inclusão de explicações humanas na construção de modelos mais justos e interpretáveis [26].

Grondahl et al. [12] demonstram que modelos treinados em dados específicos apresentam dificuldades de generalização, sendo suscetíveis a manipulações simples, como erros ortográficos ou reformulações sintáticas. Já Fortuna et al. [9] evidenciam a incompatibilidade conceitual entre categorias como 'toxicidade' e 'discurso de ódio' em diferentes bases, dificultando a replicação e comparação entre estudos. Como solução, propõem diretrizes padronizadas para anotação de dados.

O estudo de Mathew et al. [26] propõe a base HateXplain, que incorpora justificativas fornecidas por anotadores humanos. Os autores mostram que o uso dessas explicações melhora tanto a transparência quanto a mitigação de vieses, especialmente em relação a grupos marginalizados. Eles também alertam para a tendência de modelos se apoiarem em palavras isoladas, em detrimento de uma interpretação contextual.

Complementarmente, Risch et al. [32] comparam métodos de explicação como LIME, LRP e LSTMs com atenção. Os resultados sugerem que, embora modelos complexos ofereçam melhor desempenho, abordagens como LIME e LRP apresentam interpretabilidade superior com menor custo computacional.

Mencionamos também o estudo de Salles et al. [33], que propõe uma base e um estudo similares ao nosso. Contudo, o trabalho citado não aborda a mesma hipótese de pesquisa que investigamos. Em contraste, o estudo enfatiza a construção e a avaliação de uma nova base voltada à explicabilidade.

Outros estudos reforçam o papel da explicabilidade na construção de sistemas mais confiáveis e colaborativos [27], e na melhoria de desempenho através da combinação de modelos e métodos explicativos como LIME e SHAP [10].

O presente trabalho insere-se nesse panorama, com foco na análise de modelos já consolidados, utilizando insights psicológicos baseados no LIWC e um novo conjunto de sentenças tóxicas anotadas em português. Nosso objetivo é contribuir para o avanço de abordagens explicáveis que sejam robustas e culturalmente sensíveis.

3 MATERIAIS E MÉTODOS

A fim de compreender como modelos especializados em detecção de toxicidade tomam decisões e se, de fato, apresentam vieses na atenção a palavras isoladas, estruturamos nossa análise empírica a partir de dois conjuntos de dados anotados com justificativas em nível lexical. Além disso, incorporamos uma análise semântica de alto nível, com base no léxico LIWC, para investigar padrões linguísticos associados ao desalinhamento do foco dos avaliadores humanos e dos modelos.

3.1 Conjuntos de Dados

3.1.1 HateXplain. O conjunto de dados HateXplain contém 20.148 sentenças coletadas do Twitter[1] (45%) e do Gab[2] (55%). As sentenças foram anotadas por três avaliadores em quatro categorias: *Hateful*, *Offensive*, *Normal* e *Undecided*. Quando uma sentença era classificada como *Hateful* ou *Offensive*, os avaliadores também destacavam as palavras que justificavam tal anotação. A Tabela 1 resume a distribuição das classes por plataforma.

Tabela 1: Distribuição de classes no HateXplain

Classe	Twitter	Gab	Total
Toxic (Hateful+Offensive)	3.036	8.379	11.415
Normal	5.770	2.044	7.814
Undecided	249	670	919
Total	9.055	11.093	20.148

[1]Como a base foi coletada antes da mudança de nome, aqui usamos o nome original Twitter e não X

[2]Uma rede similar ao Twitter sem nenhum filtro e moderação de conteúdo

Para as nossas análises, as classes *Hateful* e *Offensive* foram agrupadas sob o rótulo *Toxic*, e as sentenças da classe *Undecided* foram descartadas. Além disso, foi realizado um processamento dos rótulos de origem da toxicidade: foram consideradas como fontes de toxicidade as palavras apontadas por, no mínimo, dois dos três avaliadores. A Tabela 2 apresenta alguns exemplos de sentenças tóxicas com suas respectivas justificativas.

Tabela 2: Exemplos de sentenças tóxicas do HateXplain. As palavras destacadas representam as justificativas dos anotadores para a escolha da classe.

*y is big baby davis a f*ggot on shameless doe*
*Dad should have told the muzr*t where to f*ck off, and went in anyway*
*A n*gress too dumb to f*ck has a scant chance of understanding anything beyond the size of a d*ck*
*Twitter is full of tween d*kes who think they're superior because of "muh oppression". News flash: No one gives a sh*t.*

3.1.2 ToLDE-Br. O conjunto de dados ToLDE-Br, disponível em https://github.com/uai-ufmg/tolde-br, é composto por 1.000 sentenças (500 tóxicas e 500 não tóxicas) extraídas da ToLD-Br, um conjunto originalmente formado por 21 mil tweets anotados segundo múltiplas classes de toxicidade. A partir das sentenças selecionadas, foram criadas as justificativas da toxicidade, nas quais cada sentença foi avaliada por três pessoas que definiram quais palavras eram mais relevantes para determinar a classificação de tóxico.

A rotulação das palavras tóxicas neste subconjunto foi realizada por 11 autores, sendo três avaliadores por sentença. Os anotadores sabiam que os textos eram tóxicos, mas não conheciam suas subcategorias. Eles deveriam destacar as palavras responsáveis pela percepção de toxicidade, podendo também optar por não marcar nenhuma palavra.

Com as justificativas do conjunto de dados criadas, aplicou-se o mesmo processamento do conjunto HateXplain, e foram utilizadas apenas as palavras que obtiveram o consenso de, no mínimo, dois dos três avaliadores.

Tabela 3: Exemplos de sentenças tóxicas do ToLDE-Br. As palavras destacadas representam os termos rotulados como relevantes pelos anotadores.

car*lho que vizinhos insuportáveis p*ta m*rda
O que vag*bundo? Tu tá em rede social fela da p*ta.... falou tem resposta. Seu p*u no c*
eu também tiro foto com vi*dos e simpatizo mas isso não quer dizer nada se eu pudesse ainda metia uma surra
eu ai me agarrava nessa infeliz e descia unhada e tapa nela vsf fã coreana não tem um pingo de simancol

3.2 Modelos de Detecção de Toxicidade

Utilizamos dois modelos amplamente adotados na literatura: o *Perspective API* e o *Detoxify*.

O *Perspective API*, desenvolvido pela Jigsaw (Google), é uma API de análise de toxicidade textual voltada a moderadores e desenvolvedores de plataformas digitais. A ferramenta fornece escores de toxicidade com base em milhões de comentários anotados e suporta diversas categorias, incluindo *Toxic*, *Severe Toxic*, *Insult*, *Threat*, *Obscene* e *Identity Hate*. Sua aplicação em pesquisas acadêmicas é vasta; Guimarães et al. [13], por exemplo, utilizaram a API em um diagnóstico em larga escala da toxicidade em mais de 4 milhões de comentários no Facebook no contexto brasileiro. Em seu estudo, os autores validaram a acurácia do modelo para o português, concluindo que sua performance pode ser tão boa quanto a de um anotador humano. Neste trabalho, utilizamos exclusivamente a pontuação da classe *Toxic*, que apresenta maior generalização. Foram usadas as versões do modelo para inglês e português de 2023, compatíveis com as bases analisadas.

A biblioteca *Detoxify*, desenvolvida pela Unitary com base em soluções vencedoras das competições promovidas pela Jigsaw no Kaggle [17–19], disponibiliza três variantes: *Original*, *Unbiased* e *Multilingual*, com AUCs superiores a 94%. A arquitetura é baseada em *Transformers* [35], e os modelos retornam escores contínuos entre 0 e 1, indicando a probabilidade de toxicidade. Adotamos as versões *Original* e *Unbiased* para os dados em inglês, e a versão *Multilingual* para os dados em português, considerando sua robustez, código aberto e facilidade de uso.

3.3 Procedimento de Análise e Métricas

Para a análise semântica e psicológica dos textos, empregamos a ferramenta *Linguistic Inquiry and Word Count* (LIWC). O LIWC é um léxico computacional que categoriza palavras em dezenas de classes, como processos afetivos (*affect*), cognitivos (*cogmech*) e sociais (*social*), permitindo a extração de insights sobre o conteúdo do texto.

Em nossa metodologia, utilizamos as versões do LIWC para português (`LIWC_2007pt`) e inglês (`LIWC_2015en`). Adotamos dois critérios de filtragem para refinar a análise: **(1)** consideramos apenas as categorias do grupo de Processos Psicológicos, por sua maior relevância para o estudo da toxicidade, acrescentando a classe *swear* (palavrões); e **(2)** em casos de categorias hierárquicas (ex.: *emoção negativa* → *raiva*), utilizamos apenas a classe mais específica (*raiva*) para evitar redundância. Palavras que não constavam nos léxicos foram agrupadas em uma classe *unknown* para análise posterior.

A análise da classe *unknown* revelou uma limitação importante do LIWC para o domínio de redes sociais. Esta categoria continha não apenas gírias e abreviações da internet (como *rt*, `user`, *pq*), mas também um volume considerável de ofensas e termos pejorativos que não fazem parte do léxico padrão (ex.: *p*tinha*, *f*der* em português; *k*ke*, *m*slem*, *n*zi* em inglês). Este achado indica que, embora útil para identificar categorias psicológicas gerais, o LIWC nas versões utilizadas não possui cobertura adequada para muitas palavras-chave cruciais em contextos de discurso de ódio online.

4 RESULTADOS

Esta seção apresenta a análise de desempenho dos modelos de detecção de toxicidade e uma investigação detalhada sobre suas falhas de interpretabilidade. A análise está dividida em três partes: (i) uma caracterização linguística das bases de dados para contextualizar os desafios; (ii) uma avaliação quantitativa do desempenho dos modelos em tarefas de classificação; e (iii) uma análise da qualidade das justificativas geradas, que representa a contribuição central deste trabalho ao revelar limitações específicas na capacidade interpretativa dos modelos.

4.1 Caracterização das bases

Para contextualizar os desafios linguísticos de cada corpus, realizamos uma análise textual com o léxico LIWC , visando identificar as categorias mais proeminentes nos dados.

O classificador LIWC foi utilizado para agrupar o vocabulário em categorias e identificar tendências gerais. As Figuras 1 e 2 mostram as 15 maiores diferenças percentuais nas classes LIWC para as bases HateXplain e ToLDE-Br, respectivamente.

Cada valor foi obtido dividindo-se o número de ocorrências das palavras de uma classe pelo total de palavras em seu respectivo grupo (tóxico ou não tóxico). A divisão considerou sentenças rotuladas como tóxicas e não tóxicas.

Na Figura 1, destacam-se as classes 'swear' e 'anger', ambas esperadas em um conjunto de dados voltado para a análise de toxicidade. A classe 'focuspresent' também apresenta alta incidência, com 11% e 10% em sentenças tóxicas e não tóxicas, respectivamente. Inspecionando os fatores que levaram a isso, foi visto que em razão de verbos comuns como '*be*', '*have*' e '*do*'. Isso sugere uma ênfase em ações imediatas. A classe 'focusfuture' inclui termos como '*will*' e '*hope*', mais frequentes em textos tóxicos, sugerindo um tom mais assertivo ou ameaçador.

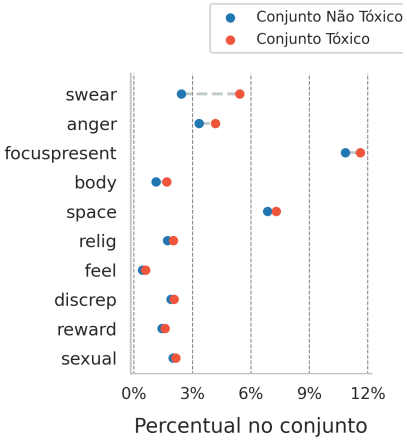


Figura 1: Top 10 classes como maior variação de ocorrências entre o conjunto de sentenças tóxicas e não tóxicas da base HateXplain

Já na Figura 2, as classes 'bio', 'body', 'swear', 'negemo', 'sexual' e 'anger' se destacam. As três primeiras se relacionam a temas sexuais e corporais, enquanto as demais a emoções negativas. Embora 'bio' e 'body' sejam neutras em outros contextos, aqui aparecem com frequência elevada, sugerindo um viés cultural e linguístico.

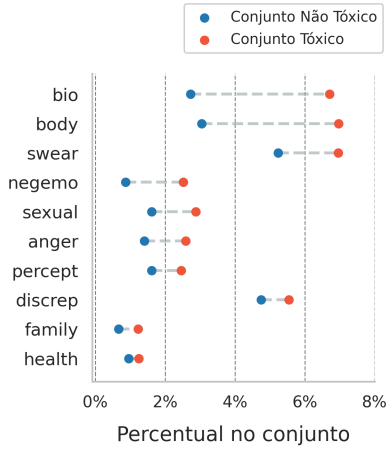


Figura 2: Top 10 classes como maior variação de ocorrências entre o conjunto de sentenças tóxicas e não tóxicas da base ToLDE-Br

Tabela 4: HateXplain - Classes com maior proporção de ocorrências e suas palavras mais comuns.

Classe LIWC	Palavras mais frequentes
swear	*n*gger, r*tard, b*tch, f*ggot, f*ck, f*cking, sh*t*
anger	*b*tch, f*ck, f*cking, sh*t, rape, hate, kill*
focuspresent	*be, have, do, can, get*
body	*sh*t, *ss, d*ck, head, fat, wear, shoe, eye*
relig	*jews, muslim, jew, islam, god, christian*
feel	*sand, feel, hard, hand, burn, skin, hot*
discrep	*if, would, want, need, should*
reward	*get, good, take, well, great*
sexual	*f*ck, f*cking, r*pe, gay, d*ke, qu*er*

As análises revelam padrões distintos entre as bases. ToLDE-Br mostra prevalência de termos ligados a aspectos fisiológicos e corporais, enquanto HateXplain destaca insultos explícitos.

As Tabelas 4 e 5 mostram as classes com maior proporção de ocorrências nas justificativas de toxicidade, acompanhadas das palavras mais representativas em cada categoria.

As tabelas mostram como a toxicidade se manifesta de formas distintas em cada corpus. No HateXplain, predominam ataques diretos com termos agressivos e identitários, notadamente nas classes *swear*, *anger*, *relig* e *sexual*. Já na ToLDE-Br, observa-se uma ênfase em insultos genéricos e expressões vulgares, muitas vezes envolvendo termos corporais e familiares, refletindo padrões específicos do português brasileiro.

4.2 Desempenho dos Modelos na Classificação de Toxicidade

Para investigar se o desempenho dos modelos de detecção de toxicidade é relevante para as bases de dados estudadas, avaliamos

Tabela 5: ToLDE-Br - Classes com maior proporção de ocorrências e suas palavras mais comuns.

Classe LIWC	Palavras mais frequentes
bio	*p*ta, tomar, filho, viar, olhar, mão*
body	*c*, ter, c*ralho, seu, p*u, carar, mão*
swear	*filho, dar, filha, fdp, como, mano, coisa*
negemo	*p*ta, chato, pedir, b*baco, problema, preocupar*
sexual	*p*rra, ficar, como, f*der, pegar, amo*
anger	*f*der, matar, desgr*çar, ódio, nego, mau, ruim*
discrep	*que, se, querer, dar, desse*
percept	*filho, carar, filha, surrar, conhecer*
family	*filho, filha, mano, mãe, pai*
health	*bocar, vida, vader, lavar, mal, ruim*

seus resultados preditivos. Como os modelos geram uma pontuação de toxicidade contínua entre 0 e 1 para cada sentença, adotamos um limiar de classificação de 0.5, conforme estabelecido por [25]. O desempenho foi avaliado por meio das métricas de Acurácia, Precisão, Revocação e AUC (*Area Under the Curve*).

As Tabelas 6 e 7 apresentam os resultados obtidos.

Tabela 6: Resultados de desempenho no conjunto de dados HateXplain.

Modelo	Acurácia	Precisão	Revocação	AUC
Detoxify (Org)	0.65	0.71	0.69	0.68
Detoxify (Unb)	0.63	0.70	0.66	0.66
Perspective (En)	0.67	0.72	0.73	0.72

Tabela 7: Resultados de desempenho no conjunto de dados ToLDE-Br.

Modelo	Acurácia	Precisão	Revocação	AUC
Detoxify (Mul)	0.86	0.89	0.82	0.86
Perspective (Pt)	0.86	0.79	0.99	0.86

A análise dos resultados revela padrões de desempenho distintos entre os dois conjuntos de dados. No HateXplain (Tabela 6), os modelos apresentaram desempenho moderado. O **Perspective (En)** obteve os melhores resultados, com AUC de 0.72. Notavelmente, o modelo Detoxify (Unb), que passou por ajuste para não propagar viés discriminativo, teve desempenho ligeiramente inferior à sua versão original.

Em contrapartida, no ToLDE-Br (Tabela 7), ambos os modelos demonstraram performance robusta, alcançando acurácia de 86% e AUC de 0.86. A principal diferença reside no *trade-off* entre precisão e revocação. O **Perspective (Pt)** alcançou revocação excepcional de 99%, indicando alta sensibilidade para identificar conteúdo tóxico, embora com precisão menor (79%). Por outro lado, o **Detoxify (Mul)** destacou-se pela alta precisão (89%), tornando-se mais eficaz em evitar classificações de falsos positivos, ainda que com revocação ligeiramente inferior (82%). Essa dicotomia sugere que a escolha do modelo pode depender da prioridade da aplicação: minimizar o conteúdo tóxico não detectado (Perspective) ou minimizar as acusações incorretas de toxicidade (Detoxify).

4.3 Avaliação da Qualidade das Justificativas

Avaliamos o alinhamento das explicações geradas pela técnica de explicabilidade SHAP com as justificativas fornecidas por anotadores humanos. O objetivo é medir o alinhamento dos explicadores combinados aos modelos em identificar as mesmas palavras que os humanos apontam como fonte de toxicidade.

Metodologia de Avaliação. Para a comparação, o conceito de **justificativa** é central. Tanto a anotação humana quanto a explicação do modelo para uma sentença são convertidas em um vetor binário, onde cada posição corresponde a uma palavra, com '1' indicando uma fonte de toxicidade e '0' caso contrário. Para permitir a comparação, os valores de contribuição contínuos que o SHAP atribui a cada palavra foram convertidos para este formato binário por meio de um processo de limiarização.

Com os vetores binarizados, comparamos a justificativa do modelo-explicador com a do avaliador humano, palavra por palavra. Cada palavra foi classificada como verdadeiro positivo (*vp*), falso positivo (*fp*) ou falso negativo (*fn*), conforme ilustra a Tabela 8.

Tabela 8: Exemplo compacto de comparação de justificativas para a sentença 'Que ideia idiota'.

Palavra	Que	ideia	idiota
Avaliador Humano	0	1	1
Modelo-Explicador	1	0	1
Classificação	*fp*	*fn*	*vp*

A avaliação final foi conduzida por meio das métricas de Precisão, Revocação e Coeficiente F1, calculadas a partir da contagem total de *vp*, *fp* e *fn* para cada sentença. A análise concentrou-se em sentenças originalmente anotadas como tóxicas por humanos, uma vez que estas são as únicas que possuem justificativas em nível de palavra para servirem como referência. Por essa limitação, a avaliação das explicações foi organizada em dois cenários distintos: **Verdadeiros Positivos (VP)**, quando o modelo acerta a classificação da sentença, e **Falsos Negativos (FN)**, quando o modelo falha na detecção da toxicidade. Consequentemente, não foi possível conduzir uma análise comparativa similar para os casos de Falsos Positivos (FP) e Verdadeiros Negativos (VN), dada a ausência de anotações-base para as sentenças não tóxicas.

Cabe destacar que, ao longo deste trabalho, adotamos a convenção de utilizar siglas em maiúsculas (VP, FN) para nos referirmos aos acertos e erros na classificação de sentenças como um todo, enquanto as siglas em minúsculas (*vp*, *fp*, *fn*, *vn*) indicam a avaliação no nível de palavras, no contexto das justificativas de toxicidade.

Adicionalmente, para contextualizar os resultados e validar o desempenho dos explicadores, foi gerado um baseline de comparação denominado **Justificativas Aleatórias**. Para garantir uma comparação justa, a quantidade de palavras marcadas em cada justificativa aleatória foi determinada por um modelo de regressão linear, que aprendeu a correlação entre o comprimento da sentença e o número de palavras que os humanos tipicamente marcam como tóxicas. Em seguida, as palavras eram selecionadas aleatoriamente de acordo com a quantidade estimada.

4.3.1 Resultados no Dataset HateXplain. A Tabela 9 apresenta o desempenho das justificativas para este dataset. A análise dos dados mostra que o SHAP supera o baseline aleatório, validando sua utilidade.

Tabela 9: HateXplain - Métricas das explicações geradas pelo SHAP.

Modelo e Cenário	Precisão	Revocação	F1-Score
Justificativas Aleatórias	0.17 ± 0.01	0.13 ± 0.00	0.13 ± 0.00
VP Detoxify (Org)	0.47 ± 0.01	0.74 ± 0.01	0.51 ± 0.01
VP Detoxify (Unb)	0.45 ± 0.01	0.71 ± 0.01	0.49 ± 0.01
VP Perspective (En)	0.54 ± 0.01	0.79 ± 0.01	0.58 ± 0.01
FN Detoxify (Org)	0.28 ± 0.01	0.44 ± 0.01	0.28 ± 0.01
FN Detoxify (Unb)	0.33 ± 0.01	0.44 ± 0.01	0.32 ± 0.01
FN Perspective (En)	0.51 ± 0.01	0.68 ± 0.01	0.51 ± 0.01

Comparando os dois explicadores, o **SHAP consistentemente apresentou maior revocação**, indicando melhor capacidade de identificar o conjunto de palavras tóxicas. O **Perspective (En)**, combinado com SHAP no cenário VP, alcançou o maior F1-Score (0.58). Como esperado, o desempenho de ambos os explicadores é consideravelmente inferior nos casos FN, onde o modelo falha em classificar a sentença, prejudicando também a qualidade da explicação.

4.3.2 Resultados no ToLDE-Br (Português). Os resultados para o dataset ToLDE-Br estão na Tabela 10. O cenário FN para o modelo Perspective (Pt) não foi avaliado devido ao número insuficiente de amostras, resultado de sua alta revocação (99%) na classificação de sentenças.

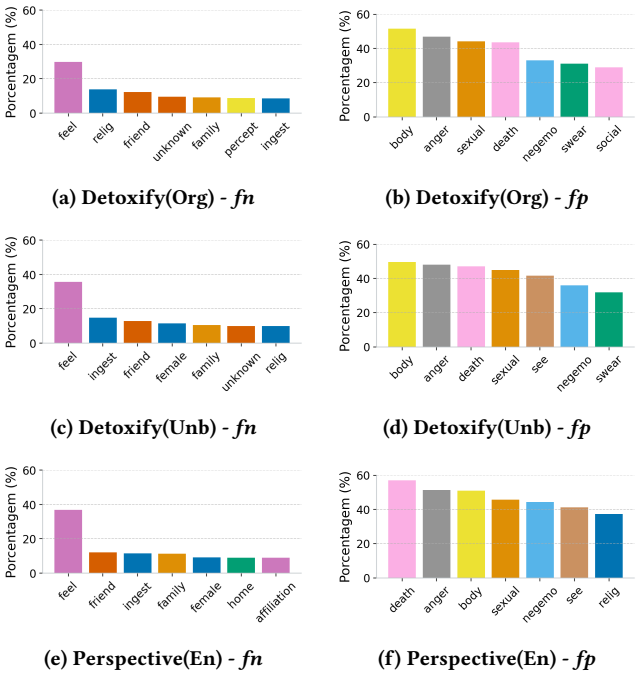
Tabela 10: ToLDE-Br - Métricas das explicações geradas pelo SHAP.

Modelo e Cenário	Precisão	Revocação	F1-Score
Justificativas Aleatórias	0.28 ± 0.03	0.24 ± 0.02	0.24 ± 0.02
VP Detoxify (Mul)	0.59 ± 0.03	0.66 ± 0.03	0.56 ± 0.02
VP Perspective (Pt)	0.48 ± 0.03	0.51 ± 0.03	0.43 ± 0.02
FN Detoxify (Mul)	0.63 ± 0.08	0.62 ± 0.07	0.54 ± 0.07

O achado mais relevante é a desconexão entre o desempenho do modelo e a qualidade de sua explicação. O modelo **Perspective (Pt)**, apesar de sua performance quase perfeita na detecção de sentenças tóxicas (revocação de 99%), não produziu consistentemente as melhores justificativas. Isso sugere que a alta acurácia de um modelo não garante que ele fundamente suas decisões em razões alinhadas com a percepção humana.

4.4 Análise das Justificativas pelo LIWC

Nesta seção, investigamos os padrões de erro apresentados por modelos de detecção de toxicidade a partir da análise de suas justificativas. Para isso, examinamos as explicações geradas pela técnica SHAP para dois cenários distintos: sentenças corretamente classificadas como tóxicas (VP) e sentenças tóxicas que os modelos não conseguiram identificar (FN). Com o auxílio do léxico LIWC, categorizamos as palavras em que houve divergência entre a justificativa



(a) Detoxify(Org) - *fn* **(b) Detoxify(Org) - *fp***

(c) Detoxify(Unb) - *fn* **(d) Detoxify(Unb) - *fp***

(e) Perspective(En) - *fn* **(f) Perspective(En) - *fp***

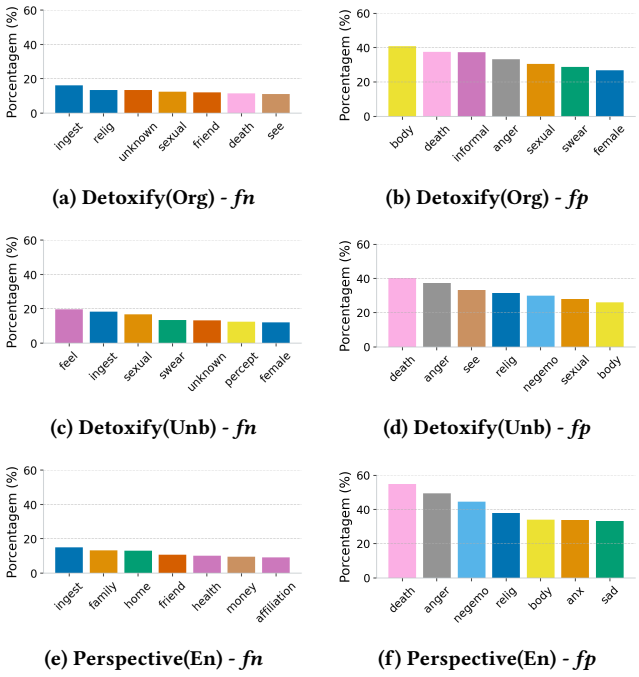
Figura 3: Proporção de erros de justificativa por classe LIWC em sentenças VP no HateXplain. Colunas da esquerda: erros *fn*; Direita: erros *fp*.

do modelo e a anotação humana, com foco em dois tipos principais de erro em nível lexical. O primeiro caso corresponde aos *fn*, que ocorrem quando palavras reconhecidas como tóxicas pelos anotadores humanos são ignoradas pelo modelo, evidenciando casos que o sistema falha em reconhecer. O segundo caso abrange os *fp*, nos quais o modelo atribui toxicidade a palavras que não foram apontadas como fonte da ofensa pelos humanos, revelando padrões de atenção indevida e possíveis vieses nos critérios de decisão do modelo.

4.4.1 Análise dos Erros no Dataset HateXplain. A Figura 3 apresenta as classes LIWC com maior proporção de erros de justificativa (*fn* e *fp*) para sentenças VP no HateXplain.

Para os erros *fn*, a classe 'feel' se destaca. Uma análise mais profunda revelou que este resultado é artificialmente inflado pela alta frequência da palavra 'sand', utilizada na ofensa racial 'sand n*gger', que os modelos consistentemente falham em identificar como uma unidade tóxica. Outras classes com erros *fn* notáveis, como 'ingest' e 'relig', continham gírias pejorativas como 'beaner' e ofensas a grupos religiosos ('jew', 'islam'), indicando dificuldade dos modelos em reconhecer toxicidade contextual e culturalmente específica.

Para os erros *fp*, os resultados são mais consistentes e reveladores. Nos três modelos, as classes com maior taxa de erro foram 'anger', 'body', 'death', 'negemo' (emoção negativa) e 'sexual'. Isso demonstra forte tendência dos modelos em focar excessivamente em palavras com carga negativa ou sexual inerente, mesmo que não sejam a real fonte da ofensa na sentença.



(a) Detoxify(Org) - *fn* (b) Detoxify(Org) - *fp*

(c) Detoxify(Unb) - *fn* (d) Detoxify(Unb) - *fp*

(e) Perspective(En) - *fn* (f) Perspective(En) - *fp*

Figura 4: Proporção de erros de justificativa por classe LIWC em sentenças FN no HateXplain. Colunas da esquerda: erros *fn*; Direita: erros *fp*.

Resultados similares foram observados ao analisar as justificativas de sentenças FN, conforme a Figura 4.

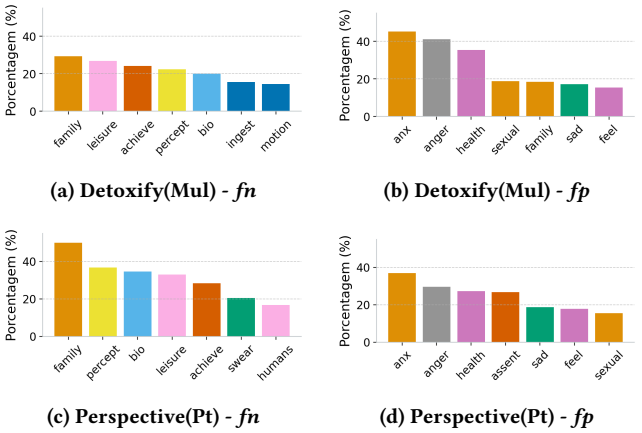
A Tabela 11 apresenta exemplos nos quais há divergência entre o foco identificado pelos anotadores humanos e o fornecido pelo explicador SHAP. Na linha 1, observa-se que os avaliadores humanos marcaram como fonte da toxicidade os termos *'gay'*, *'jew'* e *'f*ggot'*, enquanto o modelo-explicador focou apenas em *'gay'* e *'f*ggot'*, desconsiderando o termo de classe religiosa *'jew'* como parte da expressão ofensiva. Na linha 2, o termo *'beaner'*, uma ofensa direcionada a pessoas de ascendência mexicana, não recebeu destaque do modelo, que atribuiu maior importância a palavras de menor carga ofensiva, como *'is'* e *'bitching'*. Na linha 3, a sentença contém termos altamente ofensivos como *'k*ke'* (ofensa a pessoas de ascendência judaica) e *'sand n*ggers'* (ofensa a pessoas de origem árabe); no entanto, o modelo-explicador falhou ao não considerar *'sand'* como parte relevante da construção, o que compromete a interpretação correta da ofensa. Esses casos evidenciam fragilidades na capacidade do modelo em replicar as preferências humanas quanto aos elementos que tornam uma sentença tóxica.

4.4.2 Análise dos Erros no Dataset ToLDE-Br. A análise para o ToLDE-Br segue a mesma metodologia usada no HateXplain, identificando as categorias semânticas (classes LIWC) onde os erros de justificativa dos modelos são mais frequentes. A investigação se concentrou em sentenças classificadas como Verdadeiro Positivo (VP), pois não havia dados suficientes para uma análise robusta das sentenças do tipo Falso Negativo (FN).

A Figura 5 apresenta as classes LIWC com maior proporção de erros de justificativa (*fn* e *fp*) para os modelos em português.

Tabela 11: Exemplos de falhas de justificativa no HateXplain.

Nº	Sentença com Justificativa (Modelo)
1	*better call the anti defamation league you gay jew f*ggot* (Detoxify Org.)
2	*like i care < tong > is b*tching to a stupid beaner like me going to fix that nope good luck* (Detoxify Unb.)
3	*either way k*kes and sand n*ggers will most likely die this is very good* (Perspective En.)



(a) Detoxify(Mul) - *fn* (b) Detoxify(Mul) - *fp*

(c) Perspective(Pt) - *fn* (d) Perspective(Pt) - *fp*

Figura 5: Proporção de erros de justificativa por classe LIWC em sentenças VP no ToLDE-Br. Colunas da esquerda: erros *fn*; Direita: erros *fp*.

Analisando primeiramente os erros de **falso negativo (*fn*)**, destaca-se a incapacidade dos modelos de entenderem a toxicidade composicional, específica da língua portuguesa. Classes como 'family', 'leisure' e 'percept' apresentaram altas taxas de erro porque são frequentemente associadas às palavras 'filho' e 'tomar'.

Tabela 12: Exemplos de falhas de justificativa no ToLDE-Br.

Nº	Sentença com Justificativa do Modelo
1	a quantidade de incêndios a acontecerem é absurdo . aos filhos da p*ta que põem fogos , só espero q um dia ardam vivos lá no meio também
2	@user vai toma no c* raça desgraçada
3	*alguém segura esse v*ad*nho do c*ralho pq se não eu vou dar na cara dele*

As sentenças exemplificadas na Tabela 12 mostram casos de sentenças VP, ilustrando diferentes tipos de falhas na atribuição de justificativas por parte do modelo, especialmente em contextos nos quais a toxicidade é expressa por meio de construções compostas ou contextuais.

Na sentença 1, observa-se uma identificação incompleta da expressão ofensiva composta 'filhos da p*ta'. Embora o termo 'da

p*ta' seja corretamente identificado, o termo 'filhos' não recebeu a mesma atenção, o que sugere que o modelo não reconhece a expressão como uma unidade semântica ofensiva, e se focou apenas em um marcador. Além disso, a atenção errada dos termos 'põem' e 'fogos' indica uma interpretação literal da ação, desconsiderando o contexto da sentença que está gerando o sentido tóxico.

Na sentença 2, repete-se o padrão de foco isolado em termos com maior carga negativa. A expressão composta 'tomar no c*' é desmembrada, com 'c*' sendo o único termo identificado como ofensivo. Isso reforça a limitação do modelo em compreender expressões idiomáticas ou insultos compostos. Além disso, o termo 'raça' é ignorado, apesar de, em conjunto com 'desgraçada', contribuir para o teor discriminatório da sentença.

Já na sentença 3, observa-se uma falha com o termo 'vi*dinho', comumente associado a discurso homofóbico, não é identificado pelo par modelo-explicador, o que indica falha na identificação de termos ofensivos contextualmente relevantes. Por outro lado, os termos 'dar' e 'cara' são classificados como *fp*, apesar de estarem inseridos em uma ameaça explícita, sugerindo um viés do modelo na supervalorização de palavras associadas à violência física, mesmo quando seu uso pode ser ambíguo.

De forma geral, as sentenças analisadas reforçam a limitação dos modelos em capturar a *toxicidade contextual*, com forte tendência a focar em palavras com carga negativa isolada, como *c**, *p*ta* ou *c*ralho*. Este comportamento está alinhado com os resultados quantitativos discutidos anteriormente, que apontaram elevadas taxas de erro para classes como *anger*, *swear* e *sexual*. Além disso, expressões idiomáticas, gírias ou construções compostas são frequentemente tratadas de forma fragmentada, comprometendo a precisão das justificativas atribuídas.

A análise agregada dos resultados revela um paradoxo central na capacidade dos modelos de detecção de toxicidade: a distinção entre o desempenho classificatório e a qualidade da justificação. Embora os modelos, especialmente no conjunto ToLDE-Br, demonstrem alta performance em métricas como acurácia e AUC, a análise de suas justificativas expõe uma compreensão superficial e frágil do que constitui a toxicidade.

Fica evidente que os modelos se apoiam fortemente em marcadores lexicais isolados, comumente associados a categorias como `anger`, `sexual` e `swear`. Essa dependência excessiva leva a duas falhas principais: (i) a incapacidade de reconhecer toxicidade que emerge do contexto composicional, como visto nas expressões em português ("filho da p*ta", "tomar no c*") e em gírias culturais no inglês ("sand n*gger"); e (ii) a atribuição incorreta de toxicidade a palavras que, embora possuam carga negativa, não são a fonte da ofensa na sentença.

O achado mais significativo é a desconexão observada no modelo Perspective (Pt), que, apesar da revocação quase perfeita (99%), produziu justificativas de qualidade inferior ao Detoxify (Mul). Isso confirma que a alta acurácia de um modelo não é garantia de que suas decisões sejam fundamentadas em um raciocínio alinhado à percepção humana. Em suma, os modelos são eficazes em "o quê" classificar, mas falham em explicar "porquê" o fazem de maneira coerente.

5 CONCLUSÕES

Este estudo investigou as limitações de modelos de detecção de toxicidade, confirmando a hipótese de que eles apresentam falhas significativas ao atribuir atenção desproporcional a palavras-chave isoladas, em detrimento da interpretação contextual. Por meio da aplicação de técnicas de explicabilidade (SHAP) e análise semântica (LIWC), demonstramos que, embora os modelos alcancem alto desempenho classificatório, seu processo de decisão subjacente é superficial e não reflete o raciocínio humano.

Nossa principal contribuição foi revelar os padrões sistemáticos de erro desses modelos. Identificamos uma forte tendência em focar excessivamente em termos com carga negativa ou sexual inerente (e.g., categorias LIWC `anger`, `body`, `sexual`), resultando em falsos positivos na atribuição de importância. Além disso, expusemos a incapacidade dos modelos em lidar com a toxicidade composicional, onde o sentido ofensivo emerge da combinação de palavras, um problema evidente em expressões idiomáticas no português, e com a toxicidade culturalmente específica, como gírias e insultos raciais no inglês que não são universalmente óbvios. A desconexão entre a alta acurácia classificatória e a baixa qualidade das justificativas reforça que otimizar apenas para a classificação correta pode mascarar uma compreensão fundamentalmente falha do problema.

Uma limitação deste trabalho foi a escassez de amostras de falsos negativos com justificativas humanas no dataset ToLDE-Br, o que restringiu uma análise mais aprofundada dos erros de omissão para os modelos em português. A criação de conjuntos de dados com anotações mais ricas e diversas continua sendo um desafio para a área.

Como trabalhos futuros, sugerimos duas frentes principais. Primeiramente, aprofundar a análise da estrutura semântica das ofensas, indo além das palavras para compreender as construções sintáticas e pragmáticas que geram toxicidade. Em segundo lugar, explorar o potencial dos *Large Language Models* (LLMs), cuja capacidade avançada de compreensão contextual pode superar as limitações dos modelos baseados em classificadores tradicionais. O desenvolvimento da próxima geração de ferramentas de moderação de conteúdo necessita de uma abordagem que integre não apenas avanços técnicos em Processamento de Linguagem Natural, mas também uma compreensão sofisticada das nuances sociais e culturais da comunicação humana.

AGRADECIMENTOS

O apoio a esta pesquisa foi fornecido pelos projetos CNPq INCT-TILD-IAR (nº 408490/2024-1), CNPq INCT-IACiber (nº 408432/2024-1) e CNPq IRADA (nº 405572/2025-5). Agradecemos também à Capes pelo apoio ao Programa de Pós-Graduação em Ciência da Computação da UFMG.

REFERÊNCIAS

[1] Hind Almerekhi, Haewoon Kwak, Bernard J Jansen, and Joni Salminen. 2019. Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. 291–292.

[2] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. 2020. Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference 2020*. 3033–3040.

[3] Sultan Alshamrani, Mohammed Abuhamad, Ahmed Abusnaina, and David Mohaisen. 2020. Investigating Online Toxicity in Users Interactions with the Mainstream Media Channels on YouTube.. In *CIKM (Workshops)*.

[4] Kofi Arhin, Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Moninder Singh. 2021. Ground-Truth, Whose Truth?–Examining the Challenges with Annotating Toxic Text Datasets. *arXiv preprint arXiv:2112.03529* (2021).

[5] Danah M Boyd and Nicole B Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication* 13, 1 (2007), 210–230.

[6] Athus Cavalini, Thamya Donadia, and Giovanni Comarela. 2024. Characterizing the toxicity of the Brazilian extremist communities on telegram. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC, 370–374.

[7] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

[8] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[9] Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 6786–6794.

[10] Priya Garg, MK Sharma, and Parteek Kumar. 2024. Improving Hate Speech Classification Through Ensemble Learning and Explainable AI Techniques. *Arabian Journal for Science and Engineering* (2024), 1–14.

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

[12] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is"love"evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*. 2–12.

[13] Samuel S. Guimarães, Filipe N. Ribeiro, Julio C. S. Reis, and Fabrício Benevenuto. 2020. Characterizing Toxicity on Facebook Comments in Brazil. In *Proceedings of the 26th Brazilian Symposium on Multimedia and the Web (WebMedia '20)*. Association for Computing Machinery (ACM), São Luís, Brazil, 1–10. https://doi.org/10.1145/3428658.3430974

[14] David Gunning. 2017. Explainable artificial intelligence (xai). darpa. *I20 (DARPA 2017)* (2017).

[15] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

[16] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[17] Jigsaw/ConversationAI. 2018. Toxic Comment Classification Challenge. https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

[18] Jigsaw/ConversationAI. 2019. Jigsaw Unintended Bias in Toxicity Classification. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification

[19] Jigsaw/ConversationAI. 2020. Jigsaw Multilingual Toxic Comment Classification. https://www.kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification

[20] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921* (2020).

[21] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 299–318.

[22] Ritesh Kumar, Atul Kr Ojha, Marcos Zampieri, and Shervin Malmasi. 2018. Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

[23] Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3197–3207.

[24] Joao A Leite, Diego F Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543* (2020).

[25] Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. Self-detoxifying language models via toxification reversal. *arXiv preprint arXiv:2310.09573* (2023).

[26] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14867–14875.

[27] Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms* 15, 8 (2022), 291.

[28] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*. Springer, 51–56.

[29] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com. 22 pages.

[30] Alexandra Olteanu, Kartik Talamadupula, and Kush R Varshney. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on web science conference*. 405–406.

[31] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report. University of Texas at Austin. https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content

[32] Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*. 137–143.

[33] Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*. 6659–6669.

[34] Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*. 1058–1065.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[36] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. ACM, 1391–1399.

[37] Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*. 500–507.