Emojis e Discurso de Ódio no Contexto Brasileiro: Uma Análise a Partir de Diferentes Plataformas Sociais

Thúlio M. O. Fernandes[1], **Paula T. M. Gibrim**[1], **Julio C. S. Reis**[1]

[1]Departamento de Informática – Universidade Federal de Viçosa (UFV) – Brasil

thulio.fernandes@ufv.br, paula.gibrim@ufv.br, jreis@ufv.br

Abstract. *The widespread adoption of social media in Brazil has established platforms such as Instagram, YouTube, and Twitter/X as central spaces for interaction, but also as channels for the dissemination of hate speech. While computational approaches for its detection have advanced, the role of emojis remains underexplored in the national context. This work investigates the relationship between emoji usage and hate speech in Brazilian Portuguese (PT-BR) using a dataset of over 30,000 labeled messages from different platforms, analyzed with Natural Language Processing (NLP) techniques. The results indicate that some emojis are consistently associated with hate speech, while others exhibit context-dependent variability. These findings highlight the potential of emojis as complementary signals in automated detection models.* ***Warning! This work and the referenced data contain examples of potentially offensive and hateful language.***

Resumo. *A massificação das plataformas sociais no Brasil consolidou ambientes como Instagram, YouTube e Twitter/X como espaços centrais de interação, mas também de disseminação de discurso de ódio. Embora haja avanços em abordagens computacionais para sua detecção, o papel dos emojis ainda é pouco explorado no cenário nacional. Este trabalho investiga a relação entre emojis e discurso de ódio em português (PT-BR) a partir de um conjunto de mais de 30 mil mensagens rotuladas, analisadas com técnicas de Processamento de Linguagem Natural (PLN). Os resultados indicam que alguns emojis apresentam associação consistente com o discurso de ódio, enquanto outros variam conforme o contexto. Essas descobertas destacam o potencial dos emojis como sinal complementar em modelos automáticos de detecção.* ***Atenção! Este trabalho e os dados referenciados contêm exemplos de linguagem potencialmente ofensiva e odiosa.***

1. Introdução

A presença massiva de usuários nas plataformas sociais constitui um fenômeno marcante da comunicação digital contemporânea, impulsionado pela ampliação do acesso à Internet [Bertot et al. 2012]. No Brasil, esse cenário também é evidente, consolidando plataformas como Instagram, YouTube e Twitter/X como importantes espaços de interação, produção de conteúdo e debate público [de Santana et al. 2009, Guimarães et al. 2022, Caetano et al. 2022, Pinto et al. 2024]. Assim como em qualquer interação humana, a comunicação nessas plataformas ocorre de forma heterogênea e, muitas vezes, conflituosa [Moreira et al. 2026]. Nesse contexto, o discurso de ódio — compreendido como manifestações que atacam indivíduos ou grupos com base em características como raça,

gênero, religião ou orientação sexual [Fortuna and Nunes 2018] — tem se proliferado de forma preocupante. Para enfrentar esse problema, as plataformas e a comunidade científica têm adotado abordagens baseadas em Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN), com resultados promissores em diferentes idiomas [Biere et al. 2018, Silva and Serapião 2018].

No entanto, as interações digitais deixaram de ser exclusivamente textuais, passando a incorporar elementos visuais como imagens, stickers [de Freitas Melo et al. 2025] e, principalmente, emojis [Aldunate and González-Ibáñez 2017]. Com origem no Japão, os emojis são símbolos utilizados em mensagens eletrônicas que evoluíram de simples representações faciais para abranger objetos, lugares e conceitos abstratos [Rodrigues 2025]. Funcionando como extensões das palavras, esses símbolos permitem a transmissão de emoções e intenções, suprindo limitações inerentes à comunicação escrita e atuando como mecanismo de autoexpressão que influencia a percepção social do emissor [Rodrigues 2025]. Estudos recentes demonstram que emojis podem ser explorados como características complementares em sistemas de detecção de discurso de ódio, contribuindo para a interpretação de termos ambíguos e a distinção entre ofensas reais e usos irônicos [Ibrohim et al. 2019, Althobaiti 2022]. Apesar desse potencial, ainda são escassas as pesquisas sobre a relação entre emojis e discurso de ódio no contexto brasileiro, especialmente em português (PT-BR). A maior parte da literatura nacional foca em modelos textuais de classificação binária [Leite et al. 2020, Trajano et al. 2024, Salles et al. 2025], dedicando pouca atenção ao papel de elementos visuais — lacuna ainda mais acentuada em análises comparativas entre plataformas, que potencializam o uso de emojis para fins distintos.

Diante desse cenário, esta pesquisa busca analisar a relação entre o uso de emojis e o discurso de ódio em plataformas sociais no Brasil, contribuindo para o desenvolvimento de modelos de detecção mais adequados ao contexto nacional. Mais especificamente, este trabalho tem como objetivo responder às seguintes questões de pesquisa (QPs): ***QP1:*** *Quais emojis aparecem com maior frequência em mensagens rotuladas como discurso de ódio?* ***QP2:*** *A associação entre emojis e discurso de ódio varia entre plataformas (Instagram, Twitter/X, YouTube)?* ***QP3:*** *Existem emojis que se comportam como indicadores consistentes de ódio independentemente da plataforma?*

A partir dessas questões, este trabalho apresenta as seguintes contribuições principais: (i) uma análise empírica do uso de emojis em mensagens rotuladas como discurso de ódio em três grandes plataformas sociais utilizadas no Brasil, a saber: Instagram, Twitter/X e YouTube; (ii) uma comparação entre essas plataformas quanto à associação entre emojis e semântica odiosa; e (iii) a identificação de símbolos que atuam como indicadores consistentes de discurso de ódio independentemente do ambiente digital.

Em resumo, os resultados mostram que determinados emojis aparecem com maior frequência em mensagens classificadas como discurso de ódio, evidenciando padrões de associação relevantes e funcionando como marcadores contextuais que reforçam a intenção ofensiva das mensagens (QP1). Observa-se, ainda, que essa relação varia entre plataformas, refletindo diferenças contextuais entre ambientes. Enquanto o Instagram apresenta a maior diversidade e intensidade de uso de emojis ofensivos, no Twitter/X o discurso de ódio baseia-se mais em elementos textuais do que em sinais visuais (QP2).

Por fim, embora não tenham sido identificados emojis universalmente associados ao ódio nas três plataformas simultaneamente, alguns símbolos apresentam consistência parcial entre os ambientes analisados (QP3). Em conjunto, essas descobertas reforçam o papel dos emojis como elementos relevantes na caracterização do discurso de ódio e evidenciam seu potencial como sinal complementar no desenvolvimento de abordagens de detecção mais robustas e sensíveis às nuances da comunicação digital contemporânea.

O restante do trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados destacando a lacuna de pesquisa preenchida pelo presente estudo; a Seção 3 descreve a metodologia adotada, incluindo a estratégia de seleção, tratamento e análise de dados; a Seção 4 apresenta e discute os principais resultados; e, por fim, a Seção 5 conclui o trabalho e aponta direções para pesquisas futuras.

2. Trabalhos Relacionados

A literatura recente sobre detecção de discurso de ódio tem evoluído de análises puramente textuais para abordagens multimodais que reconhecem a importância de elementos visuais, como os emojis, na construção do sentido e da intenção do autor [Grosz et al. 2023]. Nesta seção, revisamos os principais trabalhos relacionados a esta pesquisa, organizados em três dimensões: (i) o papel psicológico e identitário dos emojis na comunicação digital; (ii) experiências de detecção multimodal em outros idiomas, e (iii) iniciativas de detecção de discurso de ódio em português brasileiro.

2.1. Perspectivas Psicológicas e Identitárias dos Emojis

Os emojis têm sido amplamente estudados como ferramentas de autoexpressão e construção de identidade na comunicação mediada por tecnologia, compensando a ausência de pistas não verbais como tom de voz e expressões faciais [Grosz et al. 2023]. Nesse sentido, [Rodrigues 2025] destaca que a utilização de símbolos considerados amigáveis — como rostos sorridentes ou corações — impacta diretamente a percepção social do remetente, fazendo com que ele seja visto como mais extrovertido, agradável e empático. Entretanto, o autor ressalta que essa percepção é sensível ao contexto: o uso excessivo pode transmitir uma impressão de insinceridade, enquanto o uso estratégico serve para suavizar tensões e reforçar entusiasmo, atuando como mecanismo de direcionamento do impacto emocional da mensagem [Rodrigues 2025]. Essa ambiguidade semântica é central para o presente trabalho, uma vez que emojis aparentemente inofensivos podem adquirir conotações ofensivas dependendo do contexto em que são empregados.

2.2. Detecção Multimodal em Outros Idiomas

No âmbito do indonésio, [Ibrohim et al. 2019] trataram a detecção de discurso de ódio como um problema de classificação multi-rótulo, permitindo que uma mesma mensagem fosse simultaneamente classificada como discurso de ódio e linguagem abusiva. Os autores combinaram modelos tradicionais de aprendizado de máquina — SVM, *Random Forest* e Regressão Logística — com características extraídas de unigramas, etiquetas morfossintáticas (PoS) e emojis. Os resultados demonstraram que a inclusão de emojis e PoS auxiliou a abordagem na interpretação de termos ambíguos, como nomes de animais utilizados como insultos, e na distinção entre ofensas reais e interações jocosas entre usuários. O modelo de Regressão Logística com todas as características alcançou acurácia de 79,85%.

Em árabe, [Althobaiti 2022] investigou a detecção de discurso de ódio em nível fino (*fine-grained*), categorizando ofensas em classes como raça, religião, ideologia e gênero. O estudo diferenciou-se ao utilizar o modelo de aprendizado profundo BERT (`AraBERTv0.2-Twitter-base`) e ao incorporar descrições textuais de emojis e análise de sentimento como recursos adicionais. Ao converter emojis em representações textuais, o modelo obteve limites mais claros para distinguir tweets normais de ofensivos, atingindo F1-score de 84,3% na detecção ofensiva. Ambos os estudos evidenciam que emojis constituem características distintivas relevantes para abordagens de monitoramento automático de conteúdo.

2.3. Detecção de Discurso de Ódio em Português Brasileiro

No contexto nacional, diferentes esforços têm sido realizados para a criação de recursos e modelos voltados à detecção de discurso de ódio em PT-BR [Braga et al. 2020, Moreira et al. 2026]. [Fortuna et al. 2019], por exemplo, disponibilizaram um dos primeiros repositórios de dados hierarquicamente anotados em português, coletado do Twitter/X. Por outro lado, [Leite et al. 2020] ampliaram esse esforço com o TOLD-BR, um corpus focado em toxicidade e grupos minoritários, também oriundo do Twitter/X. Mais recentemente, [Trajano et al. 2024] e [Salles et al. 2025] contribuíram com bases de dados voltadas ao YouTube e ao Instagram, respectivamente, ampliando a cobertura para além do Twitter/X. Por fim, [de Oliveira et al. 2024] apresentaram análise abrangendo múltiplos domínios temáticos.

Apesar desses avanços, a maior parte das abordagens nacionais concentra-se em modelos textuais de classificação binária, dedicando pouca atenção ao papel de elementos visuais como os emojis. Análises comparativas entre plataformas são ainda mais escassas, o que representa uma lacuna relevante considerando as diferenças estruturais e comunicativas entre ambientes como Instagram, YouTube e Twitter/X, que consistem em plataformas sociais propostas para diferentes finalidades.

2.4. Síntese e Lacunas

Em suma, os estudos revisados evidenciam que emojis podem atuar como sinais complementares relevantes na detecção de discurso de ódio, tanto em cenários multilíngues quanto em contextos culturalmente específicos [Ibrohim et al. 2019, Althobaiti 2022]. No entanto, sua exploração no contexto brasileiro permanece incipiente. Este trabalho busca preencher essa lacuna ao investigar, de forma empírica e comparativa, a associação entre emojis e discurso de ódio em três plataformas amplamente utilizadas no Brasil, contribuindo tanto para a caracterização desse fenômeno quanto para o desenvolvimento, no futuro, de modelos de detecção mais adequados ao contexto nacional.

3. Metodologia

Esta seção descreve os procedimentos metodológicos adotados no estudo. A Figura 1 apresenta uma visão geral do pipeline seguido, que compreende: (i) seleção e padronização das bases de dados; (ii) extração e filtragem de emojis; (iii) normalização e cálculo da razão de associação; e (iv) análise comparativa entre plataformas.

Figura 1. Visão geral da metodologia proposta para o estudo.

3.1. Seleção e Padronização dos Dados

A partir de uma revisão da literatura, foram identificados e selecionados repositórios públicos rotulados em português do Brasil, previamente disponibilizados por pesquisadores que investigaram discurso de ódio em plataformas digitais no contexto brasileiro. A Tabela 1 apresenta os seis conjuntos inicialmente considerados.

Tabela 1. Sumário das bases de dados inicialmente consideradas. Linhas em verde indicam as bases selecionadas para a análise final.

Base de Dados	Plataforma	Descrição	Ano	Mensagens
Fortuna [Fortuna et al. 2019]	Twitter/X	Tweets coletados com base em vocabulário de termos associados a discurso de ódio; anotados como ódio ou não ódio.	2019	5.670
OffComBR3 [De Pelle and Moreira 2017]	G1	Comentários do portal G1 abrangendo diversos temas; anotados como ofensivos ou não ofensivos.	2017	1.033
TuPy [de Oliveira et al. 2024]	Twitter/X	Tweets de diferentes domínios temáticos (política, esportes); anotados como discurso de ódio ou não ódio.	2024	10.000
HateBRXplain [Salles et al. 2025]	Instagram	Comentários de contas de seis políticos brasileiros; anotados como ofensivos ou não ofensivos.	2025	7.000
OlidBR [Trajano et al. 2024]	YouTube	Comentários de vídeos sobre temas controversos (política, direitos LGBTQ+); anotados como ofensivos ou não ofensivos.	2024	6.952
TOLD-BR [Leite et al. 2020]	Twitter/X	Tweets com palavras-chave relacionadas a grupos minoritários; anotados como tóxicos ou não tóxicos.	2020	21.000

Três conjuntos foram excluídos da análise por apresentarem baixa ou inexistente ocorrência de emojis, o que inviabilizaria os objetivos deste estudo: o OffComBR3, composto por comentários do portal G1, ambiente textual com uso marginal de emojis; a base de dado Fortuna, cujas mensagens do Twitter/X apresentaram frequência de emojis inferior a 2%; e o TuPy, que, apesar da cobertura temática ampla, também não apresentou emojis associados ao conteúdo textual, inviabilizando uma análise comparativa. Os três conjuntos selecionados para a análise final estão destacados — em verde — na Tabela 1.

Em razão da heterogeneidade dos rótulos originais — que incluíam variações terminológicas como *tóxico*, *ofensivo* e *discurso de ódio* — foi realizada uma padronização das categorias, consolidando os rótulos em uma classificação binária unificada: **discurso de ódio** e **não discurso de ódio**, ou simplesmente (*ódio* / *não ódio*). Essa decisão visa garantir comparabilidade entre os conjuntos nas análises subsequentes, embora implique perda de granularidade semântica.

3.2. Extração de Emojis

A extração de emojis foi realizada em Python, utilizando bibliotecas de processamento textual para identificar e registrar a associação entre cada emoji e o rótulo da mensagem

correspondente. O pipeline foi desenvolvido para processar arquivos nos formatos `.csv`, `.parquet` e `.tsv`, garantindo aplicação uniforme aos diferentes bases de dados e assegurando reprodutibilidade via ambiente Jupyter Notebook[1].

Para cada mensagem, realizou-se a varredura do conteúdo textual com base em um conjunto de caracteres Unicode válidos reconhecidos como emojis, registrando todas as ocorrências e suas respectivas classes. Após essa etapa, os dados foram organizados em uma estrutura tabular consolidada, contendo emoji, plataforma, classe (*ódio* / *não ódio*) e contagem de ocorrências, que serviu de base para as análises comparativas.

3.3. Normalização e Razão de Associação

Como as bases de dados possuem tamanhos distintos e distribuições desbalanceadas entre classes, a comparação baseada em frequências absolutas é inadequada. Para contornar esse problema, as frequências foram normalizadas proporcionalmente ao total de mensagens de cada classe em cada plataforma. A partir dessas proporções normalizadas, calculou-se a razão de associação $R(e)$ para cada emoji e, definida como:

$$R(e) = \frac{\texttt{freq\_ódio\_norm}(e)}{\texttt{freq\_não\_ódio\_norm}(e)} \tag{1}$$

onde $\texttt{freq\_ódio\_norm}(e)$ representa a proporção de mensagens de ódio contendo o emoji e, e $\texttt{freq\_não\_ódio\_norm}(e)$ representa a proporção correspondente nas mensagens não odiosas. Logo, valores $R(e) > 1$ indicam sobrerrepresentação do emoji em conteúdo odioso; valores próximos de 1 indicam distribuição semelhante entre as classes; e valores $R(e) < 1$ indicam maior presença relativa em mensagens não odiosas. Adotamos $R(e) \geq 2$ como limiar para 'fortemente ódio' e $R(e) < 0{,}5$ para 'fortemente não ódio', escolha simétrica pelo inverso — enquanto $R(e) \geq 2$ indica que o emoji aparece ao menos duas vezes mais em ódio do que em não ódio, $R(e) < 0{,}5$ indica o inverso, isto é, ao menos duas vezes mais em não ódio. Essa simetria garante comparabilidade na interpretação dos dois extremos da distribuição. Entre esses extremos, adotamos faixas intermediárias para capturar associações mais sutis, conforme detalhado na Tabela 2.

Tabela 2. Classificação dos emojis por intensidade de associação.

Intervalo de R	Classificação	Interpretação
$R \geq 2{,}0$	Fortemente ódio	Emoji aparece ao menos duas vezes mais em mensagens de ódio
$1{,}2 \leq R < 2{,}0$	Levemente ódio	Maior frequência em ódio, sem forte predominância
$0{,}8 \leq R < 1{,}2$	Neutro	Frequência semelhante entre as classes
$0{,}5 \leq R < 0{,}8$	Levemente não ódio	Maior frequência em não ódio, diferença não acentuada
$R < 0{,}5$	Fortemente não ódio	Emoji aparece ao menos duas vezes mais em mensagens não odiosas

3.4. Análise Comparativa entre Plataformas

Com os dados normalizados consolidados em uma única tabela, foram conduzidas quatro análises complementares, que respondem as nossas questões de pesquisa definidas para

[1]Disponível em: `https://github.com/thulioufv/brasnam2026_emojis`

o estudo: (i) **emojis consistentes**, identificando símbolos com comportamento uniforme entre plataformas ($R > 1$ para ódio ou $R < 1$ para não ódio em ao menos duas das três plataformas) – QP1; (ii) **intensidade de associação**, classificando os emojis conforme os intervalos da Tabela 2 – QP2; (iii) **interseção entre plataformas**, por meio de diagrama de Venn construído a partir dos dez emojis com maior R em cada plataforma – QP3; e (iv) **distribuição por plataforma**, analisando volume e diversidade de emojis em cada ambiente, destacando diferenças estruturais de uso – QP3.

4. Resultados

Esta seção apresenta os resultados obtidos a partir da análise das mensagens rotuladas nas três bases de dados exploradas (apresentadas na Tabela 1). Eles estão organizados de forma a responder progressivamente às questões de pesquisa definidas:primeiro caracterizamos os dados e o uso de emojis por plataforma (QP1); em seguida, analisamos a associação entre emojis e discurso de ódio (QP1 e QP2); e por fim avaliamos a consistência desses padrões entre plataformas (QP3).

4.1. Caracterização das Bases de Dados e Uso de Emojis

A Tabela 3 apresenta uma sumarização das principais métricas descritivas obtidas a partir três bases de dados analisadas neste estudo. Conforme ilustrado na Figura 2, o Twitter/X é o maior conjunto, com 21.000 mensagens, seguido pelo Instagram (7.000) e pelo YouTube (6.952), que possuem um volume aproximado de mensagens. Essa diferença de escala reforça a necessidade do uso de proporções e percentuais nas análises subsequentes, em detrimento de valores absolutos.

Tabela 3. Resumo das mensagens e uso de emojis por plataforma.

Plataforma	Total	Ódio	Não ódio	Com emoji	%	Ocorrências	Únicos
Instagram	7.000	3.500	3.500	1.912	27,31	2.928	263
YouTube	6.952	5.936	1.016	656	9,44	1.045	221
Twitter/X	21.000	9.255	11.745	2.241	10,67	3.876	387

Em relação à distribuição de conteúdo ofensivo, observam-se diferenças relevantes entre as plataformas (Figura 3). A comparação entre os rótulos de cada base, apresentada na Figura 5, evidencia que o Instagram é balanceado (50% ódio / 50% não ódio), enquanto

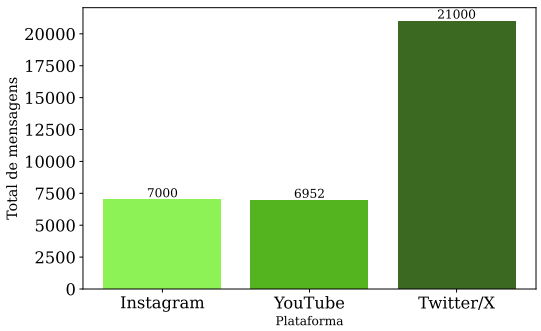

Figura 2. Total de mensagens por plataforma.

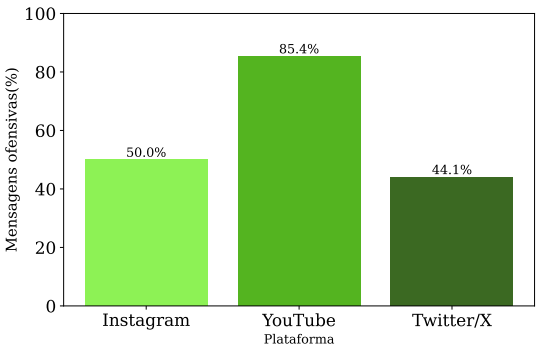

Figura 3. Percentual de mensagens ofensivas por plataforma.

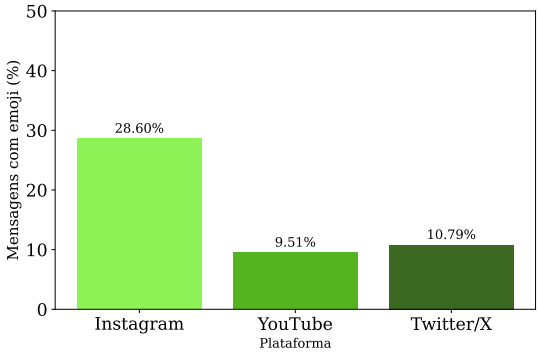

Figura 4. Percentual de mensagens ofensivas por plataforma.

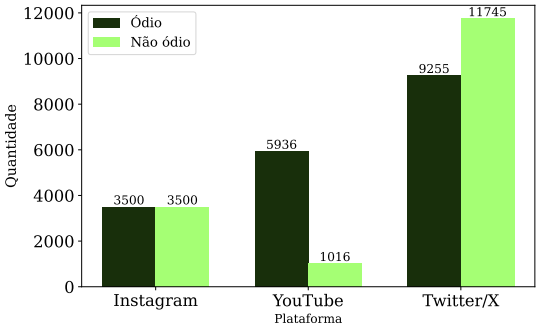

Figura 5. Distribuição dos rótulos de mensagens por plataforma.

o YouTube concentra a maior proporção de mensagens odiosas ($\approx 85\%$), o que reflete o critério mais amplo de rotulação adotado pelo OlidBR, que abrange ofensas em sentido geral e não exclusivamente discurso de ódio estrito. O Twitter/X apresenta proporção intermediária ($\approx 44\%$). Quanto ao uso de emojis, a Figura 4 evidencia que o Instagram se destaca com 27,31% das mensagens contendo ao menos um emoji, valor significativamente superior ao YouTube (9,44%) e ao Twitter/X (10,67%), sugerindo maior integração entre recursos visuais e texto nessa plataforma.

De modo geral, os resultados desta seção evidenciam que as plataformas não diferem apenas em volume e proporção de conteúdo ofensivo, mas também em padrões de uso e diversidade de emojis. Essas diferenças estruturais são fundamentais para uma interpretação adequada das análises subsequentes, especialmente no que se refere ao papel dos emojis na construção discursiva em contextos de linguagem ofensiva.

4.2. Distribuição da Razão de Associação R

As Figuras 6, 7 e 8 apresentam os histogramas da razão R (Equação 1) para cada plataforma. Em todas elas, a maior parte dos emojis concentra-se em valores próximos a 1, indicando que a maioria dos símbolos é utilizada de forma relativamente neutra em relação ao discurso de ódio. No entanto, observam-se diferenças na dispersão entre plataformas: o Instagram apresenta maior variabilidade, com mais emojis em faixas extremas da razão, enquanto o Twitter/X mostra distribuição mais concentrada em torno da neutralidade. Já o YouTube apresenta a maior concentração de discurso de ódio, possivelmente refletindo vieses da base de dados.

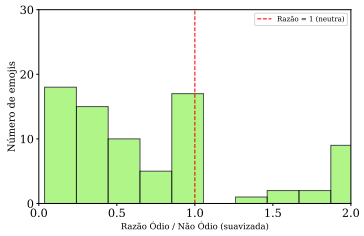

Figura 6. Distribuição da razão ódio/não ódio por plataforma (Instagram).

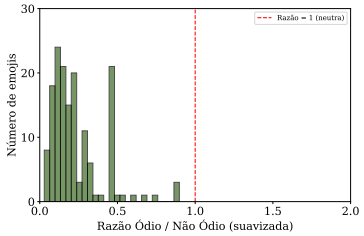

Figura 7. Distribuição da razão ódio/não ódio por plataforma (Twitter/X).

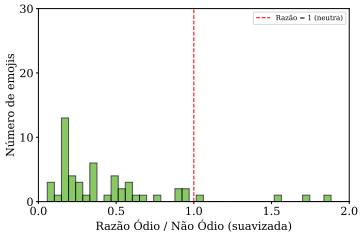

Figura 8. Distribuição da razão ódio/não ódio por plataforma (YouTube).

4.3. Emojis Mais Associados ao Discurso de Ódio

Para responder à QP1, foram selecionados os emojis com maior razão R em cada plataforma, conforme a Tabela 4. No Instagram, diversos emojis apresentam valores elevados de R, indicando forte associação com mensagens de ódio. No YouTube, os dois valores mais extremos — $R = 5222{,}37$ e $R = 2190{,}03$ — merecem atenção especial: com aproximadamente 85% das mensagens dessa base classificadas como ódio, emojis raros na classe não ódio tendem a produzir R artificialmente elevado, de modo que esses valores podem refletir o desbalanceamento estrutural da base tanto quanto uma associação semântica real. Desconsiderando esses casos extremos, os demais emojis do YouTube ainda apresentam $R > 1$, mantendo padrão de associação com conteúdo odioso. No Twitter/X, por sua vez, identificou-se um número consideravelmente menor de emojis com $R > 1$, evidenciando que o uso de emojis nessa plataforma tende a ser mais distribuído entre diferentes tipos de mensagem. Esse resultado sugere que, no Twitter/X, o discurso de ódio pode depender mais de elementos textuais do que de marcadores visuais.

Tabela 4. Top emojis com maior razão de associação ao ódio por plataforma.

Instagram

Emoji	Razão
😈	3142,86
💩	50,82
🤮	13,83
😡	7,00
🤬	6,49
😠	4,60
♂️	4,33
👎	4,33
😒	4,00
🤣	3,50

YouTube

Emoji	Razão
🤮	5222,37
💩	2190,03
🤬	3,25
🤢	2,74
🙄	2,39
🟫	2,05
😠	1,71
😡	1,31
🤣	1,09
😭	1,08

Twitter/X

Emoji	Razão
😌	3,78
😅	1,89
💚	1,68
💜	1,40

O caractere 🟫 atua como um modificador de tom de pele. Na prática, ele não é utilizado isoladamente, mas sim combinado nos bastidores com emojis base (como 👏 ou 🙌) para renderizar a variação étnica escolhida pelo usuário (resultando em 👏🏽 ou 🙌🏽). A aparição do quadrado bege isolado ocorre porque o processamento segmentou a composição sequencial do caractere original.

A Tabela 5 apresenta exemplos extraídos do corpus para os três emojis com maior razão R em cada plataforma (Tabela 4). Os exemplos ilustram como esses símbolos aparecem em mensagens classificadas como discurso de ódio, evidenciando seu papel como marcadores contextuais do conteúdo ofensivo.

4.4. Consistência e Interseção entre Plataformas

Para responder às QP2 e QP3, analisou-se o comportamento de cada emoji nas três plataformas simultaneamente. A Tabela 6 apresenta os emojis com maior número de plataformas em que apresentam associação positiva com ódio ($R > 1$), e a Tabela 7 apresenta os consistentemente associados a conteúdo não odioso ($R < 1$). Os resultados indicam que não há emojis com associação consistente ao ódio nas *três* plataformas simultaneamente.

Tabela 5. Exemplos de uso dos top-3 emojis associados ao ódio por plataforma.

Plataforma	Emoji	Exemplo
Instagram	😈	*"Com o molusco solto, até o dolar não cai!!! Quem confia num país que coloca o maior bandido do mundo na rua???? 😈😈😈😈"*
	💩	*"Que dia é o impeachment desse senhor mesmo? Já deu 💩💩💩💩💩💩💩💩💩💩 💩💩💩"*
	🤮	*"O pior foi um abraço pros Machos e um beijo nas Mulheres, um Escroto....🤮🤮 🤮🤮🤮"*
YouTube	🤮	*"Que nojo desse casal monstruoso 🤮🤮"*
	💩	*"O pior é ver um monte de babaca que defende o COMUNISMO que é pior ou igual ao nazismo e achando normal 💩💩💩💩💩"*
	🤬	*"Deixem nossas crianças em paz, bando de capeta🤬"*
Twitter/X	😒	*"@user porra, uma dessa eu viro, pai 😒😒😒😒😒😒😒😒😒"*
	😅	*"doutor eu não me engano, fdp é corinthiano. eu não sabia mais oq fazer mandei corinthiano e pra casa de fuder 😅"*
	💚	vai tomar no cú , isso aqui é brasil!!! 💚🇧🇷💛

No entanto, alguns símbolos apresentam consistência parcial, com $R > 1$ em duas das três plataformas, sugerindo certa estabilidade de uso ofensivo mesmo em contextos distintos. Por outro lado, alguns emojis apresentam comportamento consistentemente não odioso em todas as plataformas, indicando associação estável com comunicação não agressiva.

Tabela 6. Emojis com associação ao conteúdo odioso em mais de uma plataforma ($R > 1$).

Emoji	Plataformas (Não Ódio)	Plataformas (Ódio)
🤬	1	2
🤮	1	2
🤣	1	2
💩	0	2
😡	1	2
😠	1	2
🙄	1	2
🤢	0	2
♂️	1	1
👎	0	1

Tabela 7. Emojis com associação ao conteúdo não odioso em mais de uma plataforma ($R < 1$).

Emoji	Plataformas (Não Ódio)	Plataformas (Ódio)
🖤	3	0
😍	3	0
👏	3	0
👍	3	0
🙏	3	0
🏻	2	0
💖	2	0
💕	2	0
🏽	2	0
♀️	2	0

A Figura 9 apresenta o diagrama de Venn construído a partir dos dez emojis com maior R em cada plataforma. Observa-se a presença de intersecções parciais entre pares de plataformas, bem como emojis exclusivos de cada ambiente. Em particular, o Twitter/X apresenta maior concentração de emojis exclusivos (seis), ou com associações menos intensas, reforçando a hipótese de que a expressão de ódio nessa plataforma depende menos de emojis. Em contraste, Instagram e YouTube apresentam maior sobreposição e concentração de emojis associados ao ódio, indicando um uso mais consistente desses símbolos nesses ambientes.

4.5. Classificação por Intensidade de Associação

A Tabela 8 sintetiza a distribuição percentual dos emojis por categoria de intensidade em cada plataforma, conforme os intervalos definidos na Tabela 2. O Twitter/X concentra 95,5% dos emojis na categoria 'fortemente não ódio', reflexo da distribuição neutra observada na Seção 4.2. O Instagram apresenta distribuição mais heterogênea, com 41,6%

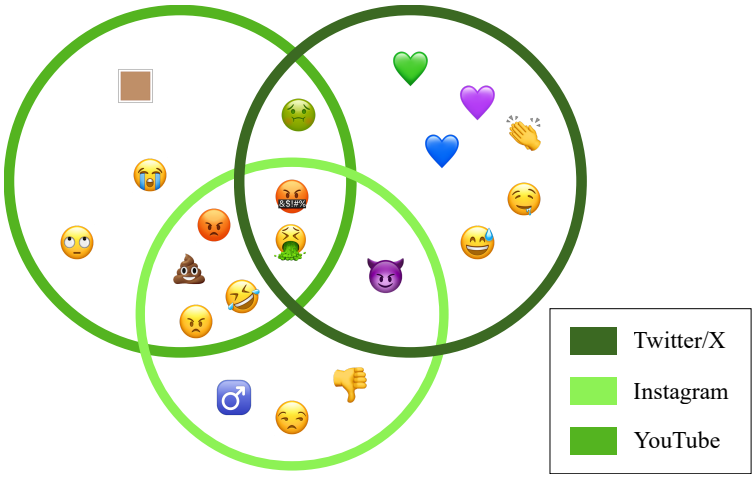

Figura 9. Interseção dos dez emojis com maior razão R entre plataformas.

fortemente não ódio e 29,7% fortemente ódio, confirmando sua maior variabilidade. O YouTube ocupa posição intermediária, indicando que a associação entre emojis e discurso de ódio nessa plataforma é mais equilibrada, refletindo um uso contextual e menos polarizado em comparação às demais.

Tabela 8. Distribuição percentual das classes de associação dos emojis por plataforma.

Plataforma	Fortemente não ódio	Fortemente ódio	Levemente não ódio	Levemente ódio	Neutro
Instagram	41.6%	29.7%	5.9%	5.9%	16.8%
Twitter/X	95.5%	0.0%	2.5%	0.0%	1.9%
YouTube	60.4%	1.9%	22.6%	5.7%	9.4%

5. Conclusão e Trabalhos Futuros

Este estudo analisou a relação entre o uso de emojis e a presença de discurso de ódio em diferentes plataformas sociais utilizadas no Brasil, com foco no Instagram, Twitter/X e YouTube. A partir de uma abordagem empírica baseada na análise de mensagens rotuladas quanto à presença de discurso de ódio, foi possível investigar padrões de associação semântica entre emojis e esse tipo de conteúdo, considerando diferentes contextos de interação digital.

Em relação às questões de pesquisa, os resultados indicam que determinados emojis aparecem com maior frequência em mensagens classificadas como discurso de ódio (QP1), evidenciando padrões de associação que sugerem o uso recorrente desses símbolos como marcadores contextuais desse tipo de conteúdo. Observou-se também que essa relação varia entre plataformas (QP2), indicando que o significado e a função dos emojis não são fixos, mas influenciados pelas dinâmicas de interação e pelas características sociotécnicas de ambientes como Instagram, Twitter/X e YouTube. Por fim, embora não tenham sido identificados emojis universalmente associados ao discurso de ódio, alguns símbolos apresentam consistência parcial entre plataformas (QP3), sugerindo a existência de padrões recorrentes que, ainda que não universais, podem contribuir para a identificação de comportamentos similares em diferentes contextos digitais.

Esses achados indicam que emojis podem atuar como sinais relevantes na caracterização do discurso de ódio, contribuindo para uma compreensão mais ampla e multimodal da comunicação digital. Nesse sentido, a análise evidencia que a interpretação desses elementos não pode ser dissociada do contexto em que ocorrem, reforçando a necessidade de abordagens que considerem não apenas o texto, mas também aspectos simbólicos e paralinguísticos das interações online. Além disso, os resultados apontam para o potencial de incorporação de emojis em modelos automáticos de detecção, possibilitando o desenvolvimento de soluções mais robustas e sensíveis às nuances da linguagem utilizada nas redes sociais.

Embora os conjuntos de dados utilizados apresentem diferenças em termos de coleta, anotação e contexto, o que pode influenciar os padrões observados, os resultados obtidos ainda revelam associações relevantes e consistentes. Nesse sentido, mesmo diante dessas limitações, o estudo contribui para a identificação de padrões interessantes no uso de emojis em contextos de discurso de ódio, reforçando sua relevância para a compreensão da comunicação digital contemporânea.

Como trabalhos futuros, pretende-se investigar como Grandes Modelos de Linguagem (LLMs) e modelos multimodais podem analisar emojis em conjunto com o contexto textual das mensagens, endereçando diretamente as limitações desta análise. Em particular, a presente abordagem não distingue usos irônicos ou sarcásticos de emojis — fenômeno relevante, dado que símbolos aparentemente neutros, como rostos sorridentes, aparecem em mensagens de ódio no corpus analisado. Estudos como o de [Grover and Banati 2024], que exploram a relação entre emojis e sarcasmo via mecanismos de atenção, apontam caminhos promissores para tratar essa ambiguidade. Além disso, a incorporação de emojis como features complementares em classificadores automáticos, combinada a técnicas de aprendizado de máquina que capturem a natureza multimodal da comunicação digital, representa uma direção natural de extensão deste trabalho.

Também se mostra promissora a integração dessas informações em modelos automáticos, explorando o uso de emojis como recursos complementares para aprimorar a detecção em cenários reais, especialmente em abordagens que considerem a natureza multimodal da comunicação digital, além da exploração de técnicas de aprendizado de máquina para a proposição de abordagens automatizadas.

Declaração de uso de Inteligência Artificial

O modelo *Claude Sonnet 4.6 (Anthropic)* foi utilizado apenas para fins de revisão textual. Todo o conteúdo analítico e intelectual deste estudo é de autoria exclusiva dos autores.

Agradecimentos

Este trabalho foi realizado com apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e do INCT em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT TILD-IAR). Os autores também agradecem ao apoio da Universidade Federal de Viçosa (UFV), especialmente ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC).

Referências

Aldunate, N. and González-Ibáñez, R. (2017). An integrated review of emoticons in computer-mediated communication. *Frontiers in psychology*, 7:2061.

Althobaiti, M. J. (2022). Bert-based approach to arabic hate speech and offensive language detection in twitter: exploiting emojis and sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 13(5).

Bertot, J. C., Jaeger, P. T., and Hansen, D. (2012). The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government information quarterly*, 29(1):30–40.

Biere, S., Bhulai, S., and Analytics, M. B. (2018). Hate speech detection using natural language processing techniques. *Master Business AnalyticsDepartment of Mathematics Faculty of Science*.

Braga, M. L. P., Nakamura, F. G., and Nakamura, E. F. (2020). Criação e caracterização de um corpus de discurso sexista em português. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 97–107.

Caetano, J., Guimarães, S., Araújo, M. M., Silva, M., Reis, J. C., Silva, A. P., Benevenuto, F., and Almeida, J. M. (2022). Characterizing early electoral advertisements on twitter: A brazilian case study. In *International Conference on Social Informatics*, pages 257–272.

de Freitas Melo, P., Kansaon, D., Couto, J. M., Reis, J. C., and Benevenuto, F. (2025). A sticker is worth a thousand words: Characterizing the use and abuse of stickers on whatsapp political groups in brazil. In *Proc. of the Int'l AAAI Conference on Web and Social Media*, volume 19, pages 1210–1223.

de Oliveira, F. R., Reis, V. D., and Ebecken, N. F. F. (2024). Detecting hate speech on brazilian social media: New dataset and analysis. In *Ibero-Latin American Congress on Computational Methods in Engineering (CILAMCE)*.

De Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)*, pages 510–519.

de Santana, V. F., Melo-Solarte, D. S., de Almeida Neris, V. P., de Miranda, L. C., and Baranauskas, M. C. C. (2009). Redes sociais online: desafios e possibilidades para o contexto brasileiro. In *Seminário Integrado de Software e Hardware (SEMISH)*, pages 339–353.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Grosz, P. G., Greenberg, G., De Leon, C., and Kaiser, E. (2023). A semantics of face emoji in discourse. *Linguistics and Philosophy*, 46(4):905–957.

Grover, V. and Banati, H. (2024). An attention approach to emoji focused sarcasm detection. *Heliyon*, 10(17):e36398.

Guimarães, S., Silva, M., Caetano, J., Araújo, M., dos Reis, J. C. S., da Silva, A. P. C., Benevenuto, F., and Almeida, J. M. (2022). Análise de propagandas eleitorais antecipadas no twitter. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

Ibrohim, M. O., Setiadi, M. A., and Budi, I. (2019). Identification of hate speech and abusive language on indonesian twitter using the word2vec, part of speech and emoji features. In *Proceedings of the International Conference on Advanced Information Science and System (AISS)*, pages 1–5.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). 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 Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.

Moreira, L. S., Gibrim, P. T. M., Rocha, L., and Reis, J. C. S. (2026). Anatomy of data repositories for the analysis and detection of toxicity in portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR) - Vol. 1*. Association for Computational Lingustics.

Pinto, S. L., Campolina, J. J., Sena, J. P. M., Félix, G., Ferreira, L. N., and Reis, J. C. (2024). Caracterização e predição de usuários tóxicos no twitter/x durante as eleições brasileiras de 2022. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 61–74.

Rodrigues, F. F. (2025). *A morfologia dos emojis: explorando a linguagem visual na comunicação digital em inglês*. Trabalho de Conclusão de Curso. Universidade Estadual do Piauí.

Salles, I., Vargas, F., and Benevenuto, F. (2025). Hatebrxplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in brazilian portuguese. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 6659–6669.

Silva, S. C. and Serapião, A. B. (2018). Detecção de discurso de ódio em português usando cnn combinada a vetores de palavras. In *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, pages 1–8.

Trajano, D., Bordini, R. H., and Vieira, R. (2024). Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.