ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/380534391

The Role of News Source Certification in Shaping Tweet Content: Textual and Dissemination Patterns in Brazil's 2022 Elections

Conference Paper · May 2024

DOI: 10.1145/3658271.3658303

CITATIONS
2

READS
86

3 authors, including:

Carlos Henrique Gomes Ferreira
Federal University of Ouro Preto
57 PUBLICATIONS **415** CITATIONS
SEE PROFILE

Julio Reis
Federal University of Viçosa
78 PUBLICATIONS **1,299** CITATIONS
SEE PROFILE

All content following this page was uploaded by Carlos Henrique Gomes Ferreira on 13 May 2024.

The user has requested enhancement of the downloaded file.

The Role of News Source Certification in Shaping Tweet Content: Textual and Dissemination Patterns in Brazil's 2022 Elections

Luis Guilherme G. da Fonseca
luis.fonseca@aluno.ufop.edu.br
Universidade Federal de Ouro Preto
Brazil

Carlos H. G. Ferreira
chgferreira@ufop.edu.br
Universidade Federal de Ouro Preto
Brazil

Julio C. S. Reis
jreis@ufv.br
Universidade Federal de Viçosa
Brazil

RESUMO

Context: Social media's rise has reshaped information sharing and consumption, particularly in elections, resulting in more toxic content, like hate speech. *Problem*: While there are studies on news content on social media, they overlook the influence of external news sources, especially uncertified sources in the Brazilian context, on the spread of toxic and extremist content on Twitter/X. This gap makes it difficult to fully understand their role in the dissemination of information on social media platforms. *Solution*: This study investigates how different journalistic sources, certified or not by competent bodies, influence the nature and dissemination of tweets related to the Brazilian elections of 2022. In particular, it examines the influence of the toxic and emotional content of headlines on the amplification of tweets by users. *IS Theory*: This research draws on the Social Media Engagement (SME) theory, which emphasizes the role of user interaction and external elements in shaping social media communication and engagement. *Method*: We employ natural language processing techniques to analyze the toxicity and sentiment of tweets and their associated headlines, including quantitative and descriptive analysis of the content of tweets in relation to news sources. *Summarization of Results*: The results show a dominance of negative and toxic tweets, which are significantly shaped by the type of headlines and the certification of sources. Headlines from non-certified sources tend to generate toxic tweets, which indicates a tendency to spread extremist content. *Contributions and Impact in the area of SI*: Our findings shed light on the potential of integrating principles of social media engagement and online news production into information systems to create a better and healthier informative online environment.

KEYWORDS

Social Media, Political Communication, News Source, Twitter (X)

ACM Reference Format:
Luis Guilherme G. da Fonseca, Carlos H. G. Ferreira, and Julio C. S. Reis. 2024. The Role of News Source Certification in Shaping Tweet Content: Textual and Dissemination Patterns in Brazil's 2022 Elections. In *XX Brazilian Symposium on Information Systems (SBSI '24), May 20–23, 2024, Juiz de Fora, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3658271.3658303

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SBSI '24, May 20–23, 2024, Juiz de Fora, Brazil
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0996-8/24/05. . . $15.00
https://doi.org/10.1145/3658271.3658303

1 INTRODUÇÃO

As plataformas de mídias sociais, com sua natureza pervasiva, revolucionaram o modo como a sociedade acessa e compartilha informações, consolidando-se como canais rápidos e eficientes para o acesso e disseminação de informações [2, 19, 26, 33, 34, 40]. No entanto, o aumento do uso de tais plataformas trouxe consigo alguns desafios como a disseminação de desinformação, teorias da conspiração, extremismo e violência, especialmente notórios durante períodos eleitorais, com foco em difundir ideias e promover ataques antidemocráticos [10, 14, 23, 27–30]. Estes conteúdos tóxicos, abrangendo a promoção de violência, discurso de ódio e extremismo ideológico, acarretam danos significativos à sociedade, contribuindo para a polarização e distorcendo o debate político. Exemplos notórios dessas consequências incluem a invasão do congresso americano em 2018 e os ataques aos principais órgãos governamentais brasileiros em 2023, além de tentativas recorrentes de influenciar as eleições em diversos países do mundo [3, 4, 25]. Consequentemente, todos esses eventos têm impulsionado discussões no sentido de regulamentação e responsabilidades dessas plataformas no mundo inteiro e, mais recentemente, no Brasil[1][2].

Dentre as diversas plataformas de mídias sociais disponíveis, destaca-se o Twitter, recentemente redesenhada como sendo X. Essa plataforma amplamente utilizada permite que os usuários consumam e compartilhem conteúdo de interesse de forma rápida e em grande escala. Apesar de contar com estratégias próprias de moderação, isto é, com políticas e diretrizes de uso que visam combater a disseminação dos diversos conteúdos extremistas como, por exemplo, discurso de ódio, violência e assédio, o Twitter tem sido amplamente explorado para deste conteúdo, especialmente, no âmbito político [3, 8, 10, 25, 38, 40]. Esforços anteriores na literatura existente focaram na dinâmica do conteúdo e do debate online no Twitter durante períodos eleitorais [7, 10, 21, 25], bem como sua interação com fontes jornalísticas externas [22, 37, 42, 43, 46].

No entanto, observa-se uma lacuna no que tange a compreensão de como diferentes tipos de fontes externas, em particular, aquelas certificadas e não certificadas, moldam este debate e sua disseminação na plataforma, especialmente, no contexto brasileiro, que ainda é pouco explorado. A certificação, neste contexto, refere-se ao reconhecimento por entidades como a Associação Nacional de Jornais (ANJ)[3], que estabelece códigos de conduta e padrões éticos para a prática jornalística. De fato, fontes externas de notícias têm um papel significativo na moldagem da percepção pública e na dinâmica do debate político [22, 31, 39]. Mesmo assim, tais esforços não as distinguem em categorias específicas para avaliar o estímulo e a

[1]https://www.globaltimes.cn/page/202312/1303925.shtml
[2]https://www.reuters.com/world/americas/brazil-lawmakers-vote-controversial-bill-clean-up-social-media-2023-05-02/
[3]https://www.anj.org.br/

reação dos usuários contribuindo para o debate e a disseminação de informações na plataforma. Por fim, tais análises não investigam esse fenômeno à luz da Teoria de Engajamento em Mídias Sociais [13], que visa investigar a interatividade inerente ao engajamento dos usuários com conteúdos de mídias sociais, reconhecendo que tal engajamento vai além da mera participação. Ele é, de fato, um reflexo de conexões sociais e da percepção de credibilidade. Nessa perspectiva, torna-se viável analisar como fontes de notícias, sejam elas verificadas ou não, e suas características textuais influenciam o engajamento dos usuários e a disseminação de variados conteúdos produzidos por essas fontes.

Visando preencher essa lacuna, este trabalho propõe investigar como o conteúdo de diferentes fontes jornalísticas - sejam elas certificadas ou não - influencia na criação e propagação de *tweets* em torno as eleições brasileiras de 2022. Nosso objetivo central é entender o impacto do teor tóxico e do tom emocional das manchetes de notícias na amplificação dos *tweets*, elucidando o papel dessas fontes na disseminação de informações na plataforma. Nossa pesquisa abrange uma análise detalhada de *tweets* coletados no período eleitoral e nos eventos políticos subsequentes, cobrindo desde a votação e apuração nos dois turnos até os ataque ao Congresso Nacional, Senado e Superior Tribunal Federal em 8 de janeiro de 2023. Para isso, utilizamos técnicas de processamento de linguagem natural para medir a toxicidade e o sentimento tanto dos *tweets* quanto das manchetes jornalísticas dos *links* neles compartilhados. Adicionalmente, exploramos a interação entre as manchetes e os *tweets*, com foco nos *tweets* que incluem *links* para fontes de notícias classificadas quanto à sua certificação. Assim, examinamos como as características textuais das manchetes influenciam no teor dos *tweets* e no engajamento e disseminação por parte dos usuários na plataforma. Nossos principais achados são:

- Em relação a presença de conteúdo tóxico, notamos um forte destaque e presença de *tweets* contendo ataques à identidade e ameaças entre as várias formas de toxicidade analisadas. Este achado revela a presença a disseminação massiva deste tipo de conteúdo durante e após as eleições brasileiras de 2022. A predominância desses tipos específicos de toxicidade sugere uma dinâmica de comunicação voltada para a polarização e o confronto, que é especialmente relevante no contexto político conturbado do período analisado.
- Ao analisar *tweets* com referências para domínios jornalísticos, notamos que *tweets* referenciando fontes não certificadas normalmente atraem mais engajamento. Porém, em casos de conteúdos sensíveis ou radicais, observa-se uma inversão deste padrão, com *tweets* vinculados a fontes certificadas obtendo maior engajamento. Este fenômeno ressalta a importância da credibilidade e da influência social, conforme sugerido pela Teoria de Engajamento em Mídias Sociais, indicando que os usuários exercem cautela ao se associar ou promover notícias de teor mais forte de fontes não certificadas.
- Os padrões textuais da manchete influenciam significativamente o conteúdo dos *tweets* relacionados. Manchetes sem ataques ou ameaças tendem a gerar *tweets* que mantêm esse padrão de não toxicidade. Por outro lado, manchetes com alto teor de toxicidade, criadas especialmente por fontes não certificadas, têm maior probabilidade de gerar *tweets* igualmente tóxicos, indicando uma propensão dessas fontes em contribuírem para a criação e disseminação de conteúdo extremista.
- Sentimento expressos nos *tweets* são predominantemente negativos, com uma proporção ligeiramente menor em *tweets* vinculados a fontes certificadas. Mesmo assim, há uma influência significativa do tom emocional das manchetes na conotação emocional dos *tweets*, novamente a partir de fontes não certificadas.

Em resumo, nossos resultados revelam a influência das manchetes e da certificação das fontes na toxicidade e no sentimento dos *tweets*, além de seu papel na amplificação destes conteúdos na sociedade. Evidenciamos que plataformas como o Twitter/X podem aprimorar suas estratégias de moderação de conteúdo para refinar algoritmos que proativamente identifiquem e gerenciem conteúdos altamente tóxicos, promovendo um ambiente online mais seguro. Dessa forma, desenvolvedores de ferramentas de moderação podem aplicar nossos achados para auxiliar tanto moderadores automatizados quanto humanos a detectar rapidamente noticiais potencialmente problemáticas, especialmente em períodos sensíveis como eleições, caminhando para um espaço digital melhor. Por fim, organizações jornalísticas podem empregar nossas descobertas para ajustar estratégias de publicação, mitigando o impacto negativo de manchetes no debate público e contribuindo para um discurso mais saudável e informado, reforçando a integridade do debate público e a democracia.

O restante do artigo está organizado da seguinte forma. A próxima seção discute os principais esforços das literatura intimamente relacionados aos nossos (Seção 2). A Seção 3 descreve o processo de coleta dos dados. A Seção 4 detalha a metodologia proposta e empregada no trabalho enquanto a Seção 5 apresenta e discute os resultados e principais achados. Por fim, a Seção 6 apresenta as conclusões e as direções futuras.

2 TRABALHOS RELACIONADOS

Diversos estudos na literatura têm focado em compreender características do debate em plataformas de mídias sociais como Facebook [12, 16, 20], Instagram [17, 19] e WhatsApp [29, 35], especialmente durante eventos críticos como eleições e crises globais. De fato, tais plataformas têm sido palco para a disseminação de conteúdo tóxico, incluindo notícias falsas e discursos extremistas. Por exemplo, Zannettou *et al.* exploraram a disseminação de mensagens tóxicas na plataforma Gab [45], enquanto Chandrasekharan *et al.* caracterizaram comunidades banidas do Reddit devido a discursos de ódio [5]. No trabalho de Silva *et al.*, os autores discutem a propagação de conteúdo inapropriado frente à liberdade de expressão em plataformas de mídias sociais [11].

Especificamente sobre o Twitter, estudos como o de Rosa *et al.* e Saldanha *et al.* [10, 36] abordaram a presença e evolução do discurso radical, com foco nas eleições americanas de 2018 e o ataque ao congresso americano. Já Chowdhury *et al.* compararam os *tweets* de usuários suspensos com um grupo de controle, descobrindo que os primeiros eram mais propensos a violar várias regras do Twitter, como compartilhamento de insultos e discurso de ódio [7]. Zannettou estudaram os padrões de compartilhamento associados a *tweets* que receberam um alerta político do Twitter, com foco

no envolvimento de usuários suspensos [44]. Já Grimminger *et al.* analisaram a presença de discurso de ódio entre diferentes grupos de apoiadores de candidatos [21].

Focando na interação dos usuários com notícias externas no Twitter, Vosoughi *et al.* destacaram como as desinformações oriundas de fontes externas se espalham mais rapidamente do que verdades factuais na plataforma [42]. Wischnewski *et al.* examinaram como notícias hiper-partidárias de fontes externas são compartilhadas em redes sociais, ressaltando a influência dessas características no engajamento dos usuários [43]. Salehabadi *et al.* investigaram a relação entre a toxicidade de notícias externas e o engajamento em conversas tóxicas no Twitter [37]. Por fim, Zhang *et al.* analisaram o papel de grupos extremistas como o QAnon na ampliação de sua influência na plataforma utilizando notícias externas para aumentar o engajamento e a visibilidade [46]. Estes estudos ressaltam a importância de entender como notícias de fontes externas moldam a interação e o engajamento dos usuários no Twitter, especialmente, em um contexto não explorado, que é o brasileiro.

Em contraste aos estudos anteriores, nosso trabalho propõe analisar como padrões textuais de notícias de diferentes fontes impactam no engajamento e na disseminação de *tweets*. Em particular, nosso trabalho busca elucidar como as características textuais de notícias de fontes diversas, tanto certificadas quanto não certificadas, influenciaram a dinâmica de engajamento e disseminação de *tweets* durante as eleições brasileiras de 2022. Adicionalmente, nosso estudo adota uma perspectiva fundamentada na Teoria de Engajamento em Mídias Sociais para explorar como as interações entre usuários e conteúdos, moldadas pela credibilidade de fontes jornalísticas, se manifestam nos padrões de engajamento dentro da plataforma. Esta perspectiva teórica, até então não explorada em estudos anteriores sobre o tema, nos permite uma compreensão mais profunda de como a natureza e a credibilidade das fontes de notícias afetam a propagação de conteúdo e, consequentemente, a formação de opinião pública, especialmente em um contexto politicamente sensível como as eleições brasileiras.

3 COLETA E CARACTERIZAÇÃO DOS DADOS

A base de dados explorada neste estudo é composta por *tweets* coletados da plataforma Twitter[4], recentemente redesenhada como X[5]. A coleta foi realizada através de uma Interface de Programação de Aplicação (do inglês, *Application Programming Interface* - API) voltada para pesquisa acadêmica[6], cujo acesso foi aprovado e disponibilizado pela própria plataforma. Em seguida, conduzimos a coleta entre 1º de outubro de 2022 e 10 de fevereiro de 2023, período que abrangeu eventos chave das eleições gerais brasileiras de 2022, como os dois turnos de votação, apuração dos resultados, pesquisas eleitorais, debates e o ataque aos prédios da Câmara dos Deputados, Senado e do Superior Tribunal Federal em 8 de janeiro de 2023. Durante este período, já se observava a propagação de conteúdos antidemocráticos no Twitter[7]. Visando identificar conteúdos políticos radicais e relacionados a atos anti-democráticos, exploramos o seguinte conjunto de palavras-chave: *atos antidemocráticos, intervenção militar, manifestações Brasília, atos pela democracia, invade Brasília, invasão Brasília, ocupa brasília, ocupa STF, ocupa congresso, ocupa palácio alvorada, terroristas brasília, ataque Brasília, invade STF, invade congresso, invade câmara dos deputados, ocupa câmara dos deputados*. A princípio, nossa base de dados incluía um total de 2.959.732 mensagens, abrangendo duplicatas de *tweets*, *retweets* (Compartilhamentos), *quotes* (Citações) e *replies* (Respostas) com identificadores únicos fornecidos pelo Twitter. Assim, utilizamos identificadores únicos da mensagem fornecidos pela própria API, juntamente com a data e hora da coleta de cada mensagem, para eliminar duplicações, mantendo, no entanto, as mensagens mais recentemente coletadas. Esta estratégia permite capturar as mensagens com índices de engajamento mais atualizados, essenciais para avaliar a interação entre os usuários e o conteúdo postado. Tais índices incluem:

- # *Likes* (Curtidas): Reflete o número de usuários que curtiram o tweet, indicando concordância ou apreciação pelo conteúdo;
- # Retweets (Compartilhamento): Mostra quantas vezes um *tweet* foi compartilhado, ampliando seu alcance;
- # *Replies* (Respostas): Contabiliza as respostas ou comentários recebidos, refletindo o nível de engajamento direto dos usuários com o tweet;
- # *Quotes* (Citação): Registra quantas vezes um *tweet* foi compartilhado com comentários adicionais, revelando a interação e perspectiva dos usuários;
- # *Impressions* (Impressões): Indica o número de vezes que um *tweet* apareceu na linha do tempo dos usuários, medindo sua visibilidade.

Depois disso, foi necessário tratar a natureza peculiar de textos escritos em português e disponibilizados em plataformas sociais que naturalmente impõe uma série de desafios ao lidar, por exemplo, com mensagens curtas para várias tarefas de processamento de linguagem natural (PLN), que é a base para o nosso trabalho e será discutida na próxima seção. Ademais, baseado em estudos anteriores [17, 27], que avaliam o impacto do tamanho das mensagens em tarefas correlatas às que serão realizadas aqui, removemos *tweets* cujo tamanho estava abaixo de 30 caracteres. Por fim, removemos mensagens que não estavam em português usando uma API do Google[8] que retorna o idioma de um texto. Após essas etapas de filtragem, chegamos a um conjunto final de 733.219 *tweets* únicos.

Uma vez que estamos interessados em avaliar a interação desse conteúdo com as notícias externas, consideramos também a relevância das manchetes dos links compartilhados nas plataformas de mídias sociais. Estudos anteriores ressaltam a importância deste elemento no processo de interação e engajamento dos usuários com este conteúdo [32]. Dessa forma, coletamos também todas as manchetes associadas aos links presentes no conjunto final de *tweets*.

A Tabela 1 apresenta a média, desvio padrão, valor máximo e mediana das métricas de engajamento do conjunto de *tweets* finais. A partir do número de *Impressions* (impressões), que representa a número de vezes com que um *tweet* aparece na linha do tempo dos usuários, observa-se uma média de aproximadamente 798 usuários com um desvio padrão significativamente alto de $38.351,21$. Isso indica que, em média, os *tweets* coletados têm uma exposição considerável ao atingir quase 800 usuários. No entanto, o desvio

[4]https://twitter.com
[5]www.x.com
[6]https://developer.twitter.com/en/use-cases/do-research/academic-research
[7]https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=508935&tip=UN,https://www.bbc.com/portuguese/brasil-63990040
[8]https://perspectiveapi.com/

Tabela 1: Descrição do conjunto de *tweets* finais.

Categoria	Média	Desvio Padrão	Máximo	Mediana
Impressions	798,55	38351.21	16760867	0
Likes	50.45	1306.34	419142	0
Retweets	8.80	208.29	53280	0
Quotes	1.05	63.03	23372	0
Replies	3.38	147.81	64329	0

padrão e o valor máximo de 16.760.867 destacam que alguns *tweets* alcançaram uma exposição excepcionalmente alta. Já a mediana de 0 sugere que a maior parte dos *tweets* teve uma exposição nula. Focando nas demais métricas (*# Likes*, *# Retweets*, *# Quotes* e *# Replies*), observa-se que, apesar da exposição significativa indicada pelo número de *Impressions*, a interação média dos usuários com os *tweets* é relativamente baixa, embora o desvio padrão e o valor máximo encontrado em cada métrica mostrem uma enorme variação. Por fim, a mediana igual a zero mostra que grande parte dos *tweets* não tiveram nenhuma interação.

4 METODOLOGIA

Esta seção apresenta a metodologia proposta para realização deste estudo. Na Seção 4.1, detalhamos o processo de classificação das fontes de notícias, diferenciando-as entre certificadas e não certificadas. Em seguida, na Seção 4.2, descrevemos a metodologia empregada na extração e análise das propriedades textuais dos *tweets* analisados.

4.1 Processo de Classificação da Fonte

Relembre que o foco deste trabalho é avaliar o papel do Twitter como canal de disseminação de padrões textuais, particularmente, com um enfoque na amplificação de notícias de fontes externas. Neste contexto, nossa estratégia incluiu inicialmente a extração de todos os localizadores uniforme de recursos (do inglês, *Uniform Resource Locator* - URL) do conjunto final de *tweets*, utilizando expressões regulares. Nós também consideramos os encurtados, fazendo requisições e coletando a URL original para assegurar consistência[9]. Posteriormente, focamos no domínio principal de cada URL, removendo subdomínios e extensões. Por fim, identificamos e selecionamos os 100 domínios mais compartilhados com base na frequência de compartilhamento.

Para determinar se um domínio é uma fonte de notícia, adotamos a autodeclaração feita pelo próprio domínio em sua página no Facebook[10]. Esta abordagem, previamente utilizada em outros estudos, permite uma identificação eficiente de um domínio como fonte de notícias [22, 41]. Dos 100 domínios mais compartilhados, 70 foram reconhecidos como fontes de notícias por terem se autodeclaradas como sendo, por exemplo, empresa de mídia/notícias, domínio de notícias e mídia, personalidade de notícias, empresa de produção de mídia e transmissão, personalidade de notícias e jornal.

A próxima etapa do nosso estudo envolve a classificação das fontes de notícias em `certificadas` ou `não certificadas`. Definimos que uma fonte é considerada `certificada` se estiver relacionada e/ou associada à Associação Nacional de Jornais[11] (ANJ), que estabelece padrões de responsabilidade e ética jornalística para empresas brasileiras. Esta estratégia tem sido explorada em esforços anteriores [9]. Observamos, no entanto, a presença significativa de fontes internacionais como CNN[12], Reuters[13] e BBC[14], que têm se popularizado no Brasil, conforme indicado pelo Google Trends[15]. Logo, para essas fontes internacionais, utilizamos o *Media Bias/Fact Check*[16], uma plataforma amplamente empregada em estudos anteriores e reconhecida por sua avaliação da confiabilidade de domínios de notícias, levando em conta aspectos como viés político e precisão factual [6, 31]. Fontes vinculadas à ANJ ou classificadas com alta credibilidade pelo *Media Bias/Fact Check* são categorizadas como `certificadas`. As fontes `não certificadas` compreendem as demais dentre as 70 analisadas. Em resumo, 13 (≈18%) fontes foram classificadas como `certificadas` e 57 como `não certificadas`, contabilizando ao todo 40054 *tweets* com menções a essas mídias. A Tabela 2 apresenta a lista dos domínios em cada uma das categorias.

Finalmente, adotamos um grupo de controle, denominado `outras`, que refere-se aos demais *tweets* não classificados com base nos critérios anteriormente adotados, contendo ou não fontes de notícias, e que são representativos do discurso sobre o tema de estudo na plataforma. O objetivo principal deste grupo é oferecer um ponto de comparação para as propriedades textuais e os padrões de engajamento dos *tweets* que incluem *links* de fontes de notícias `certificadas` e `não certificadas`, contrastando-os com outros *tweets* compartilhados na plataforma relacionados ao nosso estudo. Na prática, os *tweets* categorizados como `outras` fontes representam uma forma de discurso mais livre no Twitter, estando sujeitos unicamente à moderação da própria plataforma. Este grupo não está sujeito às restrições de conteúdo inerentes às fontes de notícias, sejam elas as `certificadas` ou as `não certificadas`, que possuem suas próprias políticas internas de gerenciamento de conteúdo.

4.2 Análise de Propriedades Textuais

Para analisar as propriedades textuais das notícias compartilhadas no Twitter, iniciamos com a investigação da presença de toxicidade no conteúdo. Esta análise é conduzida utilizando o modelo *Perspective*[17], que foi treinado com milhões de comentários de variadas fontes, incluindo Wikipédia e comentários em domínios de notícias diversos, e abrange múltiplos idiomas, incluindo o português. O modelo avalia o texto e determina a probabilidade de ele ser percebido como tóxico por usuários da Web sob diferentes dimensões. Estudos anteriores sugerem a adoção de um limiar de 0,8 para considerar um conteúdo efetivamente tóxico, indicando que tal conteúdo seria percebido como tóxico por 8 em cada 10 usuários [15, 24]. O *Perspective* avalia as seguintes dimensões de toxicidade[18]:

- Toxicidade (*Toxicity*): Avalia a probabilidade de um comentário ser percebido como rude, desrespeitoso ou irracional, de

[9] Para realização das requisições foi utilizada a biblioteca *requests*, disponível em: https://pypi.org/project/requests/.
[10] www.facebook.com
[11] https://www.anj.org.br/
[12] https://www.cnn.com/
[13] https://www.reuters.com/
[14] https://www.bbc.com/
[15] https://trends.google.com.br/trends/explore?date=2020-10-18%202023-11-18&geo=BR&q=BBC,CNN,Reuters&hl=pt
[16] https://mediabiasfactcheck.com/
[17] https://perspectiveapi.com/
[18] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

Tabela 2: Lista de domínios certificados e não certificados.

Certificadas	correiodopovo, estadao, folha, reuters, cnnbrasil, bbc.com, globo, opovo, intercept, otempo, jornaldebrasilia, gazetadopovo, uol
Não Certificadas	cartacapital, abril, brasildefato, antenapoliticabr, conexaopolitica, vistapatria, pleno, caldeiraopolitico, aosfatos, jornaldacidadeonline, r7, plantaobrasil, revistaforum, horabrasilia, nsctotal, atarde, saibamais, folhadestra, metropoles, apublica, istoe, diariodocentrodomundo, terrabrasilnoticias, terra, poder360, agoranoticiasbrasil, tvi.iol, contrafatos, gazetabrasil, jovempan, clicrbs, msn, sbtnews, sapo, wordpress, brasil247, aliadosbrasiloficial, dw.com, portaltocanews, conjur, diariodopoder, revistaoeste, noticiasaominuto, atrombetanews, brasilsemmedo, dunapress, apostagem, ebc, correiobraziliense, jornalggn, sputniknewsbrasil, change, exame, yahoo, redebrasilatual, expresso

uma maneira que possa fazer as pessoas abandonarem uma discussão;
- Ameaça (*Threat*): Avalia se um comentário contém ameaças, incluindo a intenção de infligir dor, lesão ou violência contra um indivíduo ou grupo;
- Ataque à Identidade (*Identity Attack*): Comentários negativos ou de ódio direcionados a alguém por causa de sua identidade;
- Toxicidade severa (*Severe Toxicity*): Semelhante à toxicidade, mas foca em comentários muito odiosos, agressivos e desrespeitosos ou que de outra forma provavelmente fará um usuário abandonar uma discussão ou desistir de compartilhar sua perspectiva. Este atributo é muito menos sensível a formas mais leves de toxicidade, como comentários que incluem usos positivos de palavrões;
- Insulto (*Insult*): Comentário insultuoso, inflamado ou negativo em relação a uma pessoa ou grupo de pessoas;
- Obscenidade (*Profanity*): Detecta o uso de palavras obscenas, palavrões ou outra linguagem obscena ou profana.

Adicionalmente, usamos a análise de sentimento visando mensurar a tonalidade do discurso. Entre as várias ferramentas de análise de sentimento existentes na literatura, exploramos o VADER [1, 18], uma ferramenta de análise de sentimento disponível para português que possui bom desempenho por ser construída especificamente a partir de dados plataformas de mídias sociais. Em suma, dado um texto de entrada, o VADER computa quatro variáveis para um texto: *positiva*, *negativa*, *neutra* e *composta*. A variável *composta*, normalizada entre -1 (extremamente negativo) e 1 (extremamente positivo), reflete o sentimento predominante no texto. Baseando-nos nas recomendações dos autores, classificamos o sentimento de um *tweet* da seguinte maneira:

- **Negativo:** Quando o escore composto é menor ou igual a -0,05;
- **Neutro:** Quando o escore composto fica no intervalo entre -0,05 e 0,05;
- **Positivo:** Quando o escore composto é maior ou igual a 0,05.

Por fim, nossa análise estende-se às manchetes tanto de mídias certificadas quanto não certificadas, examinando sua toxicidade e sentimento. As manchetes, frequentemente o primeiro contato que um usuário tem com o conteúdo de um link externo no Twitter, podem ter um impacto significativo na percepção e na decisão do usuário de interagir com o *tweet*. Ao serem exibidas junto aos *tweets*, as manchetes funcionam como complemento do conteúdo, moldando a expectativa e potencialmente influenciando a reação dos usuários à aquele *tweet*.

A partir da categorização proposta para as fontes de notícias e das propriedades textuais analisadas, exploramos também as métricas de engajamento descritas na Seção 3 para avaliar como o tipo da fonte e as propriedades textuais dos *tweets* e das manchetes impactam na amplificação deste tipo de conteúdo no Twitter. O foco é explorar como o tipo da fonte de notícias e as propriedades textuais dos *tweets* e das manchetes impactam na amplificação destes conteúdos no Twitter, à luz da Teoria de Engajamento em Mídias Sociais. Especificamente, propomos avaliar como o engajamento do usuário nas plataformas de mídia social é profundamente influenciado pela interação entre o conteúdo consumido e as dinâmicas sociais subjacentes, como a credibilidade da fonte e a ressonância emocional do conteúdo. Ao empregar métricas de engajamento — como impressões, curtidas, *retweets*, citações e respostas — nossa abordagem metodológica capta o engajamento ativo dos usuários com o conteúdo relacionado às eleições brasileiras de 2022, refletindo diretamente os conceitos fundamentais da teoria. As dimensões de toxicidade e sentimentos analisadas, em conjunto com essas métricas de engajamento, permitem-nos examinar como o conteúdo tóxico e emocionalmente carregado, ancorado pela veracidade e natureza das fontes de notícias, engaja os usuários e promove a disseminação de informações na plataforma.

5 RESULTADOS

Esta seção apresenta e discute os resultados do nosso estudo. Na Seção 5.1, focamos na análise da presença de toxicidade nos *tweets* e como diferentes tipos de fontes de notícias influenciam essa toxicidade. Em seguida, na Seção 5.2, aprofundamos na investigação dos sentimentos expressos nos *tweets*, examinando-os à luz das fontes jornalísticas.

5.1 Análise de Toxicidade

Nossa análise inicial foca na presença geral de toxicidade nos *tweets*. A Figura 1 exibe a função de distribuição acumulada (CDF) da toxicidade para cada dimensão avaliada pelo modelo *Perspective*, conforme discutido na Seção 4.2. Podemos observar uma prevalência significativa de *tweets* tóxicos, especialmente naqueles com probabilidade acima de 0.8, nas dimensões de ataques à identidade (*identity attack*) e ameaças (*threat*). Em particular, aproximadamente 20% dos *tweets* são fortemente identificados como contendo ataques ou

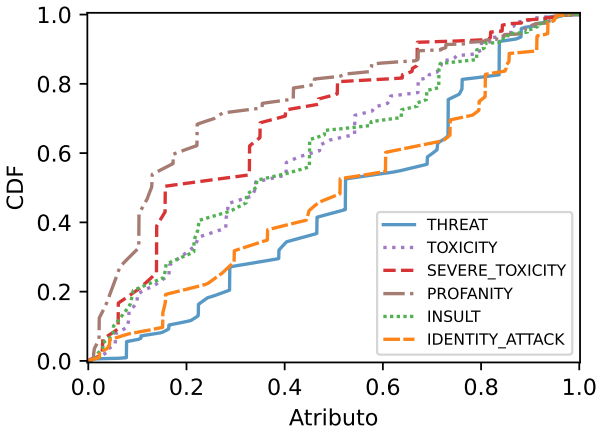


Figura 1: Distribuição dos atributos de toxicidade nos *tweets*.

ameaças. Para as outras dimensões de toxicidade, a proporção de *tweets* tóxicos varia entre 9% e 11%[19].

Diante da presença e relevância de conteúdo de ataque à identidade e ameaça em nossa análise, propusemos uma avaliação focada em capturar *tweets* que representam conteúdos tóxicos e radicais no contexto das eleições brasileiras de 2022. Para isso, propomos um escore combinado, calculado como a média desses dois atributos.

Novamente, um *tweet* é categorizado como tóxico e radical se esse escore médio exceder o limiar de 0,8. Esta abordagem nos permite identificar *tweets* que simultaneamente contêm ataques à identidade e ameaças, características intimamente relacionadas a este estudo, e que são facilmente percebidos por outros usuários. A Tabela 3 exemplifica *tweets* da nossa base de dados que alcançaram as maiores médias nesse escore, evidenciando o tipo de discurso que está sendo analisado.

Tabela 3: Exemplos de top *tweets* com maior presença de ataque à identidade e ameaça.

Texto	Escore
Quando o exército toma conta e fazer a intervenção militar e acabar de vez com o petismo comunista tudo voltará ao normal porque o lula e o maior terrorista criminoso que vive na sociedade brasileira	0.930
Vamos parar o Brasil por pessoas morrendo, crianças por esse terroristas do lula e sua turma da #esquerda	0.908
Queria jogar uma bomba nos bolsominions q tao manifestando hoje pedindo a p***da intervenção militar e explodir todos eles	0.994
Todo nazista e todos que defendem manifestações nazistas merecem morrer queimados lentamente. TODOS, NAO SALVA UM	0.989

Depois, a próxima etapa consistiu em entender os padrões de engajamento em *tweets* associados a conteúdos de mídias `certificadas`, `não certificadas` e `outras`, especialmente no que tange a presença de ataque à identidade e ameaça (i.e., métricas que compõem o escore combinado proposto). Em relação aos 22.123 *tweets* fazendo referência a conteúdos de mídias `certificadas`, apenas cerca de 5% (1.134) são classificados como contendo altos índices de ataque à identidade e ameaça (escore => 0.8). Similarmente, em um total de 17.931 *tweets* referenciando mídias `não certificadas`, aproximadamente 6% (1.098) fazem alusão a este tipo de conteúdo. Por outro lado, aqueles conteúdos provenientes de `Outras` mídias, que representam a maior parte da nossa base de dados com 693.165 *tweets*, cerca de 12% (87.754) fazem referência este tipo de conteúdo.

A Tabela 4 detalha as métricas de engajamento por categoria de *tweets*. Observamos padrões distintos que destacam as diferenças entre *tweets* contendo links para diferentes fontes de informação e fazendo menção ou não a conteúdos de ataque à identidade e ameaça. Na categoria de *tweets* contendo links para fontes e com menos probabilidade de serem percebidos como contendo ataques à identidade e ameaça `certificadas < 0.8`, notamos uma média de *impressions* (*impressions*) de 2131,68, com um máximo de 773.840 e uma mediana de 15, acompanhada de desvios padrões elevados nas métricas de engajamento, como *likes* (69,47), *retweets* (11,21), *quotes* (1,88) e *replies* (10,00). Isso indica que, apesar de alguns desses *tweets* alcançarem um grande número de usuários, muitos outros têm um alcance limitado.

Para os *tweets* contendo links para fontes `certificadas => 0.8`, observamos uma média de *impressions* menor, 748,09, indicando um alcance médio geralmente reduzido para *tweets* com alto teor de toxicidade. As médias de *likes* (38,38), *retweets* (8,00), *quotes* (0,55) e *replies* (1,58) são também inferiores às observadas na categoria anterior, refletindo um engajamento mais moderado.

Já os *tweets* com links para fontes `não certificadas < 0.8` exibem uma média de *impressions* superior (2589,17) e um desvio padrão ainda maior (25499,55), apontando para uma maior variabilidade e alcance. Além disso, as médias de likes (139,79), *retweets* (29,04) e replies (14,83) são relativamente mais altas do que aquelas observadas em *tweets* com links para fontes certificadas, implicando em um engajamento mais substancial dos usuários. Por outro lado, os *tweets* com links para fontes `não certificadas => 0.8` mostram uma média de *impressions* de 562,17, refletindo um alcance menor em comparação com os *tweets* menos tóxicos da mesma categoria. As médias de likes (33,07), *retweets* (7,37), quotes (0,57) e replies (1,75) também são menores do que nos *tweets* `não certificados < 0.8`.

A categoria `outras`, que representa *tweets* sem vínculos diretos com fontes de notícias certificadas ou não certificadas, revela um padrão distinto. Esses *tweets* têm médias mais baixas de engajamento como o número de *impressions* (782,51 para `< 0.8` e 331,59 para `=> 0.8`), likes (51,87 para `< 0.8` e 25,76 para `=> 0.8`) e *retweets* (9,07 para `< 0.8` e 4,04 para `=> 0.8`). Isso pode indicar que, apesar de representarem uma expressão mais autêntica seja ela mais ou menos moderada do debate, esses *tweets* não geram o mesmo nível de interesse que aqueles com links para fontes de notícias externas, resultando em menor engajamento.

Em geral, os resultados sugerem que quando um *tweet* recebe um número elevado de interações, como *likes*, *retweets*, *replies* ou *quotes*, ele tende a aparecer mais na linha do tempo dos usuários, um processo natural de disseminação de informação. Potencialmente, esta dinâmica pode criar um ciclo em que um maior engajamento culmine em mais *impressions*, o que por sua vez pode gerar ainda

[19]Esta distribuição foi analisada levando em conta as três categorias de fontes de informação — Certificada, Não Certificada e Outras — e revelou padrões semelhantes em todas elas.

Tabela 4: Métricas de Engajamento por Categoria de *tweets*

Categoria	Métrica	Média	Desvio Padrão	Máximo	Mediana
Certificados < 0.8	Impressions	2131.68	16133.18	773840	15
	Likes	69.47	622.33	39969	1
	Retweets	11.21	114.55	6995	0
	Quotes	1.88	20.98	1242	0
	Replies	10.0	110.79	4931	0
Certificados => 0.8	Impressions	748,09	10027.03	266639	11
	Likes	38.38	631.51	19077	0
	Retweets	8.00	148.35	4564	0
	Quotes	0.55	8.1	245	0
	Replies	1.58	19.45	561	0
Não Certificados < 0.8	Impressions	2589.17	25499.55	1259992	8
	Likes	139.79	1066.32	42951	1
	Retweets	29.04	242.64	9915	0
	Quotes	3.17	32.64	1981	0
	Replies	14.83	171.75	10116	0
Não Certificados => 0.8	Impressions	562.17	5994.97	134625	7
	Likes	33.07	287.93	6376	0
	Retweets	7.37	73.64	2107	0
	Quotes	0.57	5.49	132	0
	Replies	1.75	13.67	314	0
Outros < 0.8	Impressions	782.51	41275.02	16760867	0
	Likes	51.87	1381.82	419142	0
	Retweets	9.07	218.54	53280	0
	Quotes	1.08	68.81	23372	0
	Replies	3.23	159.33	64329	0
Outros => 0.8	Impressions	331.59	14191.38	3090076	0
	Likes	25.76	891.38	191030	0
	Retweets	4.04	136.14	27121	0
	Quotes	0.43	24.53	4481	0
	Replies	1.44	43.86	7205	0

mais engajamento. Nesse sentido, *tweets* contendo links para notícias de fontes não certificadas têm sido preferidos pelos usuários no Twitter e mais propícios a caírem neste ciclo, mas quando se trata de conteúdos mais sensíveis e radicais, como os analisados aqui, os *tweets* com links para fontes certificadas se sobressaem. Isso pode refletir uma hesitação dos usuários em se associar ou amplificar conteúdos controversos ou prejudiciais provenientes de fontes menos confiáveis.

Para investigar a influência do teor radical de uma manchete no conteúdo do(s) *tweet*(s) que a compartilham, analisamos a probabilidade de um *tweet* exibir ataques ou ameaças condicionada ao caráter da manchete[20]. Em termos práticos, calculamos a fração de *tweets* com escore de toxicidade menor do que 0.8 e maior ou igual a 0,8 dado que a manchete relacionada apresenta um escore também igual a 0,8, e vice-versa. As Figuras 2(a) e 2(b) ilustram essa relação condicional para manchetes oriundas de fontes certificadas e não certificadas, respectivamente.

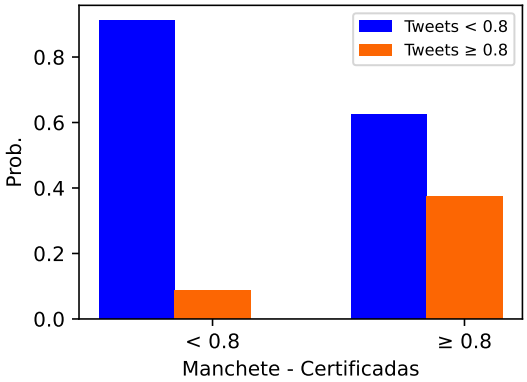
Independentemente da certificação da fonte, os resultados mostram que manchetes com um escore de ameaça e ataque abaixo de 0,8 têm uma alta probabilidade ($\approx 0{,}9$) de gerar *tweets* que refletem um teor semelhante de não toxicidade. Em outras palavras, manchetes percebidas como menos agressivas ou ameaçadoras tendem a resultar em *tweets* que mantêm essa característica, independentemente da origem da notícia.

No entanto, a situação muda significativamente quando consideramos manchetes com conteúdo mais extremista (escore => 0,8). Neste contexto, as manchetes de fontes não certificadas apresentam uma tendência maior de influenciar mais a criação de *tweets* que espelham essa mesma natureza extremista. Em contraste, as manchetes de fontes certificadas demonstram um efeito muito menor nesse sentido. Este padrão sugere que fontes não certificadas podem estar mais propensas a promover uma associação com conteúdos extremistas, especialmente aqueles caracterizados por ataques e ameaças, como identificado em nossa análise.

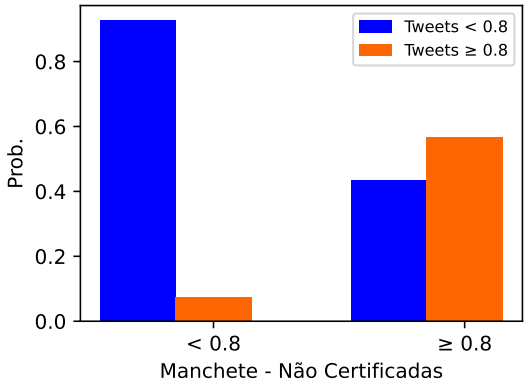
5.2 Análise de Sentimento

A Figura 3 apresenta a distribuição dos sentimentos dos *tweets* categorizados como `Certificadas`, `Não Certificadas` e `Outras`. De modo geral, nota-se uma tendência: *tweets* negativos são predominantes em todas as categorias, com variação entre entre 50% e 57%. Especificamente, a proporção de *tweets* negativos é um pouco

[20]Como explicado na Seção 4, focamos em apenas em *tweets* com pelo menos trinta caracteres.



(a) Certificada



(b) Não certificada

Figura 2: Análise do impacto da presença de ameaça e ataques na manchete na geração dos *tweets*.

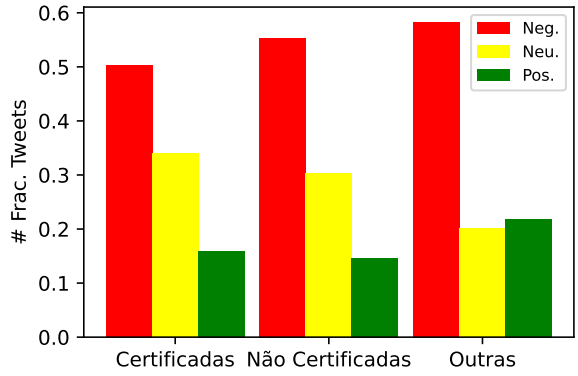


Figura 3: Distribuição dos sentimentos do texto dos *tweets* por categoria.

menor para as fontes `Certificadas` em comparação com as `Não Certificadas`, que apresentam uma proporção quase equiparável à categoria `Outras`. Isso pode indicar que os conteúdos vinculados a fontes não certificadas tendem a suscitar uma resposta emocional mais negativa.

Por outro lado, a categoria `Outras`, que não possui vínculos diretos com as fontes de notícias certificadas ou não certificadas mais populares da base analisada, possui a menor fração de sentimentos neutros comparada as outras classes bem como a maior fração de sentimentos negativos. Potencialmente, esses *tweets* abordam uma gama mais ampla de tópicos e não restritos ao contexto jornalístico que podem ocasionar em respostas emocionais mais variadas e intensas.

Da mesma forma que abordamos a toxicidade condicional, conduzimos uma investigação sobre como o tom emocional das manchetes influencia os sentimentos dos *tweets* associados, utilizando a probabilidade condicionada como nossa métrica de análise. As Figuras 4(a) e 4(b) mostram as distribuições dos sentimentos nos *tweets* em relação ao sentimento das manchetes, distinguindo entre as provenientes de fontes certificadas e não certificadas, respectivamente.

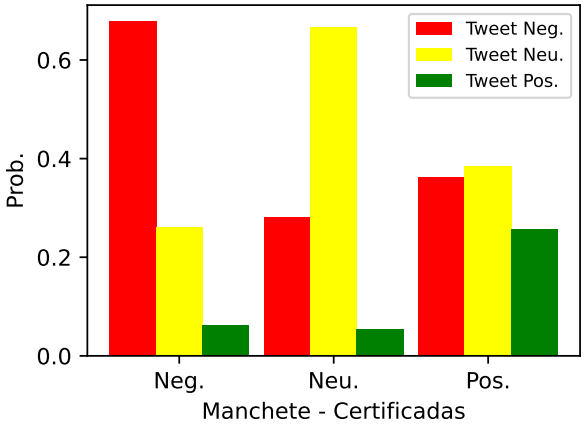
Para as manchetes de fontes certificadas (Figura 4(a)), observa-se que manchetes com conotação negativa tendem a gerar *tweets* predominantemente negativos. Este resultado corrobora com estudos anteriores [32]. Além disso, este padrão se mantém com manchetes neutras, que resultam majoritariamente em *tweets* neutros. Contudo, manchetes positivas desviam dessa tendência, levando a uma maior proporção de *tweets* negativos e neutros, em vez de positivos.

Ao analisar as manchetes de fontes não certificadas (Figura 4(b)), percebemos um padrão similar em que manchetes negativas e neutras geram *tweets* com sentimentos correspondentes. No entanto, diferentemente das fontes certificadas, manchetes positivas de fontes não certificadas tendem a resultar em *tweets* positivos. Isso indica que as manchetes de mídias não certificadas influenciam de forma mais consistente o sentimento dos *tweets* vinculados a elas. Estes resultados destacam a complexidade das reações emocionais no Twitter, particularmente em relação à natureza das fontes de notícias e sua influência no sentimento dos *tweets*.

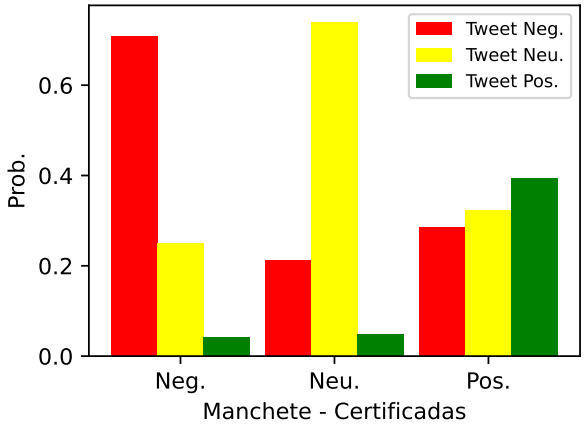
Em suma, os resultados destacam a importância da credibilidade da fonte e do impacto emocional do conteúdo no engajamento dos usuários. Esta observação corrobora com os princípios da Teoria de Engajamento em Mídias Sociais, sugerindo que a credibilidade da fonte, juntamente com a natureza do conteúdo, são fatores críticos que influenciam a decisão dos usuários de se engajar com *tweets* específicos, promovendo a amplificação seletiva de informações baseada na percepção de veracidade e relevância.

6 CONCLUSÃO E TRABALHOS FUTUROS

Este estudo investigou os padrões de interação entre usuários e conteúdos no Twitter em torno das eleições brasileiras de 2022, enfocando a influência de fontes jornalísticas certificadas e não certificadas na disseminação de *tweets* sob a ótica da Teoria de Engajamento em Mídias Sociais. Observamos não apenas uma prevalência preocupante de conteúdo tóxico, caracterizado por ataques à identidade e ameaças, mas também uma dinâmica de engajamento que reflete a credibilidade das fontes e a ressonância emocional do conteúdo.



(a) Certificada



(b) Não certificada

Figura 4: Análise do impacto do sentimento da manchete na geração dos *tweets*.

Nossos achados evidenciam um cenário de radicalismo e polarização, com um engajamento diferenciado em conteúdos sensíveis, onde as fontes certificadas são mais engajadas, e uma tendência das fontes não certificadas em promover mensagens extremistas.

À luz da Teoria de Engajamento em Mídias Sociais, nossos resultados mostram a importância da credibilidade da fonte e da natureza emocional do conteúdo na modulação do engajamento dos usuários. Isso ressalta o papel das plataformas de mídia social e das organizações jornalísticas em fomentar um ambiente de discussão mais saudável e menos extremista. A implementação de políticas de moderação de conteúdo mais eficazes e o desenvolvimento de ferramentas avançadas de monitoramento e análise podem contribuir significativamente para este objetivo. Neste sentido, as implicações práticas deste estudo são vastas, sugerindo que plataformas de mídia social adotem estratégias de moderação de conteúdo informadas pelas dinâmicas de engajamento identificadas. Além disso, organizações jornalísticas podem adotar mecanismos de monitoramento ativo das reações do público às notícias, ajustando seus padrões de comunicação para minimizar a disseminação de conteúdo polarizador e maximizar a contribuição para um debate público construtivo.

As implicações deste estudo abrem caminho para uma série de pesquisas futuras. Inicialmente, merece atenção a investigação das estratégias de certificação de fontes jornalísticas e seu efeito na propagação de informações. A criação de sistemas de informação avançados para monitorar as dinâmicas de engajamento nas redes sociais em tempo real representa um campo promissor. Além disso, a flexibilidade da nossa metodologia para se adaptar a diferentes contextos e a possibilidade de examinar as propriedades textuais pertinentes em cada caso revelam um potencial significativo para ampliar este estudo a outras áreas. Destaca-se também a versatilidade de nossa abordagem em diversos cenários e a adaptabilidade dos modelos de processamento de linguagem natural empregados. Assim, nosso estudo não apenas elucida o impacto das redes sociais nas dinâmicas políticas atuais, mas também estabelece uma base aplicável a variadas discussões online. Desta forma, sublinhamos a relevância de reconhecer a heterogeneidade do engajamento nas mídias sociais e seu efeito na formação da opinião pública. Esse entendimento, alinhado à Teoria de Engajamento em Mídias Sociais, reforça a importância de pesquisas contínuas voltadas ao aprimoramento de um espaço digital mais esclarecido e democrático.

AGRADECIMENTOS

Este estudo recebeu suporte financeiro e recursos de infraestrutura da Universidade Federal de Ouro Preto (UFOP), através do Programa de Bolsas de Iniciação Científica e Tecnológica da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). Além disso, contou com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERÊNCIAS

[1] Rafael J. A. Almeida. 2018. LeIA - Léxico para Inferência Adaptada. https://github.com/rafjaa/LeIA.

[2] Marcelo MR Araujo, Carlos HG Ferreira, Julio CS Reis, Ana PC Silva, and Jussara M Almeida. 2023. Identificação e Caracterização de Campanhas de Propagandas Eleitorais Antecipadas Brasileiras no Twitter. In *Anais do XII Brazilian Workshop on Social Network Analysis and Mining*. SBC, 67–78.

[3] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.

[4] Geisa Tamara Bugs, Agnes Silva de Araujo, Diego Saez-Trumper, and Rodrigo Firmino. 2023. Mapping Political Extremism on Twitter in Brazil. In *International Conference on Computational Science and Its Applications*. Springer, 439–454.

[5] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[6] Mingxuan Chen, Xinqiao Chu, and KP Subbalakshmi. 2021. MMCoVaR: multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[7] Farhan Asif Chowdhury, Dheeman Saha, Md Rashidul Hasan, Koustuv Saha, and Abdullah Mueen. 2022. Examining Factors Associated with Twitter Account Suspension Following the 2020 U.S. Presidential Election. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '21)*.

[8] Paulo Henrique Ribeiro Costa and Luciano Heitor Gallegos Marin. 2021. Sistema para coleta e tratamento textos brasileiros sobre polarização política. In *Anais Estendidos do XVII Simpósio Brasileiro de Sistemas de Informação*. SBC, 01–04.

[9] Joao MM Couto, Julio CS Reis, Ítalo Cunha, Leandro Araújo, and Fabrício Benevenuto. 2022. Characterizing low credibility websites in brazil through computer networking attributes. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 42–46.

[10] Jose Martins da Rosa Jr, Renan Saldanha Linhares, Carlos Henrique Gomes Ferreira, Gabriel P Nobre, Fabricio Murai, and Jussara M Almeida. 2022. Uncovering discussion groups on claims of election fraud from twitter. In *International Conference on Social Informatics*. Springer, 320–336.

[11] Gabriela Nunes Pinto Da Silva, Thiago Henrique Costa Silva, and João Da Cruz Gonçalves Neto. 2021. Liberdade de expressão e seus limites: uma análise dos discursos de ódio na era das fake news. *Revista Argumenta* 34 (2021), 415–437.

[12] Clarissa C David, Ma Rosel S San Pascual, and Ma Eliza S Torres. 2019. Reliance on Facebook for news and its influence on political engagement. *PloS one* 14, 3 (2019), e0212263.

[13] Paul M Di Gangi and Molly M Wasko. 2016. Social media engagement theory: Exploring the influence of user engagement on social media usage. *Journal of Organizational and End User Computing (JOEUC)* 28, 2 (2016), 53–73.

[14] Régis Ebeling, Jéferson Nobre, and Karin Becker. 2023. A multi-dimensional framework to analyze group behavior based on political polarization. *Expert Systems with Applications* 233 (2023), 120768.

[15] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[16] Juliana Fernandes, Magda Giurcanu, Kevin W Bowers, and Jeffrey C Neely. 2010. The writing on the wall: A content analysis of college students' Facebook groups for the 2008 presidential election. *Mass communication and society* 13, 5 (2010), 653–675.

[17] Carlos HG Ferreira, Fabricio Murai, Ana PC Silva, Jussara M Almeida, Martino Trevisan, Luca Vassio, Marco Mellia, and Idilio Drago. 2021. On the dynamics of political discussions on instagram: A network perspective. *Online Social Networks and Media* 25 (2021), 100155.

[18] CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

[19] Carlos Henrique Gomes Ferreira, Fabricio Murai, Ana Paula Couto da Silva, Jussara Marques de Almeida, Martino Trevisan, Luca Vassio, Idilio Drago, and Marco Mellia. 2020. Unveiling community dynamics on instagram political network. In *Proceedings of the 12th ACM Conference on Web Science*. 231–240.

[20] Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M Guess, et al. 2023. Asymmetric ideological segregation in exposure to political news on Facebook. *Science* 381, 6656 (2023), 392–398.

[21] Lara Grimminger and Roman Klinger. 2021. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

[22] Samuel S Guimarães, Julio CS Reis, Lucas Lima, Filipe N Ribeiro, Marisa Vasconcelos, Jisun An, Haewoon Kwak, and Fabrício Benevenuto. 2020. Identifying and characterizing alternative news media on facebook. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[23] Jonathas GD Harb and Karin Becker. 2018. Emotion analysis of reaction to terrorism on twitter. In *Anais do XXXIII Simpósio Brasileiro de Banco de Dados*. SBC, 97–108.

[24] 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.

[25] Renan S Linhares, José M Rosa, Carlos HG Ferreira, Fabricio Murai, Gabriel Nobre, and Jussara Almeida. 2022. Uncovering coordinated communities on twitter during the 2020 us election. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 80–87.

[26] Larissa Malagoli, Júlia Stancioli, Carlos HG Ferreira, Marisa Vasconcelos, Ana Paula Couto da Silva, and Jussara Almeida. 2021. Caracterizaçao do debate no twitter sobre a vacinaçao contra a covid-19 no brasil. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*. SBC, 55–66.

[27] Larissa G Malagoli, Julia Stancioli, Carlos HG Ferreira, Marisa Vasconcelos, Ana Paula Couto da Silva, and Jussara M Almeida. 2021. A look into covid-19 vaccination debate on twitter. In *Proceedings of the 13th ACM Web Science Conference 2021*. 225–233.

[28] Gabriel P Nobre, Jussara M Almeida, and Carlos HG Ferreira. 2019. Caracterização de bots no Twitter durante as Eleições Presidenciais no Brasil em 2018. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*. SBC, 107–118.

[29] Gabriel Peres Nobre, Carlos HG Ferreira, and Jussara M Almeida. 2022. A hierarchical network-oriented analysis of user participation in misinformation spread on WhatsApp. *Information Processing & Management* 59, 1 (2022), 102757.

[30] Gabriel Peres Nobre, Carlos Henrique Gomes Ferreira, and Jussara Marques Almeida. 2020. Beyond groups: Uncovering dynamic communities on the whatsapp network of information dissemination. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*. Springer, 252–266.

[31] Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. 2019. A link-based approach to detect media bias in news websites. In *Companion Proceedings of The 2019 World Wide Web Conference*. 742–745.

[32] Julio Reis, Fabrício de Souza, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *international AAAI conference on web and social media*.

[33] Julio CS Reis, Philipe Melo, Fabiano Belém, Fabricio Murai, Jussara M Almeida, and Fabricio Benevenuto. 2023. Helping Fact-Checkers Identify Fake News Stories Shared through Images on WhatsApp. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*. 159–167.

[34] Julio CS Reis, Philipe Melo, Márcio Silva, and Fabrício Benevenuto. 2023. Desinformação em plataformas digitais: Conceitos, abordagens tecnológicas e desafios. *Sociedade Brasileira de Computação* (2023).

[35] Gustavo Resende, Philipe Melo, Julio CS Reis, Marisa Vasconcelos, Jussara M Almeida, and Fabrício Benevenuto. 2019. Analyzing textual (mis) information shared in WhatsApp groups. In *Proceedings of the 10th ACM conference on web science*. 225–234.

[36] Renan S Saldanha, José M Rosa, CH Ferreira, Gabriel Nobre, Fabricio Murai, and Jussara Almeida. 2022. Uncovering coordinated communities on twitter during the 2020 us election. *Prof. of the ASONAM* (2022).

[37] Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. 2022. User Engagement and the Toxicity of Tweets. *arXiv preprint arXiv:2211.03856* (2022).

[38] Francis Spiegel Rubin, Yuri Luz de Almeida, Adriana Cesario de Faria Alvim, Vânia Félix Dias, and Rodrigo Pereira dos Santos. 2021. Analysis of the First Round of 2018 Government Election for the State of Rio de Janeiro Based on Twitter. In *XVII Brazilian Symposium on Information Systems*.

[39] Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *International AAAI Conference on Web and Social Media*.

[40] Io Taxidou and Peter M Fischer. 2014. Online analysis of information diffusion in twitter. In *Proceedings of the 23rd international conference on world wide web*. 1313–1318.

[41] Marcelo Träsel, Sílvia Lisboa, and Giulia Reis Vinciprova. 2019. Post-truth and trust in journalism: an analysis of credibility indicators in Brazilian venues. *Brazilian journalism research* 15, 3 (2019), 452.

[42] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[43] Magdalena Wischnewski, Axel Bruns, and Tobias Keller. 2021. Shareworthiness and motivated reasoning in hyper-partisan news sharing behavior on Twitter. *Digital Journalism* 9, 5 (2021), 549–570.

[44] Savvas Zannettou. 2021. I Won the Election!: An Empirical Analysis of Soft Moderation Interventions on Twitter.. In *International AAAI Conference on Web and Social Media*. 865–876.

[45] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*. 1007–1014.

[46] Yini Zhang, Zhiying Yue, Xiyu Yang, Fan Chen, and Nojin Kwak. 2022. How a peripheral ideology becomes mainstream: Strategic performance, audience reaction, and news media amplification in the case of QAnon Twitter accounts. *New Media & Society* (2022).

View publication stats