Toxicidade e Gatilhos: Um Estudo de Caso em Comunidades do Reddit no Brasil

Giovana Piorino
Universidade Federal de Minas Gerais
giovana.piorino@dcc.ufmg.br

Luiz Henrique Quevedo Lima
BEON.tech
luiz.quevedo@beon.tech

Adriana Silvina Pagano
Universidade Federal de Minas Gerais
apagano@ufmg.br

Ana Paula Couto da Silva
Universidade Federal de Minas Gerais
ana.coutosilva@dcc.ufmg.br

ABSTRACT

In this work, we aim to understand linguistic features that may indicate potential triggers fostering toxic behavior among users in Portuguese-speaking Reddit communities. Our findings show that discussions involving such triggers tend to concentrate political terms and insults, shift topics more frequently, and make greater use of subordinating conjunctions.

KEYWORDS

Toxicity Triggers, Discussion Trees, Linguistic Analysis

1 INTRODUÇÃO

Nos últimos anos, as plataformas de redes sociais se tornaram um importante meio de interação entre pessoas de diferentes lugares do mundo, tanto para troca de informações quanto para busca de suporte emocional [22]. Dentre elas, o *Reddit*[1] se destaca tanto pelo seu papel positivo quanto também como um ambiente onde estas interações podem se tornar tóxicas e prejudiciais aos seus usuários, por meio da discussão e busca por posicionamentos [1, 5, 18].

O comportamento tóxico de um ou mais usuários pode ocorrer em qualquer nível de interação da árvore de discussões ou *threads*. Alguns trabalhos na literatura buscam compreender as razões que podem levar a interações que, à principio, são saudáveis a se tornarem tóxicas [2, 3, 23] . Para tal, estes trabalhos introduzem o conceito de gatilhos, definidos como comentários que, mesmo sendo classificados como *Não Tóxicos* (por meio de anotação manual, por exemplo), apresentam respostas classificadas como *Tóxicas* em sua estrutura de árvores de discussão. Dessa forma, são interpretados como comentários com potencial de provocar respostas tóxicas, sendo identificados como pontos de relevância na estrutura das árvores de discussão.

Neste contexto, este trabalho reúne e expande estudos prévios para comparar árvores de discussão do *Reddit*, em português do Brasil, que contêm comentários gatilho, com aquelas livres desse tipo de interação. Para isso, analisamos tanto características linguísticas quanto métricas estruturais de grupos de árvores, classificados conforme a proporção de respostas tóxicas. O objetivo é identificar padrões que facilitem a detecção e a caracterização desses comentários gatilho em contraposição às conversas sem comportamento tóxico.

[1]https://www.reddit.com/

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

ISSN 2966-2753

2 TRABALHOS RELACIONADOS

Alguns trabalhos na literatura buscam entender como interações não tóxicas podem atuar como gatilhos para o comportamento tóxicos dos usuários na plataforma Reddit [2, 3, 6, 19].

Os autores em [2] investigam os possíveis gatilhos que desencadeiam comportamentos tóxicos. Para isso, define gatilho como um comentário não tóxico com respostas tóxicas (mesma definição usada no nosso trabalho). As árvores de discussão analisadas foram rotuladas por meio de um classificador LSTM (*Long Short-Term Memory*) e representação vetorial do texto usando o modelo GloVe, com F1-Score de 0,83. O estudo conclui que termos de conotação política e mudanças abruptas de temas entre comentários da árvore são fortes indícios de caracterização de gatilhos. O surgimento de toxicidade está mais ligado a mudanças locais de tópico e sentimento entre comentários adjacentes (por exemplo, a presença de termos políticos) do que a métricas globais da árvore.

De forma similar, o estudo em [3] apresenta o modelo PROVOKE, construído a partir de uma rede neural baseada em LSTM bidirecional, que integra representações de texto e informações contextuais para detectar gatilhos de toxicidade. O modelo foi treinado nas árvores de discussão do *Reddit* coletadas e alcançou F1-Score de aproximadamente 0,78. Para o conjunto de árvores analisadas no estudo, aponta-se que árvores com presença de gatilhos apresentam mais termos ligados a humor sarcástico ou política, enquanto árvores sem gatilhos concentram temas relacionados à tecnologia.

Em [19] o modelo *ToxiGen* foi utilizado para classificar um espectro de toxicidade para cada comentário de cada árvore, e organizar métricas relacionadas às árvores de discussão. O estudo conclui que a toxicidade de cada comentário é predominantemente determinada pelo comentário pai imediato, seguindo os resultados apresentados em [6]. Neste estudo, o modelo *Detoxify* foi utilizado para rotular a toxicidade das árvores. Observou-se que mais da metade das conversas mais tóxicas terminam com comentários tóxicos e são, em geral, mais longas e profundas.

Enquanto estudos prévios em inglês focaram em como comentários inofensivos podem desencadear toxicidade, nossa trabalho se distingue por analisar os aspectos linguísticos que sinalizam potenciais gatilhos em interações no Reddit em português do Brasil.

3 METODOLOGIA

A seguir descrevemos a metodologia utilizada neste trabalho.

3.1 Conjunto de Dados

Os dados utilizados neste trabalho foram previamente apresentados em [10]. O conjunto de dados é formado por postagens e comentários feitos entre janeiro e dezembro de 2022 nos dez subreddits brasileiros mais populares e foi coletado via API Pushshift.[2] Os dados analisados incluem comentários apenas em português, sendo excluídos comentários de comunidades que permitem discussões em várias línguas. Aproximadamente 600 mil comentários categorizados como deletados ou removidos foram filtrados, juntamente com comentários contendo apenas emojis ou símbolos, URLs, caracteres não alfanuméricos e reações de texto apenas de risada.[3] Por fim, também excluímos comentários gerados por contas de automoderadores e bots que detectamos em nossos dados. Assim, nosso corpus, após a aplicação dos filtros possui aproximadamente 6,6 milhões de comentários.

3.2 Anotação Manual

O processo de anotação manual de toxicidade de uma amostra do corpus foi um dos resultados apresentados em [10]. Resumidamente, uma amostragem estratificada de 2.500 comentários foi selecionada. Os comentários foram divididos em cinco lotes de 500 comentários cada. Doze estudantes de graduação e pós-graduação em Computação e Letras foram divididos em 4 grupos. Cada comentário foi classificado em uma das categorias: *Tóxico*, *Não tóxico*, *Não sei* ou *Informação insuficiente*, seguindo definições estabelecidas pela Perspective API.[4]

Como resultado final do processo de anotação, 95,72% dos comentários apresentaram concordância parcial, ou seja, ao menos dois anotadores atribuíram o mesmo rótulo à eles. Considerando esse subconjunto, 85,62% das anotações apresentaram rótulo *Não tóxico*, 11,28% para o rótulo *Tóxico*, 1,92% para *Não sei* e 1,17% para *Informação insuficiente*. Neste trabalho, consideramos apenas os comentários cuja concordância entre pelo menos dois anotadores foi atribuída aos rótulos *Não Tóxico* ou *Tóxico* (voto majoritário), resultando em um total de 2.319 comentários de interesse.

3.3 Gatilhos em Árvores de Discussão

Uma forma de analisar o conteúdo gerado através das interações dos usuários no Reddit iniciadas pelas *threads* (em termos de postagens e comentários) é a construção de árvores de discussão. Consideramos como árvore de discussão, toda a troca de informação que se inicia por um comentário em resposta à uma postagem. Ele é seguido de comentários (respostas) feitos pelo próprio usuário que gerou o comentário raiz ou por outros usuários na comunidade.

Para identificar a existência de gatilhos nas discussões analisadas, usamos a definição encontrada na literatura para um comentário ser considerado um *gatilho* [2, 3]. Mais precisamente, um *gatilho* é todo comentário que foi manualmente rotulado como *Não tóxico* e que gerou respostas (comentários) com conteúdo tóxico. Desta forma, as árvores de discussão analisadas podem ser divididas em dois grupos: árvores *tóxicas*, que possuem gatilhos, e árvores *não tóxicas*, onde comentários rotulados manualmente como *Não tóxicos* não possuem respostas tóxicas, ou seja, árvores sem *gatilhos*. Importante ressaltar que, para árvores tóxicas, uma resposta tóxica não necessariamente precisa ocorrer logo após o gatilho.

A partir dos 2.049 comentários anotados manualmente como *Não tóxicos*, reconstruímos as árvores de discussão associadas. Cada comentário rotulado manualmente está associado à uma árvore distinta, sendo que 1.194 desses comentários não possuem respostas associadas e, portanto, não seguem a definição de *gatilho* e suas árvores de discussão foram desconsideradas da nossa análise. Assim, temos 855 árvores a serem analisadas, com o total de 4.016 comentários associados. Nosso objetivo é identificar traços implícitos de linguagem que possam contribuir para mudanças de rumos nas discussões nas comunidades analisadas e identificar proativamente quando uma discussão possui o potencial de se tornar tóxica.

> Você é um assistente que classifica comentários do Reddit em Português do Brasil (PT-BR) como *Tóxico* ou *Não Tóxico*. Você receberá o texto de um comentário e a sua tarefa é classificar toxicidade do texto fornecido.
>
> Use somente as informações abaixo para fazer a predição:
>
> (1) Para cada comentario se limite a escolher apenas uma dessas duas opções, sem acrescentar texto explicativo e sem marcar outras opções que não sejam uma dessas duas; *Tóxico* ou *Não Tóxico*;
> (2) Marque somente como *Tóxico* os comentários que tiver certeza, alta confiança de que possam ser considerados tóxicos;
> (3) Marque somente como *Não Tóxico* os comentários que tiver certeza, alta confiança de que não sejam considerados tóxicos;
>
> Para cada comentário abaixo marque uma das opções: *Tóxico* ou *Não Tóxico*.

Quadro 1: Prompt utilizado para a rotulação automática de comentários por meio do GPT-4.1

3.4 Detecção Automática de Comentários Tóxicos

Para dividirmos as discussões entre árvores de discussão *tóxicas* e *não tóxicas* temos que classificar os demais 4.016 comentários que fazem parte das 885 árvores de interesse. Para a análise inicial apresentada neste trabalho, utilizamos o modelo GPT-4.1[5] para classificar um comentário como *Tóxico* ou *Não tóxico*. O Quadro 1 apresenta o *prompt* utilizado para a tarefa de classificação. Sua estrutura é pautada em instruções mais diretas e sucintas [4], e evita fornecer exemplos específicos associados a cada rótulo [14].

3.5 Caracterização das Árvores Tóxicas e Não Tóxicas

Visando encontrar possíveis diferenças entre árvores de discussão com (tóxicas) e sem gatilhos (não tóxicas), realizamos o conjunto de

[2] https://github.com/pushshift/api
[3] Em português, textos de risadas são representados pela sequência de caracteres kkkkk
[4] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US
[5] https://openai.com/index/gpt-4-1/

análises linguísticas descritas a seguir. Compreender as principais diferenças e similaridades entre discussões com e sem gatilhos é importante para aplicar mecanismos que possam garantir um ambiente mais saudável nas plataformas de redes sociais online identificando a propagação do comportamento tóxico proativamente.

3.5.1 ***Nuvens de palavras.*** Para obter uma visão geral das discussões realizadas nas árvores tóxicas e não tóxicas, utilizamos nuvens de palavras [12]. Essa abordagem possibilita identificar os termos mais frequentes nos conjuntos de comentários analisados.

3.5.2 ***Classificação de tópicos (BERTopic).*** Geramos *embeddings* para todos os comentários, tanto de árvores tóxicas quanto não tóxicas, projetando-os em um mesmo espaço vetorial. O modelo utilizado para essa extração é o *Legal-BERTimbau-sts-base-ma*[6], uma versão derivada do BERTimbau [20], com ênfase em similaridade semântica [7, 11, 17]. Em seguida, aplicamos redução de dimensionalidade de vetores via UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) e clusterização com o algoritmo HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) com auxílio da biblioteca BERTopic[9], de modo que cada comentário foi automaticamente atribuído a um dos tópicos identificados pelo modelo. Seguimos recomendações da biblioteca para a escolha de parâmetros.[7] O parâmetro de número de vizinhos utilizado para UMAP foi 10 e o número de componentes foi 8. Para o HDBSCAN, o número mínimo do tamanho de agrupamentos utilizado foi 15 e número mínimo de amostras foi 10.

Para analisar as mudanças de tema nas discussões, atribuímos um tópico dominante a cada comentário usando a similaridade de termos com a técnica c-TF-IDF [9]. Em seguida, contamos as transições de tópico entre comentários consecutivos (valores totais e normalizados) em cada árvore, sumarizando os resultados com médias e medianas.

3.5.3 ***Etiquetagem de classe de palavra (Pos Tagging).*** Para identificar as classes gramaticais predominantes em cada tipo de árvore de discussão, aplicamos POS tagging[15] com o modelo[8] disponível na biblioteca spaCy [21], baseado em uma treebank [16] e anotada segundo o padrão Universal Dependencies [8]. Entre as possíveis funções do modelo, utilizamos especificamente o componente morfologizador, responsável pela atribuição de categorias gramaticais.

3.5.4 ***Reconhecimento de Entidades Nomeadas (REN).*** Nessa análise, aplicamos o componente REN, modelo apresentado na biblioteca spaCy, adaptando-o aos nossos dados por meio do componente de POS tagging junto ao corpus WikiNER [13]. Essa abordagem classifica as entidades em quatro categorias: Pessoa (PER), Localização (LOC), Organização (ORG) e Diversos (MISC).

4 RESULTADOS

A seguir apresentamos e discutimos os nossos resultados.

[6]https://huggingface.co/rufimelo/Legal-BERTimbau-sts-base-ma-v2
[7]https://maartengr.github.io/BERTopic/index.html
[8]pt_core_news_lg

4.1 Identificação de Gatilhos

Para avaliar a eficácia do GPT-4.1 como rotulador automático de toxicidade, aplicamos o modelo e seu prompt ao conjunto de comentários rotulados por humanos para extrair métricas de desempenho. A média macro de F1-score é de 0,79, e a média ponderada é 0,91. O modelo apresenta desempenho elevado na classe *Não Tóxico*, com precisão e revocação em torno de 0,95. Para a classe *Tóxico*, no entanto, essas métricas caem para aproximadamente 0,62. Esse resultado reflete o desbalanceamento do conjunto de dados, mas indica que o modelo ainda mantém uma capacidade razoável de identificar toxicidade, sem comprometer significativamente a precisão global.

Das 855 árvores que tiveram os comentários classificados automaticamente, em 703 árvores não há comentários tóxicos subsequentes ao comentário rotulado manualmente como *Não Tóxico*, ou seja, não houve gatilho, sendo, portanto, consideradas árvores não tóxicas. Por outro lado, 152 árvores continham ao menos uma resposta tóxica subsequente ao potencial gatilho, consideradas árvores tóxicas. As árvores não tóxicas concentram-se em poucos comentários: aproximadamente ≈ 76% das árvores desse conjunto encerram-se com até cinco comentários, sendo que apenas dois a três comentários já respondem por ≈ 46% de todos os casos. A partir da 6.ª resposta a participação cai rapidamente, tornando árvores profundas menos frequentes. Nas árvores tóxicas, a distribuição é mais uniforme. Os picos ainda ocorrem em dois (≈ 10%) e três (≈ 16%) comentários, mas a queda é mais suave; cerca de metade dessas árvores ultrapassa cinco respostas. Assim, árvores tóxicas apresentam mais frequentemente interações com muitos comentários. Em suma, árvores com gatilhos tendem a atrair mais comentários.

É interessante observar que, nos casos em que o comentário analisado não atua como gatilho, o número médio de respostas subsequentes em sua subárvore é de 1,99, com uma mediana de 1 comentário. Por outro lado, quando o comentário funciona como gatilho, a média de comentários subsequentes aumenta para 3,47, com uma mediana de 3 respostas.Esses valores reforçam que, na presença de um gatilho, as conversas tendem a ser mais longas e a se estender por mais interações antes de se encerrarem.

4.2 Caracterização das Discussões Tóxicas e Não Tóxicas

Nas análises a seguir, consideramos as árvores com mais de dois comentários com o objetivo de analisar discussões com maior nível de interação entre os usuários. Assim, o nosso conjunto final de análise é formado por 543 árvores não tóxicas com 2.713 comentários associados e 136 árvores tóxicas, com 951 comentários associados.

Nuvens de palavras: Elementos mais frequentes em árvores tóxicas incluem termos e figuras públicas do âmbito político, como *Lula*, *Bolsonaro*, *fascista*, dentre outros. Por outro lado, árvores não tóxicas destacam termos relacionados à área do entretenimento, como música, filmes, jogos. A nuvem de termos exclusivos para árvores não tóxicas apresenta nomes de artistas, como *Caetano Veloso* e *Chico Buarque*.

Outros temas comuns se relacionam a finanças, como *poupança*, *financeira*. Já a nuvem de palavras exclusivas para árvores tóxicas apresenta termos de escopo político, como *venezuela*, *petê*, *ministros*, assim como termos de cunho negativo, como *ódio*, *repressão* e *extermínio*.

Tabela 1: Tópicos e seus termos mais frequentes

Tópicos	Termos mais frequentes
0	time, flamengo, bola, copa, futebol, palmeiras, corinthians, clube, mundial, brasil
1	mulher, mim, amigos, situação, algum, graça, sentido, comigo, ex, coisas
2	viu, fascista, the, esquerda, comentário, sub, ódio, inútil, brasil, sendo
3	m*rda, p*ta, pesquisas, primeiros, burro, tomar, receita, bolo, aceitar, chato
4	bolsonaro, famoso, sabia, veio, bolsonarista, lula, processo, pressão, época, preso
5	valeu, amigo, sinto, entendi, ajudar, amo, parabéns, vou, tranquilo, assistir
6	comida, gosto, ruim, comer, come, fome, qualidade, comprar, superior, and
7	violência, drogas, sociedade, eua, invés, droga, polícia, país, salário, guerra
8	estudo, explicar, mandar, pô, ideia, experiência, campeão, funciona, fonte, ensino
9	animal, leite, mesmos, vi, servem, comprei, atenção, animais, tô, mano
10	faculdade, curso, escola, trabalhar, média, ensino, ambiente, engenharia, aula, ótimo
11	mercado, uso, use, virou, dados, décadas, hoje, livre, morreu, disse
12	foto, rico, mostra, saiu, ar, idiota, dei, segue, conta, 15
13	jogo, jogos, vou, ler, jogar, canal, comentários, quiser, rede, li
14	inglês, verdade, lol, professor, mano, provavelmente, vcs, internet, natural, processar

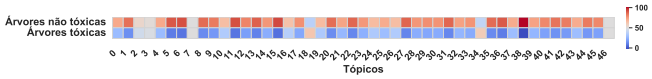

Figura 1: Proporção de comentários de cada árvore por tópico

Extração de tópicos: Foram extraídos 48 tópicos distintos, sendo um deles destinado a *outliers*, composto por 304 comentários. A Tabela 1 apresenta os 15 tópicos mais frequentes em ordem, que juntos cobrem 50,57% dos comentários totais. O tópico 0 possui 204 comentários e reúne termos de futebol (time, bola, copa), reflexo da Copa de 2022. Dentre as árvores não tóxicas, destacam-se os tópicos 5 e 6, relacionados à temas de amizade e culinária, respetivamente. Nas árvores tóxicas, destacam-se: os tópicos 2, que agrupa xingamentos a termos políticos (fascista, esquerda, ódio) e o 7, centrado em drogas e violência. O tópico 3, marcado por gírias e insultos, aparece tanto em contextos tóxicos quanto não tóxicos, indicando seu uso em ambos casos. A Figura 1 compara a proporção dos grupos de árvores por tópico. Além dos já mencionados, os tópicos sobre sexualidade e pornografia (18), moralidade sexual e religiosa (46) e as eleições de 2022 (34) são mais frequentes em árvores tóxicas, atuando como potenciais gatilhos.

Considerando as mudanças de tópicos, nas árvores tóxicas mudanças estão relacionadas ao tema sendo discutido: tópicos 42 (relacionado à gravidez e maternidade), 14 (educação e ensino) e 34 (dados de eleições) geram maior desvio de assunto ($\geq$ 50% das respostas mudam de tópico), enquanto tópicos 28 (bem-estar pessoal e relações sociais) e 41 (questões culturais na China) permanecem sem desvios. Por outro lado, em árvores não tóxicas, as mudanças estão mais frequentemente relacionadas aos tópicos 25 (discussões entre municípios), 17 (fatores religiosos concernentes ao aborto e à maternidade) e 33 (discussões acerca de decisões do Supremo Tribunal Federal). Tópicos como 38 (estilos arquitetônicos) e 45 (questionamentos sobre escolhas e tomadas de decisão) não apresentam desvios.

Reconhecimento de Entidades Nomeadas (REN): Árvores tóxicas concentram termos locais (LOC): 40% das entidades contra $\approx$ 32% em árvores não tóxicas. Por outro lado, elas trazem proporções ligeiramente menores de pessoas (PER), $\approx 27,4\%$ de participação, contra $\approx 31,8\%$ em árvores não tóxicas, e também pequenos decréscimos em organizações (ORG), 15% contra $\approx 16,1\%$.

Em suma, árvores tóxicas referenciam locais com mais frequência, organizações e outras categorias de entidades, mas mencionam menos pessoas. Interessante notar que, as nuvens de palavras mostram uma diferença entre pessoas que são citadas em cada classe de árvores: nas árvores não tóxicas, nomes de artistas sobressaem. Já nas árvores tóxicas, temos a menção de nomes de políticos. Além disso, em nuvens de termos exclusivos para árvores tóxicas, temos a presença de locais como *roraima* e *venezuela*, que podem estar atrelados a contextos sensíveis no escopo das discussões.

Pos Tagging: A distribuição de classes gramaticais é semelhante em ambos os conjuntos. Substantivos (NOUN) são a categoria mais frequente ($\approx 16,95\%$ em árvores não tóxicas, $\approx 17,77\%$ em árvores tóxicas), seguidos de verbos ($\approx 13,3\%$ em árvores não tóxicas, $\approx 13,28\%$ em árvores tóxicas) e de pontuação ($\approx 11,17\%$ em árvores não tóxicas e $\approx 10,53\%$ em árvores tóxicas). Contudo, em árvores tóxicas há certa predominância de conjunções subordinativas (SCONJ), que compõem $\approx 4,13\%$ dos termos em árvores tóxicas e $\approx 3,91\%$ para não tóxicas, enquanto as árvores não tóxicas trazem proporções um pouco maiores de nomes próprios (PROPN), apresentando $\approx 5,08\%$ da composição de árvores não tóxicas e $\approx 4,23\%$ para árvores tóxicas. A maior presença de conjunções subordinativas em árvores tóxicas pode refletir uma tendência a introduzir explicações, causalidades e condições, reforçando o tom argumentativo da discussão.

5 DISCUSSÃO FINAL E TRABALHOS FUTUROS

Neste estudo analisamos árvores de discussão do *Reddit* com o objetivo de distinguir padrões entre árvores tóxicas e não tóxicas. Partimos de um subconjunto rotulado manualmente quanto à toxicidade. Após validar o bom desempenho do GPT-4.1 nessa tarefa, aplicamos o modelo a todos os comentários restantes e comparamos métricas estruturais e linguísticas.

Nossos dados indicam que árvores tóxicas concentram um número maior de respostas, e que os subconjuntos de comentários posteriores aos gatilhos seguem essa tendência. Nessas árvores predominam termos relacionados à política e insultos. Adicionalmente, observamos uma maior frequência de mudanças de tópico e um uso mais frequente de conjunções subordinativas, o que sugere interações de caráter mais argumentativo. Em contrapartida, árvores de discussão não tóxicas compõem dominância em tópicos relacionados à amizade e culinária, e maior presença de pontuações nos comentários. As análises de padrões de linguagem, estilos argumentativos e tendências temáticas estruturam particularidades entre grupos de árvores com e sem gatilhos de discussão.

As principais limitações relacionados a nossas análises são as seguintes: (i) desbalanceamento de dados para o treinamento do modelo; (ii) o uso de um modelo de linguagem proprietário e; (iii) discussões realizadas somente no ano eleitoral de 2022, enviesando as discussões para tópicos políticos. No entanto, as análises deste artigo representam um passo inicial importante para o desenvolvimento de modelos preditivos capazes de identificar, em etapas iniciais, discussões com potencial de seguir para um caminho de interações agressivas e tóxicas, ou seja, identificar gatilhos.

Agradecimentos: Este trabalho foi parcialmente financiado pela FAPEMIG, CAPES e CNPq.

REFERÊNCIAS

[1] Ezgi Akar. 2025. Exploring the impact of social network structures on toxicity in online mental health communities. *Computers in Human Behavior* 165 (2025), 108542. https://doi.org/10.1016/j.chb.2024.108542

[2] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2020. Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 3033–3040. https://doi.org/10.1145/3366423.3380074

[3] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2022. PROVOKE: Toxicity trigger detection in conversations from the top 100 subreddits. *Data and Information Management* 6, 4 (2022), 100019. https://doi.org/10.1016/j.dim.2022.100019

[4] Henrico Bertini Brum and Maria das Graças Volpe Nunes. 2017. Building a Sentiment Corpus of Tweets in Brazilian Portuguese. arXiv:1712.08917 [cs.CL] https://arxiv.org/abs/1712.08917

[5] Gustavo Cunha and Ana Silva. 2024. Caracterizando Polarização em Redes Sociais: Um Estudo de Caso das Discussões no Reddit sobre as Eleições Brasileiras de 2018 e 2022. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 365–369. https://doi.org/10.5753/webmedia.2024.241688

[6] Tope Falade, Niloofar Yousefi, and Nitin Agarwal. 2024. Toxicity Prediction in Reddit. In *AMCIS 2024 Proceedings. 18.* https://aisel.aisnet.org/amcis2024/social_comp/social_comput/18

[7] E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. ASSIN: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal.* 13–15.

[8] Claudia Freitas, Paulo Rocha, and Eckhard Bick. 2008. A new world in Floresta Sintá(c)tica – the Portuguese treebank. *Calidoscópio* 6, 3 (2008), 142–148. https://doi.org/10.4013/cld.20083.03

[9] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs.CL] https://arxiv.org/abs/2203.05794

[10] Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. Toxic Content Detection in online social networks: a new dataset from Brazilian Reddit Communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, Pablo Gamallo, Daniela Claro, António Teixeira, Livy Real, Marcos Garcia, Hugo Gonçalo Oliveira, and Raquel Amaro (Eds.). Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, 472–482. https://aclanthology.org/2024.propor-1.48/

[11] Philip May. 2021. Machine translated multilingual STS benchmark dataset. https://github.com/PhilipMay/stsb-multi-mt

[12] Andreas Mueller. 2024. wordcloud. https://pypi.org/project/wordcloud/ Acessado em: 10/08/2025.

[13] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194 (2013), 151–175.

[14] Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 1779–1789. https://doi.org/10.18653/v1/2023.eacl-main.130

[15] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086* (2011).

[16] Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*. Pisa, Italy, 197–206. http://aclweb.org/anthology/W17-6523

[17] Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 406–412.

[18] Raquel Recuero. 2024. The platformization of violence: Toward a concept of discursive toxicity on social media. *Social Media+ Society* 10, 1 (2024), 20563051231224264.

[19] Vigneshwaran Shankaran and Rajesh Sharma. 2024. Analyzing Toxicity in Deep Conversations: A Reddit Case Study. arXiv:2404.07879 [cs.CL] https://arxiv.org/abs/2404.07879

[20] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

[21] spaCy. 2023. Portuguese Models. https://spacy.io/models/pt. Acessado em: 22/06/2024.

[22] Demetris Vrontis, Evangelia Siachou, Georgia Sakka, Sheshadri Chatterjee, Ranjan Chaudhuri, and Arka Ghosh. 2022. Societal effects of social media in organizations: Reflective points deriving from a systematic literature review and a bibliometric meta-analysis. *European Management Journal* 40, 2 (2022), 151–162. https://doi.org/10.1016/j.emj.2022.01.007

[23] Yulin Yu, Julie Jiang, and Paramveer Dhillon. 2024. Characterizing the Structure of Online Conversations Across Reddit. arXiv:2209.14836 [cs.SI] https://arxiv.org/abs/2209.14836