Diálogos Tóxicos: Gatilhos e Padrões de Interação no Reddit Brasileiro

Giovana Piorino[1], Marco Antônio de Alcântara Machado[1], Luiz Henrique Quevedo Lima[2], Adriana Pagano[1], Ana Paula Couto da Silva[1]

[1]Universidade Federal de Minas Gerais, [2]BEON.tech
giovana.piorino@dcc.ufmg.br, marcomachado@dcc.ufmg.br, luiz.quevedo@beon.tech,
apagano@ufmg.br, ana.coutosilva@dcc.ufmg.br

Resumo

In this paper we analyze the structural and linguistic dynamics of online toxicity in Reddit discussion trees, focusing on how trigger comments escalate conflicts in Brazilian Portuguese. Using a fine-tuned BERTAbaporu model, we show that toxic discussions are deeper, more engaging, and initially semantically cohesive, but degrade over time, while non-toxic interactions emphasize social bonding. Our findings contribute to a better understanding of toxicity escalation and support early detection of discursive conflicts.

1 Introdução

Nos últimos anos, as plataformas de mídia social online se tornaram ferramentas populares para interações globais, principalmente devido a natureza intrinsecamente interativa que promove um senso de pertencimento e conexão emocional (Vrontis et al., 2022). Estas plataformas passaram a ser um espaço para troca de informações e busca de apoio emocional para problemas do dia a dia.

Infelizmente, nem sempre estas interações são positivas. Estudos mostram que aproximadamente 5-10% das interações online são classificadas como tóxicas na maioria das plataformas moderadas, podendo atingir até 19% em plataformas não moderadas (Avalle et al., 2024). Estas interações tóxicas são caracterizadas por linguagem ofensiva, desrespeitosa ou emocionalmente prejudicial, o que pode afetar o engajamento dos usuários, diminuir a confiança dentro das comunidades e apresentar riscos à saúde mental de indivíduos mais vulneráveis (Maleki et al., 2022). Como exemplo, a pesquisa realizada com 29.593 estudantes universitários em Arif et al. (2024) fornece evidências da prevalência e do impacto desse problema na sociedade contemporânea, mostrando que a depressão e a ansiedade estão fortemente associadas a experiências de *cyberbullying*, uma forma de comportamento tóxico online.

Entre as diversas plataformas, a rede social *Reddit*[1] se destaca por suas dinâmicas de interação pautada em tópicos de discussão, bem como sua moderação voluntária. Por um lado, essa estrutura facilita discussões abertas e o compartilhamento, de forma anônima, de experiências pessoais (Proferes et al., 2021). Porém, apresenta desafios em termos de moderação e controle de conteúdo tóxico: estudos indicam que 16% dos usuários publicam postagens e comentários tóxicos, com 81% destes variando sua conduta entre comunidades para se adequarem às suas normas (Almerekhi et al., 2022a).

O comportamento tóxico de um ou mais usuários pode emergir em qualquer nível dentro de uma árvore de discussão (ou *thread*), mesmo em interações inicialmente positivas, fenômeno analisado em diferentes estudos da literatura (Yu et al., 2024; Almerekhi et al., 2020, 2022b). Esses trabalhos conduzem o conceito de gatilhos: comentários *não tóxicos* que provocam respostas *tóxicas* na estrutura das discussões. Identificar esses elementos é fundamental para a moderação no *Reddit*, permitindo aos moderadores adotar uma postura proativa frente à potenciais cenários de toxicidade.

Neste contexto, o objetivo principal deste trabalho é identificar e analisar árvores de discussão com e sem gatilho em comunidades do *Reddit* (em português brasileiro). Combinando propriedades linguísticas e métricas estruturais, buscamos identificar padrões que distingam elementos desencadeadores de toxicidade de dinâmicas conversacionais saudáveis, considerando o contexto de comunidades brasileiras na rede social. Nosso trabalho expande os resultados apresentados em (Piorino et al., 2025), buscando superar suas principais limitações: o alto custo atrelado ao uso de modelo proprietário e o tamanho reduzido da amostra de árvores de discussão. Para isso, propomos uma abordagem baseada em um modelo de código aberto, ampliamos conside-

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

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 581–590

ravelmente o volume de árvores analisadas e incorporamos novas ferramentas de processamento linguístico, proporcionando uma análise mais robusta das discussões e estrutura de comentários gatilho.

Nossas principais contribuições são as seguintes: **(i)** disponibilização de um conjunto de dados contendo 883.953 árvores de discussão do Reddit brasileiro; **(ii)** modelo de classificação de toxicidade para português brasileiro; e **(iii)** análises topológicas e linguísticas da toxicidade no contexto brasileiro. Os resultados são importantes para a proposta de mecanismos de moderação automática, como ferramenta complementar para mitigar comportamentos nocivos em plataformas sociais online.

2 Trabalhos Relacionados

Estudos sobre toxicidade priorizam a classificação de comentários e a caracterização linguística, com menor ênfase no impacto das discussões e em caracterização de gatilhos. Enfatizando o contexto do *Reddit*, observa-se predominância de pesquisas em inglês (Hiaeshutter-Rice e Hawkins, 2022; Kumar et al., 2023), sendo que modelos voltados à detecção e caracterização de toxicidade, bem como análises de impacto de comentários gatilho em discussões em português brasileiro (Cunha e Silva, 2024; Eberhart et al., 2021; Oliveira et al., 2023) ainda apresentam menor exploração.

Sob a ótica da dinâmica e métricas de toxicidade, estudos recentes focam na estrutura das conversas, no contexto do *Reddit*. Em Shankaran e Sharma (2024), o modelo *ToxiGen* foi empregado para analisar o espectro de toxicidade em textos de língua inglesa, concluindo que a hostilidade de um comentário é predominantemente determinada pelo seu antecessor imediato, corroborando achados de Maleki et al. (2022). Neste último, utilizando o modelo *Detoxify*, observou-se que discussões tóxicas tendem a ser mais longas, profundas e frequentemente terminam em toxicidade.

No contexto de detecção de gatilhos, definidos como comentários não tóxicos que suscitam respostas tóxicas, Almerekhi et al. (2020) utilizaram um classificador LSTM com representações GloVe, obtendo F1-Score de 0.83, também no contexto de discussões do *Reddit* em língua inglesa. Os autores concluem que mudanças abruptas de tópico e termos políticos atuam como fortes indicadores locais de gatilhos. Avançando nessa abordagem, os autores em Almerekhi et al. (2022b) propõem o modelo PROVOKE (Bi-LSTM), que integra *embeddings* textuais a atributos contextuais, como variação de sentimento, tópico e profundidade na árvore. Alcançando um F1-Score de 0.78, o estudo destaca que, além de termos ofensivos genéricos, o vocabulário específico de cada comunidade desempenha papel relevante na caracterização dos gatilhos.

Embora estudos anteriores tenham investigado como comentários inofensivos podem desencadear toxicidade em discussões online, essas pesquisas concentram-se predominantemente em contextos de língua inglesa. Nosso trabalho preenche essa lacuna ao analisar aspectos linguísticos que sinalizam potenciais gatilhos de toxicidade em interações no *Reddit* brasileiro, um contexto ainda pouco explorado na literatura.

3 Metodologia

Esta seção descreve a metodologia utilizada neste trabalho.

3.1 Conjunto de Dados

O conjunto de dados analisado neste trabalho foi compartilhado publicamente em Lima et al. (2024) e é formado por postagens e comentários feitos entre janeiro e dezembro de 2022 nos dez subreddits brasileiros mais populares neste período: *brasil, desabafos, futebol, saopaulo, eu_nvr, botecodoreddit, conversas, investimentos, tiodopave, brasilivre*. A sua coleta foi realizada através da API Pushshift.[2]

Resumidamente, os dados analisados incluem comentários apenas em português, excluindo comentários de comunidades que permitem discussões em múltiplos idiomas. Para o resultado final, 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 risadas.[3] Finalmente, foram excluídos comentários gerados por contas de moderação automática e bots que detectamos em nossos dados. Assim, nosso corpus, após a aplicação dos filtros, possui 6.589.541 milhões de comentários e aproximadamente 390.000 postagens.

3.2 Anotação Manual

Para a amostra rotulada manualmente, inicialmente foram selecionados 2.500 comentários de forma estratificada. Estes comentários foram divididos

[2]https://github.com/pushshift/api.

[3]Em português, textos de risadas são representados pela sequência de caracteres kkkkk

em cinco lotes de 500 comentários cada. Doze estudantes de graduação e pós-graduação em Ciência da Computação e Letras foram divididos em 4 grupos de anotação, sendo que o grupo que teve maior concordância entre os membros anotou dois lotes. Cada comentário foi classificado em uma das categorias: *Tóxico*, *Não Tóxico*, *Não sei*, ou *Informação insuficiente*, seguindo as definições estabelecidas pela API Perspective[4], que foram incluídas no conjunto de instruções fornecidas aos participantes do processo de anotação manual (Lima et al., 2024).

Como esperado, a porcentagem de comentários *Tóxicos* foi muito inferior à de *Não Tóxicos*. Para aumentarmos a quantidade de comentários anotados manualmente para esse rótulo, realizamos uma pré-anotação automática com a Perspective API[5], selecionando comentários considerados *Tóxicos* pelo modelo, e realizamos uma validação com um dos grupos de anotadores que participaram na primeira etapa de rotulação acerca do rótulo atribuído pela Perspective.

A partir dos resultados das rodadas de anotação, 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 2.961 comentários de interesse. Deles, 85,21% foram classificados como *Não tóxico*, e 14,79% como *Tóxico*. Tais comentários rotulados foram utilizados para o treinamento e validação do modelo de classificação de toxicidade proposto.

3.3 Modelo de Toxicidade para Português do Brasil

Para a rotulação automática de comentários como *Não tóxico* ou *Tóxico*, realizamos o ajuste fino do BERTabaporu, um modelo de linguagem para o português brasileiro construído sobre a arquitetura BERT e pré-treinado em aproximadamente 238 milhões de tweets (Costa et al., 2023). A escolha deste modelo foi motivada pela similaridade linguística entre seu domínio de pré-treinamento e nosso corpus alvo, sendo ambos textos informais de mídias sociais. Além disso, a ausência de custos financeiros torna essa solução mais viável em comparação aos Grandes Modelos de Linguagem (LLMs), especialmente diante da extensa base de dados a ser rotulada (Piorino et al., 2025).

Para adaptar o modelo para nossa tarefa de classificação binária de toxicidade, utilizamos a classe *AutoModelForSequenceClassification* da biblioteca Transformers do Hugging Face (Wolf et al., 2020).

Antes do treinamento, os comentários passaram por um pré-processamento de texto padrão para remover ruídos típicos de conteúdo de mídia social. Todo o texto foi convertido para minúsculas, e URLs, menções a usuários, hashtags, emojis, pontuação, números e stopwords foram removidos. *Stopwords* foram definidas de acordo com a lista de stopwords do português fornecida pelo corpus NLTK (Bird et al., 2009), garantindo consistência com as práticas de pré-processamento comumente adotadas na classificação de textos de mídia social (Leite et al., 2020; Vargas et al., 2022).

Visando mitigar o desbalanceamento das classes, adotamos a estratégia de aumento de dados anotados manualmente por humanos utilizando grandes modelos de linguagem (LLMs). Selecionamos 5.000 novos comentários do conjunto de dados, que foram classificados por 3 LLMs: Sabiá-2 (especializado em português) (Almeida et al., 2024), Llama-2(Touvron et al., 2023) e Mistral (Jiang et al., 2023). Somente os comentários com voto majoritário entre os modelos e aqueles que não foram previamente rotulados por humanos foram incluídos no conjunto de anotação automática (sintética). Assim, foram adicionados 3.385 comentários aos dados descritos na Seção 3.2. O prompt utilizado para a tarefa foi proposto em (Lima et al., 2024).

Os dados sintéticos foram incorporados exclusivamente ao conjunto de treinamento. O conjunto de teste é composto somente por dados anotados por humanos, garantindo uma avaliação livre de possíveis vieses introduzidos pelos grandes modelos de linguagem. A Tabela 1 apresenta os dados usados para definição do modelo. A inclusão dos dados sintéticos[6] quase dobrou os exemplos para treinamento, enriquecendo a classe de comentários *Tóxicos* cuja a acurácia de classificação é importante para a identificação da presença de gatilhos nas discussões analisadas. O modelo final foi disponibilizado publicamente[7].

Os hiperparâmetros de taxa de aprendizado e tamanho do lote foram definidos a partir de uma busca em grade, considerando a combinação dos valores $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$

[4]https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

[5]https://perspectiveapi.com/

[6]Foram filtrados comentários com rótulos *Não sei dizer, Falta Informação* e comentários duplicados

[7]https://huggingface.co/giovanaPcarvalho/bertabaporu-toxicity-reddit-ptbr

Tabela 1: Distribuição de comentários rotulados por fonte e classe.

Conjunto	Fonte da Anotação	Tóxicos	Não tóxicos	Total
Treino	Humanos	350	2.018	2.368
	Sintética (LLMs)	535	2.850	3.385
	Subtotal Treino	*885*	*4.868*	*5.753*
Teste	Humanos	88	505	593

e $\{16, 32, 64\}$, respectivamente. O modelo foi treinado usando validação cruzada com 5 *folds*. A configuração que maximizou o desempenho foi a taxa de aprendizado de 3×10^{-5} com tamanho de lote de 16, utilizando o otimizador AdamW e *Early Stopping* de 3 épocas.

3.4 Construção das Árvores de Discussão

Cada comentário coletado contém identificadores únicos, como o *id*, o *link_id* (que indica a postagem principal à qual pertence) e o *parent_id* (que aponta para o comentário imediatamente anterior na hierarquia). A partir dessas relações, é possível reconstruir a estrutura das conversas, conectando cada resposta ao seu respectivo comentário pai, gerando as sequências de discussões completas. Assim, consideramos como árvore de discussão toda a troca de comentários que se inicia a partir de uma postagem, seguida por um comentário resposta à ela, que é um comentário raiz. Respostas diretas e subsequentes ao comentário raiz são estruturadas e compõem uma árvore de discussão.

As árvores de discussão são divididas em dois grupos: *Não Tóxicas* e *Tóxicas*. Árvores *Não Tóxicas* são aquelas em que todos os comentários possuem rótulo *Não Tóxico*. Árvores *Tóxicas* possuem comentários *Tóxicos* e *Não Tóxicos*, mas possuem a seguinte característica: o comentário raiz deve ter rótulo *Não Tóxico*. Esta condição garante que a toxicidade na discussão sempre surja como uma resposta a um conteúdo *Não Tóxico*, definindo assim um gatilho de discussão. Importante notar que o gatilho pode estar em qualquer posição da árvore de discussão, uma vez que o comentário *Não Tóxico* deverá ser seguido por um comentário *Tóxico*, não sendo necessariamente a raiz da árvore. Árvores de discussão que não atendem a essa condição (como as 100% *Tóxicas* ou as iniciadas por uma raiz *Tóxica*) não são consideradas em nossas análises.

Nosso objetivo é identificar traços implícitos de linguagem que possam contribuir para mudanças de rumos nas discussões e identificar proativamente quando uma discussão possui o potencial de se tornar tóxica. Além disso, a partir dessa estruturação, comparamos estatísticas de profundidade entre os grupos de árvores de discussão e, em árvores *Tóxicas*, avaliar proporções de respostas tóxicas após um gatilho.

3.5 Caracterização Linguística

Para encontrar diferenças e similaridades entre árvores de discussão com gatilho (*Tóxicas*) e sem gatilho (*Não Tóxicas*), realizamos o conjunto de análises linguísticas descritas a seguir.

Similaridade Semântica. Para comparar a semântica dos comentários analisados, construímos um espaço semântico de alta dimensão $\mathcal{S}^n$. Cada comentário analisado é incorporado neste espaço através de sua representação vetorial. A codificação no espaço semântico é feita através do modelo de *embeeding paraphrase-multilingual-MiniLM-L12-v2*, um SBERT (Sentence-BERT) multilíngue, presente na biblioteca *Hugging Face*.[8] A escolha do modelo se deve ao fato dele ser treinado especificamente em tarefas de similaridade com suporte à língua portuguesa (Reimers e Gurevych, 2019).

Apesar de especificado para agrupar conteúdos com significado similar em regiões próximas, o espaço semântico $\mathcal{S}^n$ também captura, de forma indireta, similaridade de assuntos tratados, estruturas discursivas (retórica), padrões de argumentação, estilo e contexto. Ou seja, comentários similares, no sentido mais amplo da definição, estarão próximos no espaço vetorial construído. Assim, para quantificar o grau de proximidade entre pares de comentários adjacentes (um comentário de uma árvore de discussão e sua resposta imediata), calculamos a métrica de similaridade de cosseno.

Nosso interesse é observar quando a discussão em andamento muda significativamente, ou seja, possuem baixa similaridade dentro do espaço semântico construído. Para tal, definimos um limiar para determinar mudanças na similaridade entre pares de comentários, podendo ser um *proxy* para mudança de foco e estilo da discussão em análise, seguindo (Hearst, 1997). Este limiar é calculado usando a média e o desvio-padrão de todos os valores de similaridade (comentários par a par) das árvores analisadas e foi definido como 1 desvio padrão abaixo da média para

[8] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

considerar quedas de similaridade estatisticamente significativas (Hearst, 1997). Pares de comentários com similaridade abaixo desse limiar foram considerados uma tupla de mudança abrupta de discurso.

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 (Nothman et al., 2013). Essa abordagem classifica as entidades em quatro categorias: Pessoa (PER), Localização (LOC), Organização (ORG) e Diversos (MISC) e foi utilizada para identificar os principais atores citados nas interações, considerando que os dados abrangem um ano eleitoral e de grandes questões geopolíticas.

Etiquetagem de classe de palavra (Pos Tagging). Para identificar as classes gramaticais predominantes em cada tipo de árvore de discussão, aplicamos POS tagging (Petrov et al., 2011) com o modelo[9] disponível na biblioteca spaCy (spaCy, 2023), baseado em uma treebank (Rademaker et al., 2017) e anotada segundo o padrão Universal Dependencies (Freitas et al., 2008). Entre as possíveis funções do modelo, utilizamos especificamente o componente morfologizador, responsável pela atribuição de categorias gramaticais. Adicionalmente, expandimos a análise para a avaliação de n-gramas sintáticos (bigramas e trigramas) a partir das sequências de etiquetas. Essa abordagem, baseada em exemplos de estudos relacionados à n-gramas (Sidorov et al., 2014), permite capturar padrões estruturais gramaticais independentemente do conteúdo semântico ou tópico da discussão.

Classificação de Tipo de Discurso de Ódio. A classificação das categorias de discurso de ódio foi realizada utilizando o modelo *pysentimiento/bertabaporu-pt-hate-speech*, treinado para detectar diferentes tipos de discurso em Português (Fortuna et al., 2019; Pablo Botton da Costa, 2022) e disponível para uso por meio da biblioteca *Hugging Face*[10]. Fundamentada na arquitetura RoBERTa, a abordagem categoriza cada comentário em cinco classes principais: Sexismo (*Sexism*), Homofobia (*Homophobia*), Racismo (*Racism*), Ideologia (*Ideology*) calculando-se a probabilidade individual de enquadramento em cada rótulo, para cada comentário. O modelo foi aplicado exclusivamente aos comentários pertencentes às árvores *Tóxicas*, de forma a categorizar as tipologias de discurso nocivo. Com base em (Fortuna et al., 2019), os comentários foram rotulados com as categorias nas quais o modelo atribui probabilidade maior ou igual a 0,5 de presença de conteúdo associado.

[9] pt_core_news_lg

[10] https://huggingface.co/pysentimiento/bertabaporu-pt-hate-speech

Classe	Precisão	Recall	F1-Score
Não Tóxico (0)	0,93	0,95	0,94
Tóxico (1)	0,70	0,61	0,65
Acurácia			0,90
Média Macro	0,82	0,78	0,80
Média Ponderada	0,90	0,90	0,90

Tabela 2: Desempenho do modelo de toxicidade.

4 Resultados

A seguir apresentamos os resultados do trabalho.

4.1 Anotação Automática de Toxicidade

A Tabela 2 apresenta o desempenho do modelo de toxicidade descrito na Seção 3.3 e aplicado no conjunto de teste composto por 593 comentários anotados unicamente por humanos. Os resultados mostram que o modelo ajustado para a tarefa de toxicidade tem um bom desempenho ao classificar o conteúdo tóxico, principalmente quando comparado com outros modelos abertos apresentados na literatura (Leite et al., 2020), alcançando um F1-Score Médio Macro de 0,80. Vale ressaltar que o nosso objetivo neste trabalho é ter um modelo aberto, escalável e com desempenho satisfatório, a ser aplicado no grande conjunto de comentários das árvores de discussão a serem analisadas.

Complementando as métricas quantitativas de desempenho, a seguir apresentamos dois exemplos de comentários onde o nosso modelo divergiu em relação à rotulação humana.

O primeiro exemplo é um *falso negativo*, onde o modelo falhou em detectar toxicidade explícita, apesar do uso de adjetivos ofensivos direcionados a uma figura pública: *"mano esse luiz inacio ta ficando velho senil e psicopata na moral hahahaha"*. A inclusão de marcadores discursivos informais e risos atenua superficialmente o tom agressivo, provavelmente influenciando o modelo a interpretar o enunciado como não hostil. Esse comportamento é consistente com limitações observadas em tarefas que exigem inferência pragmática: expressões

de humor e marcações coloquiais podem suavizar a superfície linguística sem remover o conteúdo ofensivo (Davidson et al., 2017; Lewandowska-Tomaszczyk et al., 2023).

Por outro lado, o modelo apresentou casos de *falso positivos*, em que classificou indevidamente como tóxicos comentários que continham linguagem intensa, porém sem ataque interpessoal. Como exemplo, temos o comentário *"to bem f*-se, quero que o pau quebre"*, marcado pelo uso de palavras de xingamento, para expressar um possível sentimento de frustração. No entanto, modelos de toxicidade frequentemente associam palavrões à categoria tóxica de forma correlacional, sem distinguir entre xingamentos direcionados e manifestações gerais de emoção, resultando neste tipo de falso positivo.

Por fim, o modelo proposto foi aplicado ao conjunto de dados composto por 2.486.593 comentários das árvores candidatas a pertencerem às classes *Tóxica* e *Não Tóxica*. O total de comentários rotulados automaticamente como *Tóxicos* foi de 271.235 (10,91%) e como *Não Tóxicos* foi de 2.215.358 (89,09%), o que é consistente com a prevalência esperada de toxicidade em grandes comunidades online (Park et al., 2022).

4.2 Topologia das Árvores de Discussão

O modelo proposto foi utilizado para rotular os 2.486.593 comentários das árvores candidatas a pertencerem às classes *Tóxica* e *Não Tóxica*. Vale ressaltar que o critério para classificar uma árvore como *Tóxica* observa se o par subsequente de comentários tem um gatilho (comentário *Não Tóxico*) e uma resposta *Tóxica*, mesmo se o restante das respostas subsequentes forem rotuladas como *Tóxicas* ou *Não Tóxicas*. Após a validação dos critérios de classificação, o conjunto de árvores para análise consiste em 883.953 árvores de discussão, das quais 813.445 (92,02%) são *Não Tóxicas* e 70.508 (7,98%) são *Tóxicas*.

Todas as métricas estatísticas de profundidade avaliadas (média, mediana, máximo, quartis e desvio-padrão) revelaram-se superiores em árvores *Tóxicas*. O aumento da mediana e do terceiro quartil sugere que gatilhos fomentam réplicas sucessivas, postergando o encerramento rápido observado em árvores *Não Tóxicas*. Além disso, o maior desvio-padrão ($2,39$ em árvores *Tóxicas* contra $1,25$ em árvores *Não Tóxicas*) indica maior dispersão estrutural em discussões *Tóxicas*. A análise estrutural da discussão *Tóxica* após a ocorrência de um gatilho mostra que o volume médio de comentários subsequentes ao gatilho é de 2,11, representando em média 53,29% do volume total de comentários da discussão. Isso sugere que o gatilho não é um evento final isolado, e ocorre tipicamente na metade de uma interação média.

Além disso, o gatilho provoca um impacto significativo no andamento da conversa: as respostas subsequentes apresentam uma taxa média de toxicidade de 73,67%, revelando uma mudança expressiva no tom do diálogo após sua ocorrência. No entanto, essa escalada não se sustenta de forma contínua. Ao analisar sequências de respostas consecutivamente tóxicas após o gatilho, observamos um encadeamento médio de apenas 1,11, indicando que essas manifestações tendem a surgir de maneira pontual. Em outras palavras, embora as respostas imediatamente posteriores ao gatilho sejam majoritariamente tóxicas, elas não formam longas séries de conflito, mas aparecem isoladamente ou intercaladas com mensagens não tóxicas. Esse padrão sugere variações importantes no modo como os usuários reagem tanto ao gatilho quanto ao desenrolar da discussão.

4.3 Características linguísticas

Para a formulação das análises linguísticas, removemos ocorrências de árvores de discussão de profundidade dois, ou seja, apenas um comentário inicial com uma resposta, por não agregarem uma estrutura interessante em termos de evolução da discussão. Assim, após esse filtro, temos 321.174 árvores de discussão *Não Tóxicas* (compreendendo 1.230.816 comentários), e 47.860 árvores de discussão *Tóxicas* (com 225.939 comentários).

Similaridade Semântica. No nosso conjunto de dados, a média de similaridade dos 1.087.721 pares de comentários adjacentes foi igual a 0,395, sendo considerado como 0,21 o limiar de mudança de conteúdo entre interações sucessivas (1 desvio-padrão abaixo da média). Este valor foi utilizado para as análises descritas a seguir.

Considerando os conjuntos de árvores separadamente, para árvores de discussão *Tóxicas*, temos taxas de mudanças de similaridade menores (14,84%), entre pares, em comparação às árvores *Não Tóxicas* (17,29%). Isso sugere que as discussões tóxicas tendem a fixar mais o tópico da conversa. Além disso, a análise específica do par *Gatilho* $\rightarrow$ *Resposta* reforça essa observação, em que 14,62% desses pares apresentam similaridade

abaixo do limiar estabelecido, taxa ligeiramente menor que à do conjunto de árvores *Não Tóxicas*, indicando que o gatilho e sua resposta imediata apresentam leve tendência à manter o conteúdo semântico estabelecido na discussão, em comparação à pares de respostas em discussões *Não Tóxicas*.

A fim de observar tendências das taxas de similaridade semântica com os níveis de profundidade p das árvores, avaliamos as porcentagens médias de mudança de similaridade para cada nível isoladamente. Os resultados indicam que, para profundidades iniciais (3 a 10), as árvores *Tóxicas* apresentam porcentagens de mudança de similaridade inferiores com 16,09% contra 18,50% para $p = 3$. Essa tendência se mantém nas demais profundidades: $p = 4$ (15,77% contra 17,78%) e $p = 5$ (15,68% contra 16,49%), sugerindo uma leve inclinação à menor dispersão semântica em árvores *Tóxicas*.

Entretanto, observamos uma inversão desse padrão a partir de $p = 16$, onde a porcentagem de mudança em árvores *Tóxicas* (12,38%) ultrapassa a das *Não Tóxicas* (9,86%). Em discussões profundas ($p \geq 26$), a porcentagem de mudança em árvores *Tóxicas* atinge picos de 17,60%, contrastando com a estabilidade das discussões em árvores *Não Tóxicas* (inferior a 4%). Esse cenário, associado à maior incidência e extensão média das discussões *Tóxicas*, demonstra que a coesão semântica inicial da toxicidade se reverte conforme a árvore cresce, enquanto nas árvores *Não Tóxicas* a porcentagem de desvio tende à estabilização, independentemente da extensão da conversa.

Reconhecimento de Entidades Nomeadas. De forma mais geral, nossos resultados mostram que comentários de árvores *Não Tóxicas* apresentam uma maior diversidade de presença de entidades nomeadas, combinando, por exemplo, discussões em relação ao cenário eleitoral de 2022 e entretenimento, como discussões relacionadas ao futebol. Já as árvores *Tóxicas* centralizam as discussões em uma forte polarização política, decorrente da disputa presidencial. Neste contexto, a frequência relativa de figuras centrais como Lula e Bolsonaro, aumenta expressivamente nas árvores *Tóxicas*, em entre $\approx$ 10–13% em comparação as menções em árvores *Não Tóxicas* (6%).

Considerando os gatilhos, a Tabela 3 apresenta as dez entidades mais mencionadas nestes comentários. Embora existam menções a times de futebol, comentários gatilho são mais focados nos cenários da política e justiça (PT, STF, TSE).

Pessoas (PER)		**Organizações (ORG)**		**Locais (LOC)**	
Entidade	**%**	**Entidade**	**%**	**Entidade**	**%**
Bolsonaro	12,32	PT	6,68	Brasil	15,64
Lula	9,93	Flamengo	4,03	EUA	3,15
Ciro	1,73	Palmeiras	2,47	Argentina	1,67
Dilma	0,89	STF	2,32	Rússia	1,47
Neymar	0,88	Corinthians	1,88	Europa	1,40
Moro	0,82	TSE	1,13	China	1,39
Haddad	0,62	Vasco	1,02	Ucrânia	1,24
Putin	0,59	Globo	0,99	Portugal	0,92
Jesus	0,54	Inter	0,85	BR	0,86
Rapaz	0,39	MBL	0,82	São Paulo	0,83

Tabela 3: Entidades predominantes em comentários classificados como gatilho, ordenadas por porcentagem de ocorrência dentro do total de cada categoria.

Adicionalmente, termos associados a conflitos ideológicos ou figuras polêmicas (Moro, MBL) também estão entre as principais entidades citadas. Curiosamente, a categoria de Localidades (LOC) nos gatilhos mantém um padrão geopolítico similar ao restante do conjunto de comentários que compõem o grupo de árvores *Tóxicas*, com destaque para a polarização nacional (Brasil) e internacional (EUA, China, Rússia), sugerindo que a discussão sobre política externa também atua como ponto de toxicidade.

Etiquetagem de classe de palavra (Pos Tagging). Árvores *Tóxicas* apresentam a predominância um pouco maior de Substantivos (+0,64% mais ocorrências que nas árvores *Não Tóxicas*). Já árvores *Não Tóxicas* utilizam uma quantidade maior de Advérbios (+0,85%), indicando uma comunicação contextualmente detalhada.

Ao avaliar bigramas mais frequentes e combinações de rótulos presentes a partir deles, destacamos a combinação de bigrama das etiquetas *Substantivo + Adjetivo* (e vice-versa).[11] Podemos observar uma diferenciação no domínio semântico dos comentários: enquanto árvores *Não Tóxicas* são caracterizadas por construções voltadas à polidez, bem-estar e marcação temporal (como *boa sorte*, *dia feliz*, *ano passado*), árvores *Tóxicas* enfatizam a polarização ideológica e direciona xingamentos e palavrões. Bigramas como *lavagem cerebral*, *direita extrema* e *gente burra* demonstram que, em discussões tóxicas, há mais indícios de estruturas construídas para afetar outros usuários ou demarcar polarização político.

Ao analisarmos comentários-gatilho, nossos resultados mostram uma maior predominância

[11] Outras combinações de bigrama foram analisadas, sendo o resultado descrito o de maior relevância para a nossa análise.

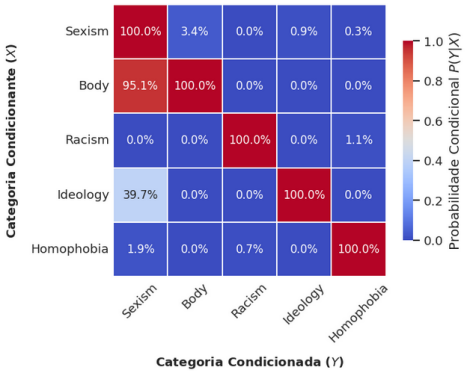

Figura 1: Matriz de co-ocorrência condicional ($P(Y|X)$) entre as categorias de discurso de ódio.

de conectivos de subordinação e marcadores de opinião, trigramas como *eu acho que*, *não sei se* e conjuntos de bigramas de etiquetas do tipo *Verbo + Conjunção Subordinativa*. Isso sugere que a toxicidade não é iniciada necessariamente por insultos, mas por expressões de dúvida ou discordância.

Classificação de Tipo de Discurso de Ódio. No conjunto de árvores de discussão *Tóxicas* foram identificados 12.669 comentários em que o modelo indicou $\geq 0,5$ de probabilidade de atribuição à pelo menos uma categoria de discurso de ódio.

Estes comentários estão distribuídos em 7.881 árvores, representando 16,47% do total de árvores analisadas. A pequena porcentagem de árvores com comentários rotulados pelo modelo possivelmente se deve à sua dificuldade em classificar com alto grau de certeza textos informais, com uso intensivo de gírias e com especificidades contextuais dos eventos ocorridos no ano de 2022.

Ao analisar a distribuição da confiança do modelo para valores acima do limiar 0,5, categorias *Body*, *Homophobia* e *Sexism* apresentam as maiores medianas de confiança ($\approx 0,83$), indicando maior assertividade do modelo para este tipo de conteúdo. Em contrapartida, a categoria *Racism* registra a menor mediana do conjunto ($\approx 0,63$), sugerindo que as manifestações racistas no *corpus* tendem a ser mais sutis, estruturalmente ambíguas, ou infrequentes, dificultando a atribuição de altas probabilidades pelo classificador. Assim, o rótulo predominante mais frequente em nosso dados é *Sexism*, com 10.176 comentários, superando consideravelmente os demais rótulos: *Homophobia* com 1.351, *Racism* com 828, *Ideology* com 162 e *Body* com 152 comentários.

Analisamos os padrões de co-ocorrência entre múltiplos rótulos (limiar $\geq 0,5$) na Figura 1. A matriz adota uma abordagem probabilística direcional ($P(Y|X)$), indicando a probabilidade de um comentário pertencer à categoria Y (coluna) dado que foi classificado na categoria X (linha). Diferentemente da correlação linear, essa análise revela assimetrias estruturais da toxicidade encontrada nos comentários. A categoria *Sexism*, por exemplo, atua como uma categoria dominante, absorvendo quase integralmente os ataques à aparência física: 95,1% dos comentários classificados como *Body* são simultaneamente sexistas. Isso sugere que, neste corpus, as críticas estéticas funcionam instrumentalmente como vetores de discurso sexista. Ressaltamos que essas tendências de interdependência se mantêm consistentes ao se observar especificamente as categorias dos comentários gatilhos.

5 Limitações

As limitações deste estudo, são as seguintes: (i) o desbalanceamento de classes no treinamento do modelo de detecção de toxicidade: apesar das técnicas de mitigação empregadas, a disparidade nas métricas de classe ainda é considerável e, embora uma validação humana das árvores de comentários classificadas como tóxicas fosse ideal para atestar a qualidade da detecção, essa alternativa mostrou-se inviável devido ao alto custo e tempo de anotação frente ao grande volume de dados processados; (ii) o viés político decorrente do recorte temporal da coleta de dados em 2022: por se tratar de um ano eleitoral de alta polarização, o léxico de comentários tóxicos avaliado está frequentemente atrelado ao debate político, como evidenciado na caracterização linguística deste trabalho, o que pode restringir a generalização do modelo para discursos de ódio em contextos mais cotidianos; e, por fim, (iii) o escopo restrito às dez comunidades analisadas: essa delimitação reduz a diversidade de assuntos e de estruturas de discussão avaliadas, uma vez que cada *subreddit* apresenta particularidades.

6 Conclusão

Este trabalho caracteriza a dinâmica discursiva do Reddit brasileiro com objetivo de distinguir as estruturas de árvores de discussões *Tóxicas* e *Não Tóxicas* considerando o impacto do fenômeno de gatilhos. Assim, elaboramos um ajuste fino ao modelo BERTAbaporu, a partir de dados rotulados manualmente e técnicas de *data augmentation*, e o aplica-

mos à totalidade dos comentários. A análise, fundamentada no conteúdo das árvores de discussão, integrou métricas de similaridade vetorial, morfossintáticas e semânticas para caracterizar a arquitetura das interações e a incidência de discurso de ódio.

Nossos resultados mostram que árvores *Tóxicas* apresentam maior profundidade e engajamento, com coesão semântica inicial que diminui em interações mais profundas. Nestas, predominam a polarização política e a sintaxe argumentativa, em contraste com a estabilidade e neutralidade temática das árvores *Não Tóxicas*. Adicionalmente, existe uma predominância de discurso de ódio sexista em discussões *Tóxicas*. Uma vez que que o tipo de árvore é definida pela presença ou não de um gatilho, nossos resultados indicam que de fato um gatilho influencia o conteúdo, estrutura e intensidade das discussões presentes no nosso conjunto de dados.

Nosso trabalho é um primeiro passo importante para fundamentar o desenvolvimento de mecanismos preditivos para identificar gatilhos, permitindo antecipar e mitigar a toxicidade com base nos padrões identificados em discussões realizadas em comunidades brasileiras no Reddit. Como trabalhos futuros, planejamos ampliar o escopo do estudo para incluir novos *subreddits* e atualizar a base de dados com discussões mais recentes, além de analisar individualmente cada comunidade e suas potenciais particulardades de estrutura conversacional.

Considerações éticas. Para a anotação manual, divulgamos apenas o texto do comentário, o *subreddit* e a data de publicação, preservando o anonimato típico dos usuários do Reddit. Metadados de *id* serviram exclusivamente para a reconstrução das árvores de discussão.

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

Referências

Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira, e Ramon Pires. 2024. Sabiá-2: A new generation of portuguese large language models. *Preprint*, arXiv:2403.09887.

Hind Almerekhi, Haewoon Kwak, e Bernard J Jansen. 2022a. Investigating toxicity changes of cross-community redditors from 2 billion posts and comments. *PeerJ Computer Science*, 8:e1059.

Hind Almerekhi, Haewoon Kwak, Joni Salminen, e Bernard J. Jansen. 2020. Are these comments triggering? predicting triggers of toxicity in online discussions. Em *Proceedings of The Web Conference 2020*, WWW '20, página 3033–3040, New York, NY, USA. Association for Computing Machinery.

Hind Almerekhi, Haewoon Kwak, Joni Salminen, e Bernard J. Jansen. 2022b. Provoke: Toxicity trigger detection in conversations from the top 100 subreddits. *Data and Information Management*, 6(4):100019.

Aahan Arif, Muskaan Abdul Qadir, Russell Seth Martins, e Hussain Maqbool Ahmed Khuwaja. 2024. The impact of cyberbullying on mental health outcomes amongst university students: A systematic review. *PLOS Mental Health*, 1(6):1–16.

Michele Avalle and 1 others. 2024. Persistent interaction patterns across social media platforms and over time. *Nature*, 628(8008):582–589.

Steven Bird, Ewan Klein, e Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Pablo Botton Costa and 1 others. 2023. BERTabaporu: Assessing a genre-specific language model for Portuguese NLP. Em *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, páginas 217–223. INCOMA Ltd., Shoumen, Bulgaria.

Gustavo Cunha e Ana Silva. 2024. Caracterizando polarização em redes sociais: Um estudo de caso das discussões no reddit sobre as eleições brasileiras de 2018 e 2022. Em *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web*, páginas 365–369, Porto Alegre, RS, Brasil. SBC.

Thomas Davidson, Dana Warmsley, Michael Macy, e Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. Em *Proceedings of the international AAAI conference on web and social media*, volume 11, páginas 512–515.

Carolina Eberhart, Luciano Ignaczak, e Márcio Martins. 2021. Text mining for cyberbullying detection: a brazilian portuguese evaluation. Em *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, páginas 92–100, Porto Alegre, RS, Brasil. SBC.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, e Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. Em *Proceedings of the Third Workshop on Abusive Language Online*, páginas 94–104, Florence, Italy. Association for Computational Linguistics.

Claudia Freitas, Paulo Rocha, e Eckhard Bick. 2008. A new world in floresta sintá(c)tica – the portuguese treebank. *Calidoscópio*, 6(3):142–148.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Dan Hiaeshutter-Rice e Ian Hawkins. 2022. The language of extremism on social media: An examination of posts, comments, and themes on reddit. *Frontiers in Political Science*, 4:805008.

Albert Q. Jiang and 1 others. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Deepak Kumar, Jeff Hancock, Kurt Thomas, e Zakir Durumeric. 2023. Understanding the behaviors of toxic accounts on reddit. Em *Proceedings of the ACM Web Conference 2023*, WWW '23, página 2797–2807, New York, NY, USA. Association for Computing Machinery.

João Augusto Leite and 1 others. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. Em *Proceedings of the 1st Conference of the Asia-Pacific and the 10th International Joint Conference on Natural Language Processing*, páginas 914–924, Suzhou, China. Association for Computational Linguistics.

Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene, e Slavko Žitnik. 2023. An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics*, 19(1):7–48.

Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, e Ana Paula Couto da Silva. 2024. Toxic content detection in online social networks: a new dataset from Brazilian Reddit communities. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, páginas 472–482, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Maryam Maleki, Mohammad Arani, Esther Mead, Joseph Kready, e Nitin Agarwal. 2022. Applying an epidemiological model to evaluate the propagation of toxicity related to covid-19 on twitter.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, e James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Felipe Oliveira, Victoria Reis, e Nelson Ebecken. 2023. Tupy-e: detecting hate speech in brazilian portuguese social media with a novel dataset and comprehensive analysis of models. *Preprint*, arXiv:2312.17704.

Pablo Botton da Costa. 2022. bertabaporu-base-uncased (revision 1982d0f).

Joon Sung Park, Joseph Seering, e Michael S. Bernstein. 2022. Measuring the prevalence of anti-social behavior in online communities. *Preprint*, arXiv:2208.13094.

Slav Petrov, Dipanjan Das, e Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Giovana Piorino, Luiz Lima, Adriana Pagano, e Ana Silva. 2025. Toxicidade e gatilhos: Um estudo de caso em comunidades do reddit no brasil. Em *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web*, páginas 575–579, Porto Alegre, RS, Brasil. SBC.

Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, e Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2).

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, e Valeria de Paiva. 2017. Universal dependencies for portuguese. Em *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, páginas 197–206, Pisa, Italy.

Nils Reimers e Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. Em *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Vigneshwaran Shankaran e Rajesh Sharma. 2024. Analyzing toxicity in deep conversations: A reddit case study. *Preprint*, arXiv:2404.07879.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, e Liliana Chanona-Hernández. 2014. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860.

spaCy. 2023. Portuguese models. `https://spacy.io/models/pt`. Acessado em: 09/11/2025.

Hugo Touvron and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Francielle Vargas and 1 others. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. Em *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, páginas 7174–7183, Marseille, France. European Language Resources Association.

Demetris Vrontis, Evangelia Siachou, Georgia Sakka, Sheshadri Chatterjee, Ranjan Chaudhuri, e 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):151–162.

Thomas Wolf and 1 others. 2020. Transformers: State-of-the-art natural language processing. Em *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, páginas 38–45. Association for Computational Linguistics.

Yulin Yu, Julie Jiang, e Paramveer Dhillon. 2024. Characterizing the structure of online conversations across reddit. *Preprint*, arXiv:2209.14836.