

Agressores e Alvos: O mapa dos ataques identitários no Reddit brasileiro

Beatriz Reis Gama Barbosa
beatriz.barbosa@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil

João Lucas Simões Moreira
joaomoreira@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil

Izadora Monken Ganem
izadoraganem@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil

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

ABSTRACT

Social media are virtual environments in which interactions are not always positive, especially since user identification is often not mandatory. Among different types of toxic content, attacks targeted at identity groups are particularly deserving of attention due to their potential to reinforce biases and marginalize already vulnerable communities. In this context, this work introduces a methodology to identify toxic content targeted at different identity groups in online platforms. This study identified approximately 73.4 thousand hostile messages on Brazilian Reddit, which allowed us to characterize the main groups targeted by the attacks. The analyses revealed that these attacks focused on subreddits with polarizing themes, such as politics and football, and the topic modeling highlighted different patterns between the attack types, with a strong recurrence of themes related to gender, mental health and political figures. Furthermore, the study of structural abuse patterns revealed an elevated prevalence of spurious abuse, which means most attacks received by a target occur in an isolated interaction with the aggressor. Our results suggest that specific and localized intervention measures are promising strategies to mitigate this behavior on social media and can support future user-targeted awareness campaigns, as well as aid the development of automatic moderation algorithms.

KEYWORDS

Toxicidade Online, Discurso de Ódio, Anotações Semânticas, Processamento de Linguagem Natural, Modelagem de Tópicos

1 INTRODUÇÃO

Na atualidade, as redes sociais são uma importante ferramenta de comunicação social e, não raro, servem de fórum

para disseminação de discursos que realizam ataques direcionados a grupos específicos. A esses ataques dá-se a denominação de “ataques identitários”, definidos por [3] como “linguagem que ataca, diminui e incita a violência ou ódio contra grupos, com base em características como aparência física, religião, descendência, origem nacional ou étnica, orientação sexual ou identidade de gênero.”

Segundo [9], esse tipo de discurso apresentou um crescimento preocupante nas redes sociais. Os autores identificaram uma tendência clara no aumento da influência de discursos tóxicos online. A gravidade desse fenômeno deve-se ao fato de que, ao contrário do discurso de ódio tradicional, ataques dessa natureza buscam incitar medo e insegurança em grupos específicos da sociedade.

Além disso, o direcionamento de ataques a grupos específicos transcende o ambiente virtual. Como afirma [2], a exposição a discursos de ódio online está associada a um aumento significativo nos sentimentos de insegurança dos alvos desses ataques, o que impacta sua percepção de segurança também em espaços físicos. Assim, a violência simbólica exercida no meio digital ultrapassa as fronteiras da internet e provoca consequências emocionais e sociais concretas no cotidiano de quem sofre os ataques.

Diante disso, este trabalho propõe a investigação de como se dão estes ataques na rede social Reddit, mais especificamente nas principais comunidades brasileiras, ou seja, nos mais populares subreddits no Brasil. Para orientar esta investigação, estabelecemos as seguintes questões de pesquisa:

QP1: *Qual é o comportamento de contas que realizam ataques identitários no Reddit e como podemos apoiar defesas específicas contra esses ataques?*

QP2: *Como podem ser classificados os ataques mais recorrentes em relação à sua natureza, seu alvo e os tópicos mencionados?*

Este trabalho introduz uma nova abordagem metodológica para identificar e caracterizar ataques identitários no ecossistema brasileiro do Reddit, unindo em um pipeline amostragem representativa em larga escala, classificação semântica de toxicidade, modelagem de tópicos e caracterização estrutural das interações abusivas. Inicialmente, coletamos e filtramos mais de três milhões de comentários dos dez subreddits brasileiros mais populares. Diante dos altos custos computacionais

de analisar grandes conjuntos de dados com modelos de linguagem avançados, implementamos uma técnica de amostragem estratificada. Essa abordagem preservou a distribuição conjunta do comprimento dos comentários e suas origens (subreddits), resultando em uma amostra altamente representativa de 100.000 comentários para análise semântica profunda.

Uma das principais contribuições técnicas deste estudo é a aplicação de um modelo ajustado (fine-tuned) da arquitetura Bidirectional Encoder Representations from Transformers (BERT) para detectar dez categorias distintas de toxicidade. Ao estabelecer um rigoroso limiar de confiança de 95%, o modelo classificou com sucesso aproximadamente 73.400 comentários como hostis. Refinamos nosso foco para os ataques especificamente baseados em identidade ao excluir obscenidades genéricas, garantindo, assim, que o conteúdo tóxico identificado estivesse estritamente alinhado com a marginalização identitária.

Por fim, nossa metodologia avança na compreensão do abuso online ao combinar a modelagem semântica de tópicos com a análise estrutural das interações. Utilizamos o modelo BERTopic apoiado por um embedding MiniLM otimizado para o português com o intuito de revelar os temas específicos que fundamentam diferentes tipos de ataques identitários. Adicionalmente, adaptamos métricas estruturais de abuso para classificar a experiência da vítima (padrões de receber experiences) em ataques espúrios, repetidos ou de inundação (flooding), fornecendo evidências acionáveis para o desenvolvimento de algoritmos de moderação automática e defesas localizadas.

2 TRABALHOS RELACIONADOS

Nosso trabalho baseia-se em estudos prévios sobre comportamentos abusivos em plataformas digitais, com destaque para a análise de [6], que investiga o comportamento de contas tóxicas no Reddit, e para [7], que propõe um modelo para a análise de sentimentos em comentários de diferentes subreddits brasileiros.

Embora o foco dos autores em [6] seja no estudo de toxicidade, sua metodologia proposta e seus resultados servem de base para a investigação de outro tipo de comportamento abusivo em ambientes virtuais: os ataques identitários. Entre as etapas apresentadas, destacamos a filtragem de corpus com regras claras para prevenção de ruído, como a seleção de comentários com comprimento entre 15 e 300 caracteres, exclusão de mensagens de baixa informação, emojis ou apenas pontuação o uso de ferramentas automatizadas para análise de discurso e detecção do alvo do ataque.

Já o trabalho de [7], por outro lado, tem como principais contribuições a coleta de subreddits Brasileiros, a proposta de um modelo de análise de sentimentos para textos em Português, bem como a caracterização dos comentários realizados.

Neste cenário, nosso trabalho amplia estas linhas de investigação ao **enfocar** uma categoria específica de comportamento abusivo (os ataques identitários), **adaptar** uma metodologia de grande escala para um contexto mais restrito e **integrar** a

Tabela 1: Número de entradas por rótulo de toxicidade.

Subreddit	Número de Entradas
r/brasil	2.136.866
r/brasillivre	1.219.265
r/futebol	1.214.412
r/desabafos	1.211.643
r/conversas	320.061
r/eu_nvr	188.620
r/investimentos	141.823
r/saopaulo	81.969
r/botecodoredit	57.298
r/tiodopave	11.584

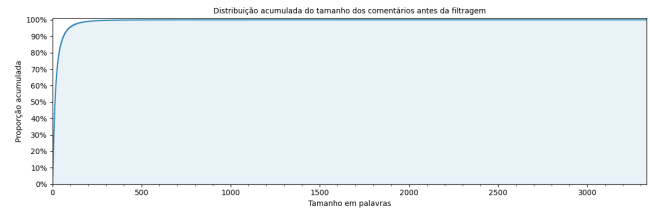


Figura 1: Distribuição do tamanho dos comentários do conjunto de dados completo.

metodologia aplicada a métodos mais atualizados, baseados em modelos de linguagem.

3 METODOLOGIA

A metodologia proposta pode ser decomposta em três etapas, sendo elas a coleta e o pré-processamento dos dados (subseção 3.1), a anotação dos textos (subseção 3.2) e a caracterização dos ataques (subseção 3.3).

3.1 Coleta e pré-processamento

Os dados utilizados nesta análise foram originalmente coletados por [7]. O conjunto original possui 6.589.541 linhas e 26 colunas, sendo que as principais variáveis incluem o nome do autor, o *subreddit* em que o comentário foi publicado, o conteúdo textual do comentário e informações relevantes, como a quantidade de palavras e de caracteres. Os dados foram coletados nos dez mais populares subreddits no Brasil considerando o total de comentários. A Tabela 1 apresenta os subreddits analisados e o total de comentários coletados, respectivamente.

Em relação ao conteúdo textual, é importante destacar que nenhum comentário da base é composto exclusivamente por emojis: esse filtro foi previamente aplicado por [7]. O subreddit mais recorrente é *r/brasil* (Tabela 1).

Na distribuição do número de palavras por comentário (Figura 1), a maioria se concentra na faixa entre 15 e 300 palavras. Assim, adotamos esse intervalo como critério de filtragem dos comentários para análises realizadas no trabalho. O conjunto resultante possui 3.160.852 comentários, cuja distribuição de tamanho está na Figura 2 e cuja média de palavras por comentário é cerca de 42.

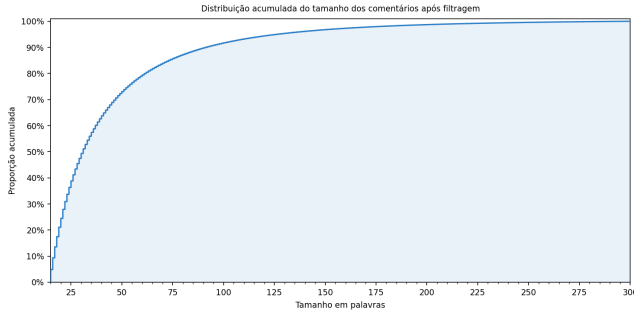


Figura 2: Distribuição do tamanho dos comentários após filtragem.

3.2 Amostragem estratificada

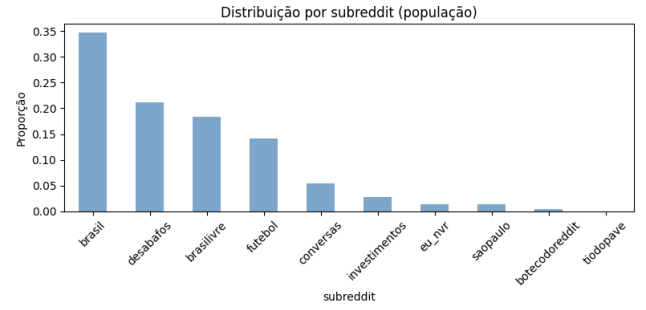
Devido ao alto custo de classificar a base completa com um modelo de linguagem, trabalhamos sobre uma amostra de $n = 100.000$ comentários. Para isso, adotamos uma **amostragem estratificada** que preserva a distribuição conjunta de duas dimensões: o *subreddit* de origem e a *faixa de comprimento* do comentário. A faixa de comprimento foi discretizada em cinco estratos por quantis, tornando a estratificação robusta à forte assimetria da distribuição de tamanhos. Cada estrato corresponde, portanto, a uma combinação *subreddit* $\times$ *faixa de comprimento*, e a quantidade de comentários amostrados em cada estrato é proporcional ao seu tamanho na população, garantindo que a amostra reproduza fielmente a composição temática e estrutural da base. Para ilustrar, um desses estratos é formado exclusivamente pelos comentários do *r/brasil* pertencentes ao primeiro quintil de tamanho (15 a 19 palavras), enquanto outro estrato agrupa os comentários do *r/futebol* situados no último quintil (62 a 300 palavras). A Figura 3 mostra a distribuição para cada amostra considerada.

3.3 Anotação dos textos

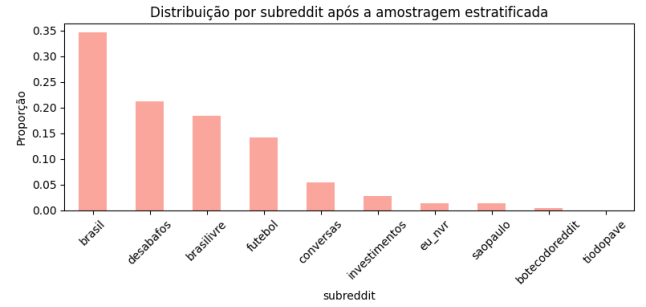
Para a identificação dos ataques identitários, adotamos métodos baseados em LLMs (Large Language Models). Utilizamos um modelo do tipo BERT (Bidirectional Encoder Representations from Transformers, [1]), treinado em textos em língua portuguesa e ajustado via fine-tuning para a tarefa específica de detecção de toxicidade.¹ O modelo classifica os comentários em dez categorias distintas de discurso tóxico (de modo que um comentário pode ser associado a mais de uma categoria), cujos rótulos são: *health*, *ideology*, *insult*, *lgbtqphobia*, *other_lifestyle*, *physical_aspects*, *profanity_obscene*, *racism*, *sexism*, *xenophobia*. Para maior precisão na identificação dos ataques identitários, foi utilizado um *threshold* de 95% na classificação dos comentários em qualquer uma das categorias de toxicidade.

Assim, para os comentários filtrados, o texto foi classificado como *hostil* ou *não hostil* pelo modelo. Aplicada à amostra estratificada de 100.000 comentários, a classificação resultou em **73.376 comentários hostis** (73,4% da amostra). Esta etapa

¹<https://huggingface.co/dougtrajano/toxicity-type-detection>



(a) Distribuição de amostras da população antes da filtragem.



(b) Distribuição de amostras da população após a filtragem

Figura 3: Distribuição de amostras.

resultou em uma base de dados contendo apenas comentários considerados hostis, cada um com seus rótulos e metadados associados.

Nesse contexto, o foco principal deste estudo é o subconjunto de comentários cujo conteúdo é direcionado a indivíduos ou grupos com base em características identitárias. Uma vez que o alvo do estudo são comentários que configuram ataques identitários, optamos por excluir o rótulo *profanity_obscene* e manter os restantes, que de fato se enquadram na definição de [3]. Além disso, mantivemos a categoria *insult* por considerarmos que esta constitui manifestações de toxicidade [4], de modo que ataques identitários são uma subcategoria, posteriormente contextualizada pelas análises temática e estrutural. A Tabela 2 apresenta a distribuição de entradas por rótulo associado.

3.4 Caracterização dos ataques

Após a coleta e a anotação dos dados, foi feita a caracterização dos ataques. Para isso, foram feitas análises relativas a subreddits (Seção 4.1), usuários (Seção 4.2) e tópicos de discussão (Seção 4.3) mais frequentes nos comentários identificados como ataques identitários. Adicionalmente, analisamos os *padrões estruturais de abuso* do ponto de vista das vítimas (Seção 4.4).

Tabela 2: Número de entradas por rótulo de toxicidade.

Rótulo de Toxicidade	Número de Entradas
insult	36.601
ideology	21.641
physical_aspects	5.660
xenophobia	4.736
other_lifestyle	2.663
sexism	1.921
racism	1.072
health	977
lgbtqphobia	857

Para a modelagem de tópicos, foi utilizado o modelo BERT-Topic ([5]), apoiado no modelo de embeddings *paraphrase-multilingual-MiniLM-L12-v2*², da família Sentence-BERT [8], que proporciona otimização do desempenho computacional sem perda significativa de qualidade. O modelo MiniLM, base desse embedding, é uma versão compacta e acelerada do BERT [10], com suporte nativo ao português e equilíbrio entre qualidade semântica e custo computacional. Para cada rótulo de toxicidade, os comentários correspondentes foram processados separadamente, de modo a revelar os assuntos predominantes de *cada* tipo de ataque. Nos casos em que o número de comentários de um rótulo excedia 10.000, sorteou-se um subconjunto de 10.000 para a extração de tópicos; caso contrário, utilizou-se o corpus completo do rótulo.

3.5 Seleção de hiperparâmetros

Os hiperparâmetros da redução de dimensionalidade (UMAP) e da clusterização (HDBSCAN) foram selecionados por *grid search*, usando a categoria *insult* como referência e a **coerência** c_v como métrica de qualidade dos tópicos. Variaram-se **n_neighbors** (10, 15, 20) e **min_cluster_size** (30, 50, 75). A melhor coerência ($c_v = 0,518$) foi obtida com **n_neighbors** = 10 e **min_cluster_size** = 50. A configuração final completa foi então aplicada às demais categorias.

4 RESULTADOS

Esta seção apresenta os resultados das nossas análises.

4.1 Subreddits

Quanto aos subreddits mais frequentes no quesito ataques identitários, o subreddit *r/brasil* destacou-se como aquele com maior quantidade de comentários considerados ataques identitários, seguido de *r/brasilivre*, *r/desabafos* e *r/futebol*. Medimos a **taxa de hostilidade** de cada subreddit apenas em comunidades com pelo menos 30 comentários para evitar taxas instáveis. A Tabela 3 apresenta os dez subreddits ordenados por essa taxa.

Tabela 3: Percentual de comentários tóxicos em cada subreddit.

Subreddit	Valor
brasilivre	0.815490
brasil	0.756905
futebol	0.738820
investimentos	0.712560
botecodoredit	0.706366
desabafos	0.679278
eu_nvr	0.634549
saopaulo	0.622348
tiodopave	0.594595
conversas	0.577421

Tabela 4: Dez usuários com mais comentários hostis na amostra, com o total de comentários hostis e a categoria de toxicidade dominante de cada um.

Usuário	Coment. hostis	Categoria dominante
Usuário 1	138	insult (70%)
Usuário 2	113	insult (49%)
Usuário 3	112	ideology (59%)
Usuário 4	111	insult (42%)
Usuário 5	110	ideology (57%)
Usuário 6	106	insult (54%)
Usuário 7	103	insult (61%)
Usuário 8	102	insult (86%)
Usuário 9	100	ideology (50%)
Usuário 10	96	insult (79%)

A concentração dos ataques em *r/brasil* sugere que muitos dos ataques identitários realizados são motivados por discussões políticas, tema central deste subreddit - assim como do segundo colocado, o *r/brasilivre*.

4.2 Usuários

A Tabela 4 lista os dez usuários com maior número de comentários hostis na amostra. O usuário mais ativo acumula 138 comentários hostis, e todos os dez ultrapassam 95, indicando que a hostilidade não se distribui uniformemente entre os autores, mas se concentra em um núcleo reduzido de contas altamente ativas. É importante ressaltar que o perfil de toxicidade desses usuários divide-se claramente entre dois eixos: contas dominadas por *insult* (como **Usuário 8**, com 86% de seus comentários hostis nessa categoria, e **Usuário 10**, com 79%), associadas sobretudo a comunidades como *r/futebol*, e contas dominadas por *ideology* (como **Usuário 3**, com 59%, e **Usuário 5**, com 57%), ligadas às discussões políticas. Isso revela que o alto volume de hostilidade em comunidades como *r/futebol*, que não lidera a taxa geral, é promovido por poucas contas especializadas em insultos, e não por uma hostilidade difusa entre todos os participantes.

²<https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>

Tabela 5: Tópicos identificados pelo BERTopic na categoria *lgbtqphobia* ($c_v = 0,3374$)

Tópico	Qtd.	Palavras mais frequentes
0	627	gay, pra, vc, mulher, homem, vai, pessoas
1	58	aborto, casos, argumento, vida, caso, pra, contra
2	56	pastor, igreja, religião, católico, padre, deus, pra

Tabela 6: Tópicos identificados pelo BERTopic na categoria *health* ($c_v = 0,3751$)

Tópico	Qtd.	Palavras mais frequentes
0	763	psicólogo, vida, ter, pessoa, bem, pra, vai
1	163	fluminense, jogo, pra, br, contra, brasil, centro

4.3 Tópicos

Para cada identificador, foi obtida uma lista das palavras mais frequentemente associadas aos tópicos extraídos do texto dos comentários. Por limitação de espaço, as tabelas contém os 7 termos mais relevantes de cada tópico.

A aplicação do BERTopic às nove categorias de ataques identitários (excluindo *profanity_obscene*) revelou padrões temáticos distintos conforme a natureza do rótulo de toxicidade. Na categoria *insult*, a mais prevalente (com taxa média de 36,9% dos comentários hostis), o modelo identificou apenas dois tópicos (Tabela 8), ambos de baixa especificidade lexical: o tópico dominante ($n = 9.805$) agrupa insultos genéricos sem âncora semântica clara, enquanto o segundo ($n = 115$) concentra referências a elementos religiosos como *deus*, *igreja* e *bíblia*. Esse resultado é coerente com a natureza difusa do insulto, que tende a ser contextualmente dependente e resistente à segmentação temática — limitação esperada em abordagens de modelagem de tópicos não supervisionadas aplicadas a linguagem hostil [5].

A categoria *ideology*, segunda mais prevalente (média de 21,7%), produziu a estrutura temática mais rica do corpus, com 21 tópicos (Tabela 7). Os tópicos refletem com fidelidade o espectro político-ideológico brasileiro: destacam-se agrupamentos sobre disputa eleitoral (T_2 : *lula*, *turno*, *ciro*, *votar*, *pt*; T_4 : *voto*, *candidato*, *eleição*; T_{10} : *bolsonaro*, *lula*, *votar*), polarização esquerda-direita (T_8 : *esquerda*, *direita*, *liberal*, *extrema*), referências ao comunismo e ao nazismo (T_9 , T_{11}), à crise Rússia-Ucrânia (T_3) e a negacionismo científico (T_{17} : *vacina*, *covid*, *pandemia*). A presença simultânea de tópicos sobre aborto (T_{18}), militares e golpe (T_{15}) e fascismo (T_{19}) corrobora a percepção de que os ataques ideológicos no Reddit brasileiro estão fortemente associados ao ciclo eleitoral de 2022 e à crescente polarização política documentada na literatura [9].

Nas categorias de identidade social (*racism*, *sexism* e *lgbtqphobia*) os tópicos identificados evidenciam como a hostilidade se ancora em marcadores específicos de grupo. Em *racism* (Tabela 11), o tópico T_0 ($n = 448$) agrupa termos

Tabela 7: Tópicos identificados pelo BERTopic na categoria *ideology* ($c_v = 0,5198$)

Tópico	Qtd.	Palavras mais frequentes
0	2.287	brasil, pra, dinheiro, vai, governo, país, ter
1	522	bolsonaro, vai, pra, cara, bozo, ainda, pt
2	493	lula, pra, vai, turno, ciro, votar, pt
3	384	rússia, ucrânia, putin, guerra, otan, eua, russia
4	370	voto, votar, votos, candidato, eleição, pra, político
5	317	lei, polícia, crime, corrupção, stf, pra, judiciário
6	289	pra, vai, lula, gado, aí, cara, ainda
7	251	cara, vai, acho, pra, fez, trump, ter
8	194	esquerda, direita, liberal, extrema, liberalismo, pessoas, liberais
9	178	comunismo, comunista, marx, comunistas, capitalismo, pra, estado
10	164	bolsonaro, lula, vai, ciro, votar, pt, lula
11	145	nazismo, nazista, hitler, nazistas, stalin, judeus, partido
12	144	reddit, news, fake, brasilivre, fake news, comments, brasilivre comments
13	141	igreja, deus, religião, igrejas, jesus, cristãos, evangélicos
14	104	homem, mulheres, homens, mulher, trans, minorias, direita
15	102	militares, forças, forças armadas, armadas, militar, exército, pra
16	85	china, coreia, chinês, país, eua, sul, taiwan
17	70	vacina, vacinas, covid, pandemia, vacinação, saúde, vacinar
18	62	aborto, vida, mãe, contra, existe necessidade, país, necessidade
19	56	fascismo, fascista, fascistas, bolsonaro fascista, vc, estado, vai
20	55	kkkkk, kkkkkkkk, kkkk, out, kkk, pra, kkkkkk

Tabela 8: Tópicos identificados pelo BERTopic na categoria *insult* ($c_v = 0,5183$)

Tópico	Qtd.	Palavras mais frequentes
0	9.805	pra, vai, ter, vc, bem, acho, fazer
1	115	deus, igreja, religião, pessoas, bíblia, pra, jesus

como *racismo*, *negro*, *branco*, *preto*, indicando discussões metalinguísticas sobre o próprio fenômeno do racismo, enquanto T_3 ($n = 59$) contém o termo *macaco*, configurando uso explicitamente depreciativo. Em *sexism* (Tabela 12), o tópico predominante (T_0 , $n = 716$) é focado em *mulher/homem*, e T_3 ($n = 56$) reflete ataques específicos a mulheres transexuais (*trans*, *mulher trans*, *cis*). Em *lgbtqphobia* (Tabela 5), além do tópico central com *gay* (T_0 , $n = 627$), identificam-se tópicos associados a argumentos religiosos (T_2 : *pastor*, *igreja*, *padre*, *deus*), apontando para uma interseção recorrente entre discurso de ódio LGBTfóbico e justificativas de cunho religioso.

Tabela 9: Tópicos identificados pelo BERTopic na categoria *other_lifestyle* ($c_v = 0,4938$)

Tópico	Qtd.	Palavras mais frequentes
0	560	pra, beber, vida, ter, bem, dia, vai
1	532	tratamento, pra, remédio, médico, psiquia- tra, vai, remédios
2	161	maconha, drogas, cocaína, droga, vai, uso, pra
3	152	rua, cidade, bairro, pra, centro, aqui, trans- porte
4	144	fumar, cigarro, pra, fumo, fuma, parar, ficar
5	133	brasil, pra, drogas, bolsonaro, cara, di- nheiro, pai
6	109	mulheres, mulher, homens, homem, homens mulheres, sobre, pornografia
7	79	arma, armas, polícia, policial, crime, polici- ais, tráfico
8	63	anticoncepcional, hiv, pílula, camisinha, toma, vc, pra
9	62	condomínio, imóvel, imobiliária, aluguel, apartamento, casa, pedreiro
10	61	maconha, cigarro, fumar, legalização, brasil, álcool, pode

Tabela 10: Tópicos identificados pelo BERTopic na categoria *physical_aspects* ($c_v = 0,4575$)

Tópico	Qtd.	Palavras mais frequentes
0	918	time, jogo, pra, flamengo, fluminense, pro, tá
1	478	peso, pra, dieta, comer, gordura, gordo, bem
2	454	mulher, beleza, feio, homem, aparência, feia, pra
3	400	academia, pra, fazer, vai, treino, bem, exer- cícios
4	256	anos, velho, idade, criança, vai, pra, vida
5	250	camisa, pra, vc, vai, porra, caralho, tá
6	233	cirurgia, médico, pra, dor, vai, fazer, dias
7	193	pra, cara, velho, ter, vai, ainda, falando
8	163	depressão, ansiedade, terapia, psicólogo, pra, psiquiatra, pessoa
9	153	pra, vc, pessoas, mental, algo, pessoa, ter
10	125	kkkkk, kkkk, kkkkkk, kkkkkkk, kkk, kkkkkkkk, pra
11	106	pra, mãe, porque, mulher, conselho, vai, cara
12	70	bike, bicicleta, carro, cidade, pra, calçada, moto
13	68	cabelo, barba, careca, crescer, pele, corte, franja
14	61	pra, dinheiro, pc, fazer, reais, vai, aí

A categoria *xenophobia* (Tabela 13) apresentou a maior variedade temática após *ideology*, com 14 tópicos. O tópico mais volumoso (T_0 , $n = 892$) agrega termos relativos à rivalidade futebolística com a Argentina (*brasil, argentina, copa, futebol*), sugerindo que o esporte motiva frequentemente a hostilidade xenofóbica no Brasil. Também se destacam tópicos

Tabela 11: Tópicos identificados pelo BERTopic na categoria *racism* ($c_v = 0,3126$)

Tópico	Qtd.	Palavras mais frequentes
0	448	racismo, pra, negro, branco, negros, preto, pessoas
1	311	brasil, pra, brasileiro, gente, branco, negro, negros
2	120	bolsonaro, pra, voto, nego, pt, cara, vero
3	59	macaco, macacos, pode, faz, lá, pra, todo

Tabela 12: Tópicos identificados pelo BERTopic na categoria *sexism* ($c_v = 0,3993$)

Tópico	Qtd.	Palavras mais frequentes
0	716	mulher, pra, mulheres, homem, vc, homens, vai
1	138	mulher, acho, cabelo, preto, homens, ho- mem, altura
2	56	brasil, mulheres, cidade, aqui, and, mulher, pra
3	56	trans, mulher trans, cis, homem, pessoa, mulheres, mulher
4	53	mulher, kkkkkk, kkkk, kkkkk, kkkkkkk, pra, kkkkkkkk

Tabela 13: Tópicos identificados pelo BERTopic na categoria *xenophobia* ($c_v = 0,4955$)

Tópico	Qtd.	Palavras mais frequentes
0	892	brasil, argentina, copa, brasileiro, futebol, pra, jogo
1	376	rússia, ucrânia, guerra, russo, russia, ucr- niano, pra
2	371	inglês, português, sotaque, aprender, língua, pra, espanhol
3	240	brasil, dólar, pra, imposto, eua, preço, aqui
4	227	rio, cidade, sp, paulo, rio janeiro, janeiro, pra
5	209	países, país, colônias, áfrica, europa, pra, europeus
6	116	china, chinês, chineses, amp, https, 32, pra
7	101	brasil, vacina, eua, casos, 000, covid, morte
8	79	japão, japonês, japoneses, japonesa, cul- tura, lá, bem
9	78	brasileiros, brasil, racismo, gt, racista, of, xenofobia
10	74	povo, curte, pra, índio, gosto, vc, aí
11	74	escola, pra, professor, escolas, educação, fa- zer, ensino
12	61	comida, comer, pizza, carne, restaurante, aqui, gosto
13	60	frio, calor, verão, inverno, sol, quente, dia

sobre o conflito russo-ucraniano (T_1), preconceito linguístico (T_2 : *inglês, português, sotaque, aprender*), comparações eco-
nômicas com os EUA (T_3 : *dólar, imposto, preço*), rivalidade
inter-regional entre São Paulo e Rio de Janeiro (T_4) e rela-
ções com o continente africano e o colonialismo (T_5 : *países*,

colônias, áfrica, europa, europeus). Esse conjunto revela que a xenofobia no corpus não se restringe a fronteiras nacionais, mas abrange também tensões regionais internas e disputas culturais e históricas.

Por fim, nas categorias **physical_aspects**, **health** e **other_lifestyle**, os tópicos identificados evidenciam que a hostilidade relativa a aspectos físicos e de estilo de vida frequentemente se manifesta em contextos de discussão sobre saúde mental, corpo e hábitos. Em **physical_aspects** (Tabela 10), emergem tópicos sobre peso e obesidade (T_1 : *peso, dieta, gordura, gordo*), aparência (T_2 : *beleza, feio, aparência, feia*), envelhecimento (T_4 : *anos, velho, idade*) e saúde mental (T_8 : *depressão, ansiedade, terapia, psicólogo*). Em **health** (Tabela 6), o tópico principal (T_0 , $n = 763$) concentra termos como *psicólogo, vida, pessoa, bem*, apontando para discussões em torno de saúde mental. Em **other_lifestyle**, destacam-se tópicos sobre uso de substâncias (T_2 : *maconha, drogas, cocaína*; T_4 : *fumar, cigarro*), tratamento médico (T_1 : *remédio, psiquiatra, tratamento*) e anticoncepção e saúde sexual (T_8 : *anticoncepcional, hiv, pílula, camisinha*). Esses resultados apontam para uma sobreposição temática entre os rótulos **health**, **physical_aspects** e **other_lifestyle**, sugerindo que ataques direcionados ao corpo, à mente e ao comportamento pessoal tendem a ocorrer em contextos semelhantes [3].

4.4 Padrões estruturais de abuso

Além de *o que é dito e por quem*, caracterizamos a forma estrutural dos ataques sob a perspectiva das vítimas, replicando a análise de *receiver experiences* de [6]. A partir da relação atacante $\rightarrow$ receptor, que foi reconstruída associando cada comentário hostil ao autor do comentário-pai. Em seguida, classificamos cada vítima segundo duas métricas: y , o número de atacantes distintos que realizaram ataques, e x , o número de alvos/posts distintos atingidos. A relação entre x e y revela os seguintes padrões:

- (1) **Abuso espúrio** ($y = x$), abuso pontual e isolado;
- (2) **Abuso repetido** ($y < x$), perseguição persistente por um mesmo agressor;
- (3) **Inundação** (*flooding*, $y > x$), múltiplos agressores convergindo sobre um mesmo alvo.

A Figura 4 apresenta a distribuição das vítimas nesses três padrões. O abuso **espúrio** domina massivamente, respondendo por **95,8%** das vítimas, seguido de **repetido** (2,8%) e **inundação** (1,4%). A direção do resultado alinha-se à do estudo original de [6] (76,1% / 15,8% / 8,1%): na maioria dos casos, a vítima sofre o ataque em uma interação pontual com o agressor, e não como alvo de perseguição sistemática ou de uma onda coordenada.

A predominância do abuso espúrio tem implicação direta na moderação: a maior parte da hostilidade não provém de perseguidores reincidentes contra alvos fixos, mas de interações isoladas e dispersas. Isso sugere a eficácia de estratégias de moderação localizadas.

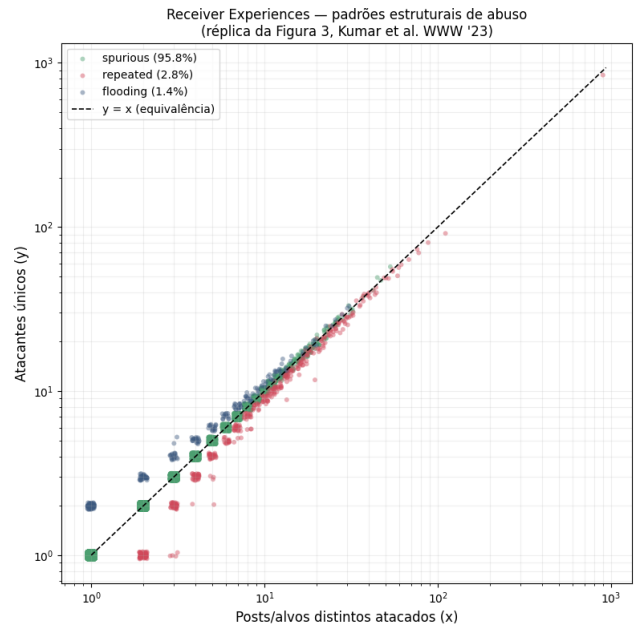


Figura 4: Padrões estruturais de abuso.

5 DISCUSSÃO

Nesta seção revisitamos as questões de pesquisa propostas e discutimos as limitações e questões éticas do nosso trabalho.

5.1 Questões de Pesquisa

Em relação à QP1 *Qual é o comportamento de contas que realizam ataques identitários no Reddit e como podemos apoiar defesas específicas contra esses ataques?*, os resultados apresentados nas Seções 4.1 e 4.2 mostram que é comum que os ataques identitários realizados nos subreddits sejam feitos por usuários que concentram seus ataques em comunidades específicas. Assim, nossos resultados fornecem indícios que medidas localizadas de combate a esse comportamento podem se mostrar eficazes para impedir a disseminação de conteúdo nocivo entre os usuários.

Em relação à QP2, *Como podem ser classificados os ataques mais recorrentes em relação à sua natureza, seu alvo e os tópicos mencionados?*, mostramos que a partir da análise dos tópicos realizada na Seção 4.3 e considerando a distribuição de ataques por tipo (Figura 2), os ataques mais recorrentes ocorrem na forma de insultos mas sem tópicos específicos definidos (à exceção do tópico relativo a futebol, provavelmente localizado em *r/futebol*).

Além disso, são massivos e se destacam os ataques identitários identificados como *ideology*, o que vai ao encontro da crescente polarização política no Brasil. Essa percepção é corroborada ainda pela presença dos tópicos relacionados à Lula e Bolsonaro e ao espectro político esquerda-direita.

Finalmente, em relação a alguns tipos menos recorrentes, como *xenophobia*, *sexism*, *racism* e *lgbtqphobia*, os tópicos

são bem definidos e revelam claras possibilidades de debate para ação combativa.

5.2 Limitações

As principais limitações do trabalho consiste na etapa de anotação dos textos, que foi realizada com o uso do modelo BERT devido à indisponibilidade de recursos para a realização de uma anotação manual por humanos. Além disso, conforme mencionado na Seção 3.2, a base de dados precisou ser reduzida de 3.160.852 comentários para uma amostra estratificada de 100.000 comentários devido ao alto custo e limitações no poder de processamento para classificação com o modelo de linguagem. Pela mesma razão, conforme mencionado na Seção 3.3, o conjunto de comentários de cada rótulo de toxicidade foram limitados por um sorteio de 10.000 amostras nos casos em que a categoria excedia esse volume para a extração de tópicos. Assim, para aumentar a precisão dos resultados, seria necessário dispor de maior poder computacional para a realização das tarefas de processamento de texto em toda a base de dados filtrada.

Cabe observar que, embora a modelagem de tópicos tenha utilizado uma lista de *stopwords* do português (via *CountVectorizer* com a lista da biblioteca NLTK), diversas palavras de alta frequência e baixo conteúdo semântico permanecem visíveis nas representações dos tópicos, em especial termos coloquiais e formas contraídas típicas da escrita informal em redes sociais. Esse fenômeno decorre de uma limitação conhecida das listas de *stopwords* padronizadas: construídas predominantemente a partir de textos formais, elas não contemplam adequadamente o registro coloquial e abreviado característico dos comentários de Reddit [3]. A presença dessas palavras é mais pronunciada nas categorias de menor especificidade temática, como *insult* (cujo tópico dominante, T_0 , apresenta como palavras mais representativas *pra*, *vai*, *ter*), e menos impactante nas categorias com vocabulário identitário mais restrito, como *racism* (*racismo*, *negro*, *branco*, *preto*, *racista*) e *lgbtqphobia* (*gay*, *pastor*, *igreja*, *aborto*), onde os termos distintivos sobressaem com maior clareza.

5.3 Cuidados éticos

Quanto aos aspectos éticos, este estudo utilizou exclusivamente postagens públicas do Reddit e foi conduzido em consonância com os princípios da Lei Geral de Proteção de Dados Pessoais (LGPD, Lei n.º 13.709/2018). A pesquisa não teve como objetivo identificar usuários. Todas as análises foram realizadas em nível agregado e os perfis com maior volume de postagens hostis foram anonimizados e tratados sob pseudônimos estruturais. Eventuais informações pessoais presentes nas postagens, como nomes, telefones, CPFs ou outros identificadores, foram desconsideradas e não integraram as análises. Devido à natureza sensível dos tópicos abordados, que incluem saúde mental e ataques a grupos marginalizados, os exemplos textuais extraídos da base foram truncados para mitigar riscos de reidentificação dos autores por meio de buscas diretas em motores de busca.

6 CONCLUSÃO

O presente trabalho apresenta a caracterização de ataques identitários em comunidades brasileiras do Reddit, com base em uma abordagem automatizada apoiada por modelos de linguagem. Ao analisar uma amostra representativa de 100 mil comentários (obtida a partir de uma base inicial filtrada de mais de três milhões), foi possível identificar aproximadamente 73,4 mil mensagens com características de discurso hostil que evidenciam ataques direcionados a identidades específicas.

Constatamos que tais ataques são frequentemente centralizados em subreddits voltados a temas polarizadores, como política e futebol, e impulsionados por um número restrito de usuários altamente ativos nessas comunidades. A extração de tópicos revelou que, enquanto algumas categorias de ataque apresentaram baixa coerência temática (como insultos), outras mostraram forte associação a temas sensíveis, como saúde mental, sexualidade, gênero, etnia e ideologia política.

Os padrões estruturais do abuso identificados concentram-se no *abuso espúrio*: a maior parte dos ataques sofridos pelas vítimas ocorre em uma única interação com o agressor. Esse achado reforça que intervenções localizadas têm potencial promissor para mitigar esse tipo de comportamento.

Os resultados indicam, portanto, que o enfrentamento dos ataques identitários deve ser orientado por estratégias contextualizadas, que considerem a especificidade das comunidades afetadas e os tópicos recorrentes nesses ataques.

Para trabalhos futuros, pretendemos aplicar a metodologia proposta em outras redes sociais, além do Reddit, com o objetivo de avaliar a generalização para diferentes contextos de interação. Adicionalmente, propomos construir e adaptar as listas de *stopwords* específicas para o português informal dessas plataformas, visando a reprodutibilidade do trabalho e o aumento da coerência e da interpretabilidade dos tópicos extraídos.

Agradecimentos. Este trabalho foi parcialmente financiado pela FAPEMIG, CAPES e CNPq. Para anotar os rótulos de toxicidade de cada comentário do Reddit, o modelo utilizado foi o *doutrajano/toxicity-typedetection*, pertencente à família de modelos BERT, disponível na plataforma Hugging Face. Como as anotações feitas por esse modelo foram fontes de análise no trabalho, a utilização desse método implica que os resultados estão sujeitos às limitações e aos possíveis vieses inerentes ao modelo. Para a modelagem de tópicos, foi utilizado BERTopic com o modelo de embeddings *paraphrase-multilingual-MiniLM-L12-v2*, pertencente à família Sentence-BERT (SBERT). Nessa etapa, a Inteligência Artificial foi empregada como ferramenta para representar semanticamente os comentários e identificar agrupamentos temáticos. Logo, os tópicos obtidos podem ter limitações ligadas à qualidade dos embeddings e às características do algoritmo de modelagem de tópicos.

REFERÊNCIAS

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019*

- Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.
- [2] Arne Dreißigacker, Philipp M"uller, Anna Isenhardt, and Jonas Schemmel. 2024. Online Hate Speech Victimization: Consequences for Victims' Feelings of Insecurity. *Crime Science* 13 (2024), 4. doi:10.1186/s40163-024-00204-y
 - [3] Paula Fortuna and S'ergio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *Comput. Surveys* 51, 4 (2018), 85:1–85:30. doi:10.1145/3232676
 - [4] P. T. M. Gibrim, M. V. G. Ribeiro, L. S. Moreira, P. de F. Melo, and J. C. S. Reis. 2025. I'm not Welcome Here – Uma Revisão de Escopo sobre Toxicidade e Discurso de Ódio em Jogos Digitais. In *Anais do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*. <https://sol.sbc.org.br/index.php/webmedia/article/view/38019>
 - [5] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to Create Easily Interpretable Topics. doi:10.5281/zenodo.4430182
 - [6] Deepak Kumar, Jeffrey Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the Behaviors of Toxic Accounts on Reddit. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. doi:10.1145/3543507.3583522
 - [7] Giovana Piorino, Vitor Moreira, Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. Análise de sentimentos de conteúdo compartilhado em comunidades brasileiras do Reddit: Avaliação de um conjunto de dados rotulados por humanos. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web (WebMedia)*. 54–62. doi:10.5753/webmedia.2024.242020
 - [8] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 3982–3992.
 - [9] Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saunabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. 2023. On the Rise of Fear Speech in Online Social Media. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2212270120. doi:10.1073/pnas.2212270120
 - [10] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, Vol. 33. <https://doi.org/10.48550/arXiv.2002.10957>