Caracterização e Comportamento de Usuários Tóxicos em Subreddits Brasileiros

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

[1] Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 – 31.270-901 – Belo Horizonte – MG – Brazil

[2]BEON.tech

{marcomachado, giovana.piorino, ana.coutosilva}@dcc.ufmg.br,

luiz.quevedo@beon.tech

Abstract. *While toxicity detection is well-studied, behavioral patterns of users generating such content in Portuguese-speaking communities remain underexplored. We analyze toxic accounts in the Brazilian Reddit ecosystem, processing 6.5 million comments from the top 10 subreddits in 2022 with a fine-tuned BERTabaporu model. A highly skewed distribution emerges: "Highly Recidivist" users (≈5% of toxic accounts) produce over 52% of harmful content via a broadcasting strategy. Toxicity concentrates in political and sports communities, spiking around real-world events, and is statistically rewarded with upvotes, suggesting engagement driven by moral outrage. These findings support behavior-focused moderation strategies.*

Resumo. *Embora a detecção de toxicidade seja bem estudada, os padrões comportamentais dos usuários que geram esse conteúdo em comunidades lusófonas permanecem pouco explorados. Analisamos contas tóxicas no Reddit brasileiro processando 6,5 milhões de comentários dos 10 maiores subreddits em 2022 com um modelo BERTabaporu ajustado. Usuários "Altamente Reincidentes" (≈5% das contas tóxicas) produzem mais de 52% do conteúdo nocivo via broadcasting. A toxicidade concentra-se em comunidades políticas e esportivas, com picos ligados a eventos reais, e é recompensada com upvotes, sugerindo engajamento movido por indignação moral. Os achados sustentam estratégias de moderação focadas no comportamento do usuário.*

1. Introdução

Redes sociais online ampliaram a comunicação e a formação de comunidades, mas também facilitaram comportamentos nocivos. Comentários tóxicos, tais como insultos, ameaças e assédio, estão entre as formas mais comuns de abuso online [Thomas et al. 2021] e podem comprometer o bem-estar dos usuários, levando à autocensura ou ao abandono de plataformas [Duggan 2017, Vogels 2021].

Entre as plataformas, o Reddit se destaca por sua organização em comunidades temáticas (*subreddits*) e pelo pseudonimato, que favorecem discussões intensas e dinâmicas particulares de conflito [Xia et al. 2020]. Embora haja ampla literatura sobre detecção automática de toxicidade, há menos atenção dedicada ao

comportamento dos usuários que a produzem, o que é essencial para estratégias de moderação mais eficazes [Kumar et al. 2023]. Os poucos trabalhos encontrados na literatura avaliam comportamento de usuários que compartilham conteúdo em inglês [Ribeiro et al. 2018, Kumar et al. 2023], deixando em aberto análises em larga escala do comportamento dos usuários no ecossistema do Reddit em português, cuja dinâmica difere de redes centradas em perfis e seguidores.

Neste trabalho, realizamos uma análise quantitativa em larga escala para caracterizar a reincidência e os padrões de atuação de contas tóxicas em *subreddits* brasileiros. Nossos resultados mostram uma assimetria pronunciada na produção de toxicidade: uma minoria hiperativa (cerca de 5% dos usuários tóxicos) responde por mais da metade do conteúdo nocivo detectado. Também observamos que surtos de toxicidade se alinham a eventos do calendário nacional e que, na maioria das comunidades analisadas, a hostilidade tende a receber maior validação social por meio de *upvotes*. Esses achados fornecem evidências quantitativas para subsidiar políticas de moderação orientadas ao comportamento do usuário.

2. Trabalhos Relacionados

Trabalhos recentes têm destacado que compreender toxicidade exige ir além do conteúdo isolado e analisar o comportamento de quem o produz [Kumar et al. 2023]. No Twitter, sinais textuais e estruturais do grafo social foram combinados para caracterizar usuários odiosos [Ribeiro et al. 2018]. Em uma perspectiva temporal, a intensificação do comportamento de ódio foi observada em plataformas com moderação permissiva [Mathew et al. 2020].

No Reddit, o histórico do usuário e o contexto do *subreddit* mostraram-se preditores relevantes para interações tóxicas [Xia et al. 2020]. Em larga escala, a toxicidade foi identificada como altamente heterogênea e concentrada em subgrupos específicos de usuários [Kumar et al. 2023].

No contexto brasileiro, conjuntos de dados anotados foram desenvolvidos para plataformas como Twitter, Instagram e portais de notícias [Leite et al. 2020, Vargas et al. 2022, Fortuna et al. 2019, de Pelle and Moreira 2017]. Para o Reddit brasileiro, Lima et al. introduziram um conjunto de dados com 2.500 comentários anotados manualmente, com foco na caracterização linguística do conteúdo tóxico [Lima et al. 2024]. Este trabalho complementa essa iniciativa ao escalar a análise para 6,5 milhões de comentários e deslocar o foco da caracterização linguística para os padrões comportamentais dos usuários tóxicos, abrangendo reincidência, dinâmica temporal e espacial, além de sinais de validação social.

3. Metodologia

3.1. Coleta e Pré-Processamento de Dados

O *corpus* compreende dados de 2022 dos dez maiores *subreddits* brasileiros, coletados via API Pushshift [Baumgartner et al. 2020]. O recorte temporal captura eventos de alta relevância como a eleição presidencial brasileira e a Copa do Mundo. Aplicamos um fluxo de filtragem para remover: (i) conteúdo indisponível (*[deleted]/[removed]*); (ii) mensagens sem valor semântico (apenas emojis, URLs, símbolos ou risadas); e (iii) automações

(*bots*). Do total inicial de 7,3 milhões, o conjunto final consolidado resultou em 6.589.541 comentários válidos, distribuídos conforme a Tabela 1.

Tabela 1. Distribuição de dados por subreddit (pós-filtragem).

Subreddit	Inscritos	Posts	Comentários
r/brasil	1.516.433	110.829	2.136.866
r/desabafos	490.049	115.876	1.211.643
r/futebol	369.925	35.826	1.214.412
r/brasilivre	210.582	67.301	1.219.265
r/conversas	247.545	21.967	326.061
r/eu_nvr	308.064	12.631	188.620
r/investimentos	232.485	9.756	141.823
r/saopaulo	358.681	7.308	81.969
r/botecodoreddit	270.451	7.059	57.298
r/tiodopave	219.926	2.371	11.584
Total	**-**	**390.924**	**6.589.541**

3.2. Detecção de Toxicidade

A identificação de conteúdo tóxico em português brasileiro exige um modelo adaptado tanto ao idioma quanto ao domínio informal de mídias sociais. Neste trabalho, adotamos a definição de toxicidade da Perspective API[1], seguindo Lima et al. (2024): um comentário é considerado tóxico quando contém linguagem rude, desrespeitosa ou não razoável, com potencial de levar um usuário a abandonar uma discussão. O modelo utilizado neste trabalho foi desenvolvido especificamente para o Reddit brasileiro [Piorino et al. 2026]; resumimos aqui seus principais aspectos, uma vez que o foco deste trabalho é a caracterização comportamental dos usuários que produzem esse conteúdo.

O conjunto de treinamento foi construído a partir do processo de anotação manual introduzido para o Reddit brasileiro, do qual este trabalho utiliza o conjunto completo de 2.961 comentários anotados (85,21% *Não Tóxico*, 14,79% *Tóxico*), particionados em treino (2.368) e teste (593). Para mitigar o desbalanceamento de classes, o conjunto de treino foi complementado com 3.385 comentários reais adicionais, cujos rótulos foram gerados automaticamente por consenso de votação majoritária [Ding et al. 2024] entre os grandes modelos de linguagem (LLMs) Sabiá-2 [Almeida et al. 2024], Llama-2 [Touvron et al. 2023] e Mistral [Jiang et al. 2023] – aplicados a comentários reais ainda não anotados por humanos. Ao final, o conjunto de treinamento totalizou 885 exemplos rotulados como *Tóxico* e 4.868 como *Não Tóxico*, enquanto o conjunto de teste permaneceu com anotação exclusivamente humana (88 *Tóxico* e 505 *Não Tóxico*).

Realizamos o ajuste fino do BERTabaporu [Costa et al. 2023], modelo pré-treinado em 238 milhões de *tweets*. A escolha do BERTabaporu foi motivada pela similaridade entre seu domínio de pré-treinamento – registro informal, de curta extensão e em português do Brasil – e o perfil linguístico dos comentários do Reddit brasileiro. Após otimização de hiperparâmetros via validação cruzada estratificada com 5 *folds*, o classificador atingiu acurácia de 0,90 e F1-Score Macro de 0,80 (F1 de 0,65 para a classe *Tóxico*), desempenho comparável ao reportado por modelos anteriores para detecção de toxicidade

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

em português [Leite et al. 2020], indicando robustez suficiente para análise de tendências em larga escala. A análise qualitativa dos erros evidencia limitações pragmáticas do classificador. Por exemplo, como *falso negativo*, o modelo não detectou toxicidade em *"mano esse luiz inacio ta ficando velho senil e psicopata na moral hahahaha"*, possivelmente devido ao tom informal e ao riso. Como *falso positivo*, classificou como tóxico *"to bem f*-se, quero que o pau quebre"*, embora o comentário expresse frustração sem ataque interpessoal direto.

Aplicado ao *corpus* completo de 6.589.541 comentários, o modelo identificou 575.486 comentários tóxicos (8,73%), prevalência consistente com estudos anteriores [Park et al. 2022].

3.3. Caracterização do Perfil dos Usuários

Para definir perfis de reincidência de comportamento tóxico, analisamos a distribuição empírica de comentários tóxicos (x_u) por usuário. Utilizamos os percentis P_{75} e P_{95} para dividir os usuários em três grupos: (i) **Pouco Reincidentes ($x_u \leq 7$):** representam 75% da população, com toxicidade esporádica; (ii) **Moderadamente Reincidentes ($8 \leq x_u \leq 44$):** usuários de atividade intermediária; e (iii) **Altamente Reincidentes ($x_u > 44$)**. O limiar de 44 comentários tóxicos corresponde ao 95-percentil da distribuição empírica de toxicidade por usuário, isolando a cauda superior de usuários mais recorrentes.

Para aprofundar a análise de comportamento dos usuários, definimos as seguintes métricas:

1. Proporção de Toxicidade (P_u): Mede a proporção do comportamento tóxico em relação ao total de comentários do usuário (v_u):

$$P_u = \frac{t_u}{v_u} \times 100\% \tag{1}$$

Essa métrica distingue usuários predominantemente hostis daqueles que, apesar de muitos comentários tóxicos (t_u), possuem majoritariamente participação construtiva.

2. Alcance de Toxicidade ($|A_u|$): Número absoluto de discussões distintas (*threads*) em que o usuário emitiu ao menos um comentário tóxico, indicando sua dispersão pelo espaço de discussões da plataforma.

3. Densidade de Toxicidade (D_u): Relaciona o volume de toxicidade com a capilaridade para verificar o padrão de ataque:

$$D_u = \frac{t_u}{|A_u|} \tag{2}$$

Valores de $D_u \approx 1$ indicam que o usuário distribui seus comentários tóxicos por muitas discussões distintas, comportamento de espalhamento (*broadcasting*), enquanto valores altos denotam concentração de múltiplos comentários tóxicos em poucas discussões específicas.

4. Validação Social: Para verificar a resposta da comunidade, ajustamos um modelo de Regressão Linear (OLS) para cada *subreddit*, estimando a associação da toxicidade no total de *upvotes*:

$$Score = \alpha + \beta \cdot (IsToxic) \tag{3}$$

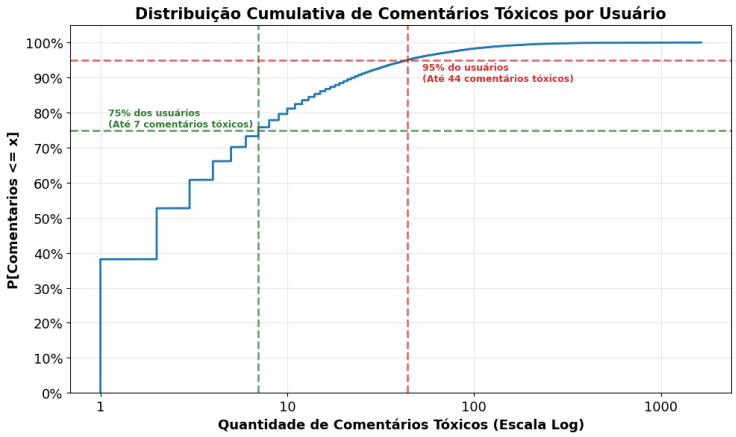

Figura 1. Distribuição Cumulativa (CDF) de Comentários Tóxicos por Usuário. As linhas tracejadas indicam os cortes percentuais que definem os perfis Pouco Reincidentes (75%) e Altamente Reincidentes (95%)

onde $IsToxic \in \{0, 1\}$ é uma variável binária que indica se o comentário foi classificado como tóxico pelo modelo, α corresponde ao *score* médio de comentários não tóxicos e β indica a diferença média de *score* associado à presença de toxicidade. Assim, $\beta > 0$ indica que comentários tóxicos têm, em média, maior *score* (maior validação), enquanto $\beta < 0$ indica punição média.

4. Resultados

4.1. Perfis de Usuários Tóxicos

Aplicamos o modelo ao *corpus* completo formado por 6.589.541 comentários e 148.779 usuários únicos participando das discussões, identificando 575.486 comentários *Tóxicos* (8,73%), prevalência consistente com estudos anteriores [Park et al. 2022]. Para a caracterização dos usuários, restringimos a análise a 561.480 comentários atribuídos a 52.267 usuários únicos, excluindo aqueles marcados como *[deleted]* por não permitirem atribuição de autoria individual. Ou seja, 35,10% da base de usuários ativa nas discussões analisadas manifestou algum grau de comportamento tóxico no período. A Figura 1 apresenta a distribuição cumulativa de comentários tóxicos por usuário. A curva evidencia a alta concentração de usuários com baixa atividade tóxica (o crescimento rápido inicial), seguida por uma cauda longa formada pelos usuários mais toxicamente ativos, típica de redes sociais [Ribeiro et al. 2018].

A análise da distribuição de comentários por usuário revela que, para a vasta maioria de usuários, a toxicidade é um evento isolado. Mais precisamente, o ecossistema de toxicidade é sustentado por uma minoria hiperativa (Tabela 2). O perfil *Altamente Reincidente*, embora represente apenas 4,97% dos usuários tóxicos, é responsável por produzir mais da metade (52,36%) de todo o conteúdo nocivo identificado. Em contraste, o grupo majoritário *Pouco Reincidentes* (75% dos usuários tóxicos) contribui com menos de 16% do volume total de toxicidade.

4.2. Dinâmica Temporal do Volume de Toxicidade

A toxicidade em plataformas online é reativa e profundamente influenciada por eventos do mundo físico [Olteanu et al. 2018]. A Figura 2 apresenta a evolução semanal do to-

Tabela 2. Caracterização dos perfis e impacto no volume total de toxicidade.

Perfil	Intervalo (t_u)	Num. Usuários	Vol. Tóxico	% do Vol.
Pouco Reincidente	≤ 7	39.618	88.574	15,78%
Moderadamente Reincidente	$8 - 44$	10.053	178.936	31,86%
Altamente Reincidente	≥ 45	2.596	293.970	52,36%
Total	**-**	**52.267**	**561.480**	**100,00%**

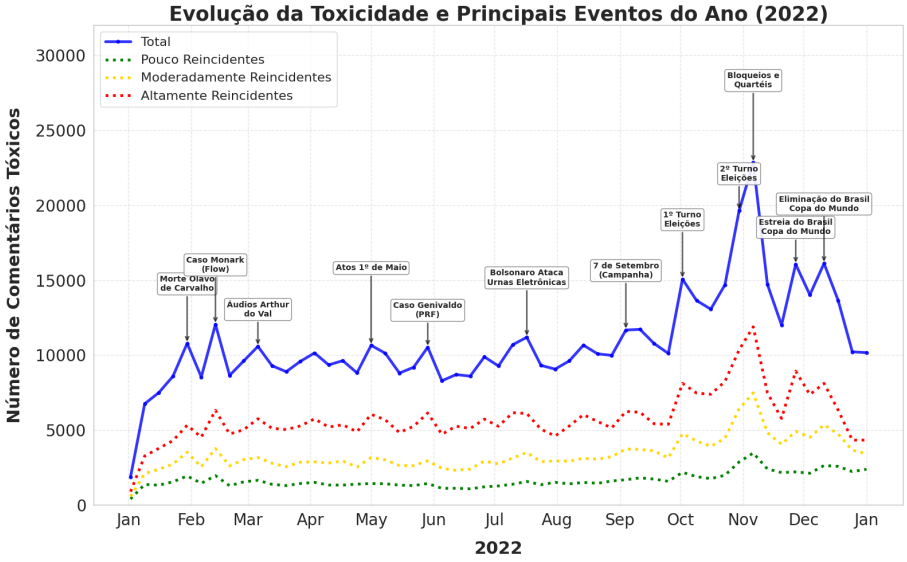

Figura 2. Série temporal semanal de comentários tóxicos.

tal de comentários tóxicos ao longo de 2022. Esta série temporal revela que os surtos de toxicidade apresentam uma forte sobreposição temporal com eventos de grande repercussão no calendário político e social brasileiro, identificados na figura. Adicionalmente, podemos observar visualmente como as atividades dos diferentes perfis de reincidência são determinantes para o volume de dados tóxicos compartilhados durante este período de análise. A seguir, discutimos estes resultados.

A Tabela 3 detalha eventos externos temporalmente sobrepostos aos principais picos de toxicidade identificados na Figura 2, sugerindo possíveis catalisadores dos surtos de hostilidade. Observa-se predominância de eventos políticos, incluindo escândalos de figuras públicas [CNN Brasil 2022b, CNN Brasil 2022a, G1 2022c], episódios de violência policial e atos de rua [Uol 2022c, Uol 2022a], além da escalada eleitoral entre julho e novembro [G1 2022a, G1 2022b, G1 2022d, Uol 2022b]. Ao final do ano, os picos se deslocam para o futebol, coincidindo com a participação do Brasil na Copa do Mundo [Globo Esporte 2022].

Adicionalmente, os números apresentados revelam que as comunidades *r/brasil* (comunidade focada em assuntos gerais) e *r/brasilivre* (comunidade com forte viés político) alternam a liderança da toxicidade em momentos políticos, com a comunidade *r/brasil* sendo a mais tóxica durante eventos de grande impacto nacional (ex: eleições e bloqueios de estradas). Já a comunidade *r/futebol* assume o protagonismo durante a Copa do Mundo. Isso demonstra que a toxicidade é temática e migra para onde está o foco da atenção pública que, no ano de 2022 no Brasil, foi marcada pelo futebol e pelas eleições presidenciais.

Tabela 3. Semanas de pico de toxicidade, eventos associados e a comunidade com maior contribuição para o volume tóxico do período.

Semana	Evento Principal	Vol. Tóxico	Comunidade Predominante (% do Total)
30/Jan	Morte de Olavo de Carvalho	10.630	r/brasilivre (38,4%)
13/Fev	Caso Monark (Flow Podcast)	11.448	r/brasil (33,3%)
06/Mar	Áudios de Arthur do Val	10.045	r/brasilivre (38,8%)
01/Mai	Atos de 1º de Maio	10.334	r/brasilivre (37,4%)
29/Mai	Caso Genivaldo (PRF)	10.354	r/brasilivre (36,4%)
17/Jul	Ataques às Urnas Eletrônicas	11.030	r/brasil (28,7%)
04/Set	Mobilizações de 7 de Setembro	11.534	r/brasil (33,5%)
02/Out	Eleições – 1º Turno	14.878	r/brasil (49,0%)
30/Out	Eleições – 2º Turno	19.372	r/brasil (42,8%)
06/Nov	Bloqueios em Estradas	22.189	r/brasil (43,0%)
27/Nov	Copa do Mundo – Estreia do Brasil	15.745	r/futebol (30,1%)
11/Dez	Copa do Mundo – Eliminação do Brasil	15.729	r/futebol (33,8%)

Considerando a série temporal apresentada na Figura 2, podemos observar que a curva de conteúdo compartilhado pelo perfil *Altamente Reincidente* (linha vermelha pontilhada) acompanha com alta fidelidade as oscilações do volume total (linha azul sólida). A análise estatística corrobora essa observação visual: tanto o perfil *Altamente Reincidente* quanto o perfil *Moderadamente Reincidente* apresentam correlações de Pearson extremamente altas com o total de comentários tóxicos ($\rho = 0.97$ e $\rho = 0.98$, respectivamente), provavelmente indicando que estes perfis tendem a discutir eventos do mundo real nesta plataforma.

Entretanto, a proximidade numérica dos coeficientes deve ser contextualizada pela magnitude de cada grupo. Enquanto a alta correlação do perfil *Moderadamente Reincidente* é sustentada por cerca de 20% da base de usuários tóxicos, o perfil *Altamente Reincidente* consegue replicar as tendências de pico da curva total dependendo de apenas 5% dos usuários. Essa desproporção indica que as oscilações abruptas de toxicidade na plataforma não refletem uma mudança de comportamento do usuário comum (perfil *Pouco Reincidente*, com $\rho = 0.89$ e curva estável), mas são fenômenos definidos majoritariamente pela intensificação da atividade desse pequeno núcleo de usuários tóxicos reincidentes.

Nossos resultados indicam que a toxicidade na plataforma não é uniformemente distribuída, mas sim concentrada. A existência desse núcleo denso de usuários altamente tóxicos sugere que a maior parte da experiência negativa nas comunidades é impulsionada por um pequeno número de atores reincidentes, validando a importância de estratégias de moderação focadas no comportamento do usuário e não apenas na análise de conteúdo de forma isolada.

4.3. Dinâmica Espacial da Toxicidade

Nos dados analisados, a toxicidade apresenta uma forte concentração em algumas comunidades. Mais precisamente, apenas quatro comunidades (*r/brasil*, *r/brasilivre*, *r/futebol* e *r/desabafos*) concentram 91,21% de todo o volume tóxico detectado, o que é em grande parte esperado, dado que essas comunidades concentram 87,75% de todo o volume de comentários (Tabela 1); por isso, analisamos a composição interna da toxicidade por perfil

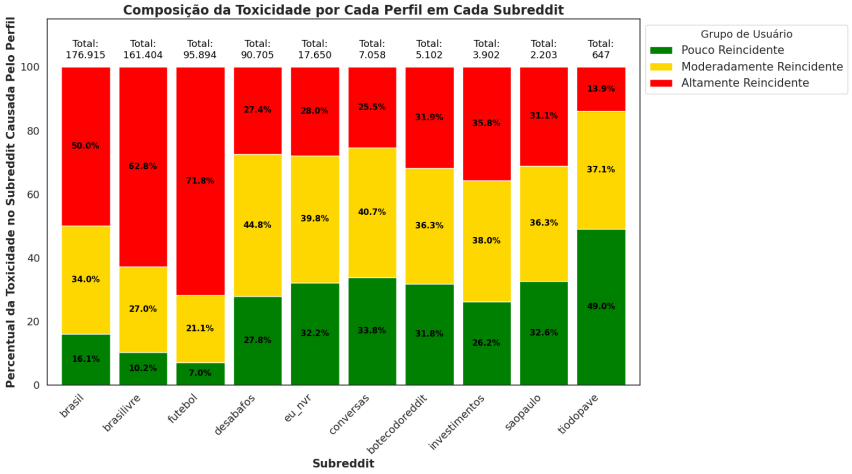

Figura 3. Quantidade de comentários tóxicos por Subreddit e porcentagem causada por cada perfil de usuário

em cada *subreddit* (Figura 3).

Nos subreddits de maior volume tóxico (*r/futebol*, *r/brasilivre*, *r/brasil*), a toxicidade é estruturalmente dominada pelo perfil *Altamente Reincidente* (barra vermelha). No *r/futebol*, por exemplo, quase 72% do conteúdo nocivo provêm desse perfil, indicando um ambiente onde a hostilidade é impulsionada pela recorrência de uma minoria engajada. Curiosamente, *r/brasilivre* é a comunidade com menos inscritos, mas figura entre as que mais geram comentários (3ª em volume total) e toxicidade (2ª em volume tóxico). Esse descompasso sugere que o volume de interações e de conteúdo tóxico não depende apenas do tamanho da audiência, mas também do nível de participação dos membros ativos.

Em contraste, comunidades como *r/tiodopave* (humor), *r/conversas* e *r/saopaulo* apresentam um perfil oposto. Nestas comunidades, a participação dos *Pouco Reincidentes* (barra verde) e *Moderadamente Reincidentes* (barra amarela) é majoritária. No *r/tiodopave*, o perfil *Pouco Reincidente* responde por quase metade da toxicidade. Essa distinção chama a atenção: ela sugere que a toxicidade em comunidades de humor ou cotidianas tende a ser acidental, reativa ou contextual (um usuário comum que fez uma piada de mau gosto ou perdeu a paciência por exemplo). Já nas comunidades políticas e de futebol, a toxicidade é sistêmica, provavelmente alimentada por usuários que utilizam a ofensa como ferramenta contínua de interação.

Complementarmente à análise por comunidade, as Figuras 4 e 5 comparam o foco de atuação não tóxica e tóxica de cada perfil. Observa-se uma forte correlação entre os ambientes de convívio e os de agressão: os usuários tendem a ser hostis nos mesmos espaços onde mantêm interações regulares. Esse padrão sugere que, no recorte analisado, a toxicidade é predominantemente endógena – fruto de atritos entre frequentadores habituais – e não apresenta um sinal consistente de deslocamentos sistemáticos para comunidades de baixo vínculo prévio (como se esperaria em episódios de *brigading*). O *r/tiodopave* ilustra o caráter episódico dessa dinâmica: apesar de ter foco tóxico e não tóxico próximo de zero para todos os perfis, 49% dos comentários tóxicos do *subreddit* são atribuídos ao perfil *Pouco Reincidente*, indicando eventos pontuais em um ambiente majoritariamente de baixa toxicidade.

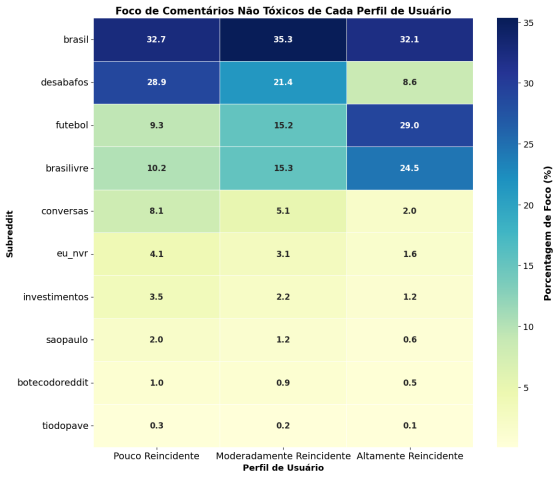

Figura 4. Foco de Comentários Não-Tóxicos

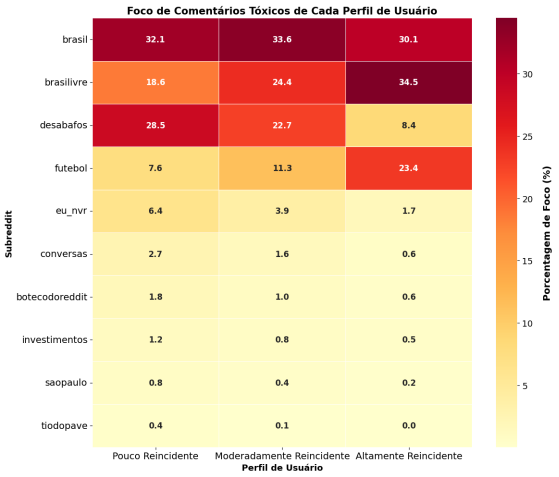

Figura 5. Foco de Comentários Tóxicos

No entanto, ao analisar o perfil dos grupos, usuários *Altamente Reincidentes* apresentam um padrão distinto. Diferente dos usuários *Pouco Reincidente* e *Moderadamente Reincidente*, que dedicam entre 21% e 29% de sua atividade não tóxica a comunidades de acolhimento como o *r/desabafos*, o núcleo hiperativo revela um desinteresse por esse tipo de interação, tendo apenas 8,6% de presença nessa comunidade em termos de comentários. Em vez disso, concentram mais de 88% de sua carga tóxica nas comunidades *r/brasil*, *r/brasilivre* e *r/futebol* (voltadas para assuntos gerais, política e esporte respectivamente). Essa distribuição evidencia que o usuário altamente tóxico prioriza ambientes de debate sobre grandes temas de interesse coletivo, em detrimento de espaços destinados à troca de experiências subjetivas e ao apoio mútuo.

Uma exceção à regra da proporcionalidade é a comunidade *r/brasilivre*, que apresenta uma dinâmica de amplificação da hostilidade a todos os perfis. Nesta comunidade, observamos que o volume de interações tóxicas é consistentemente superior à atividade regular esperada para todos os perfis. Os usuários *Pouco Reincidentes*, por exemplo, elevam sua participação de 10,2% (comportamento não tóxico) para 18,6% (comportamento tóxico). O mesmo fenômeno de amplificação ocorre com os *Moderadamente Reincidentes* (de 15,3% para 24,4%) e com os *Altamente Reincidentes* (de 24,5% para 34,5%). Essa variação positiva, variando entre 8 e 10 pontos percentuais, indica que a dinâmica interna desta comunidade tende a escalar divergências para hostilidade com maior intensidade do que nas demais comunidades, independentemente do perfil do usuário.

4.4. Comportamento dos Usuários Tóxicos

Nesta seção caracterizamos o comportamento dos perfis de usuários tóxicos considerando as métricas de P_u, A_u e D_u definidas na Seção 3.3. A análise combinada destas métricas permite compreender melhor padrões de atuação destes usuários e subsidiar políticas de moderação para mitigar a escalada de discussões tóxicas.

A Figura 6 relaciona P_u ao volume total de comentários, com o tamanho dos pontos indicando o volume tóxico, enquanto a Tabela 4 resume essas diferenças por perfil. O resultado mostra que maior reincidência não implica, necessariamente, maior proporção tóxica: o grupo *Pouco Reincidente* apresenta média mais alta de P_u (23,11%) e

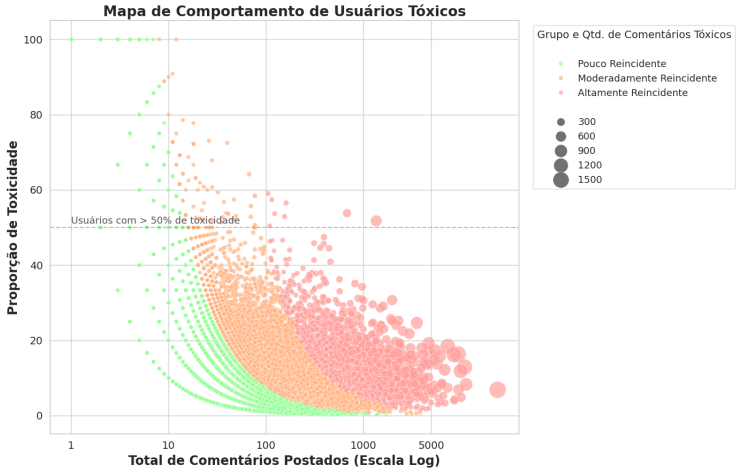

Figura 6. Relação entre Volume Total de Comentários e Proporção de Toxicidade (%). O tamanho dos pontos representa a quantidade absoluta de comentários tóxicos feitos pelo usuário.

Tabela 4. Estatísticas da Taxa de Toxicidade (%) por Grupo de Usuário.

Perfil	Usuários	Média (%)	Mediana (%)	Máximo (%)
Pouco reincidente	39.618	23,11	11,76	100,00
Moderadamente reincidente	10.053	12,99	10,71	100,00
Altamente reincidente	2.596	14,21	12,68	59,05

inclui casos próximos de 100%, compatíveis com perfis *throwaway* para ataques pontuais [Leavitt 2015]. Já o grupo *Altamente Reincidente* não atinge 100% de toxicidade (máximo de 59,05%), sugerindo que sua relevância decorre do alto volume de participação, com hostilidade diluída em muitas interações não tóxicas.

Em seguida, analisamos A_u, isto é, o número de *threads* distintas em que o usuário publicou ao menos um comentário tóxico. A partir da Tabela 5 observamos que enquanto a mediana global é baixa (2), o perfil Altamente Reincidente rompe esse padrão (mediana de 65 *threads*), indicando atuação distribuída por múltiplas discussões.

Para entender melhor o alcance tóxico dos reincidentes, investigamos se este é fruto de engajamento profundo em poucas discussões ou de pulverização através da métrica D_u. A Tabela 6 apresenta os resultados. Os valores de média e mediana são pequenos e aproximados entre os perfis de usuários.

Esse resultado é crucial para desambiguar a toxicidade do mero volume de atividade. Embora o perfil *Altamente Reincidente* tenha um alcance absoluto (A_u) muito maior devido à sua hiperatividade, o modo de operação (D_u) é semelhante ao dos usuários menos reincidentes: a estratégia de *broadcasting*. Em todos os níveis de reincidência, a norma é emitir poucos comentários tóxicos por discussão e partir para a próxima, em vez de monopolizar um tópico com longas cadeias de ofensas.

A análise conjunta das métricas P_u, A_u e D_u revela as principais características de um usuário tóxico no Reddit brasileiro. O perfil *Altamente Reincidente* não se caracteriza por uma *fúria concentrada* (alta P_u ou alta D_u), mas sim pela *onipresença*. Eles possuem baixa Proporção de Toxicidade ($P_u \approx 14\%$, em média) e baixa Densidade de Toxicidade ($D_u \approx 1,29$), comportando-se como usuários comuns na maior parte do tempo.

Tabela 5. Estatísticas de Threads Únicas com Comentários Tóxicos.

Grupo	25% (Q1)	Mediana	75% (Q3)	Máximo
Geral (Todos)	1,0	2,0	6,0	1.117
Pouco Reincidente	1,0	1,0	3,0	7
Moderadamente Reincidente	9,0	13,0	20,0	44
Altamente Reincidente	47,0	65,0	101,0	1.117

Tabela 6. Densidade de Comentários Tóxicos por Thread (Média, Desvio Padrão, Mediana e Quartil Q3).

Grupo	Média	Desvio P.	Mediana	Q3 (75%)
Geral (Todos)	1,13	0,48	1,00	1,04
Pouco Reincidente	1,09	0,36	1,00	1,00
Moderadamente Reincidente	1,25	0,79	1,09	1,23
Altamente Reincidente	1,29	0,44	1,16	1,33

No entanto, sua frequência extrema de uso amplifica essa pequena fração de hostilidade em um Alcance (A_u) massivo, permitindo atingir muitas discussões distintas com baixo investimento por interação.

4.5. Aceitação Social da Toxicidade

A Tabela 7 apresenta os resultados da regressão do *score* de validação social ordenados pelo coeficiente β. De forma geral, os resultados indicam que a toxicidade não é majoritariamente punida: em quase todas as comunidades, β é positivo e estatisticamente significativo, sugerindo recompensa social para comentários *Tóxicos*. Esse padrão é consistente com evidências da literatura de que conteúdos associados a indignação e emoção negativa tendem a gerar maior engajamento [Crockett 2017, Brady et al. 2017].

O maior incentivo à toxicidade ocorre no *subreddit r/brasil* ($\beta = 3.019$), indicando que comentários *Tóxicos* recebem, em média, cerca de três *upvotes* a mais do que comentários *Não Tóxicos*. Esta comunidade trata de temas gerais, com um grande engajamento dos usuários e o comportamento tóxico pode ser usado como estratégia para atrair atenção durante as discussões, principalmente em tópicos polarizados. Em comunidades de suporte e conversação (*r/desabafos* e *r/conversas*), o efeito também é positivo, o que pode refletir validação de ataques direcionados a possíveis *antagonistas* narrados em relatos pessoais. Já em comunidades de humor e memes (*r/eu_nvr* e *r/botecodoreddit*), coeficientes elevados podem estar associados a linguagem informal/intensa que é socialmente aceita, embora sinalizada como tóxica pelo classificador.

Em contraste, *r/brasilivre* e *r/futebol* exibem ganhos menores ($\beta < 0.6$), possivelmente por efeito de saturação em ambientes onde a hostilidade é mais frequente e, portanto, menos distintiva. A única exceção é *r/tiodopave*, onde o efeito é negativo e não significativo, sugerindo que, em um nicho de humor leve, a agressividade tende a violar a expectativa de interação.

5. Conclusão

Este trabalho caracterizou a dinâmica da toxicidade no ecossistema brasileiro do Reddit, processando mais de 6,5 milhões de comentários. A partir da definição de perfis de re-

Tabela 7. Impacto da Toxicidade no Score por Subreddit (Regressão OLS).

Subreddit	**Coeficiente (β)**	**Erro Padrão**	**P-valor**
r/brasil	3,019	0,091	$< 0,001^{***}$
r/eu_nvr	2,183	0,269	$< 0,001^{***}$
r/botecodoreddit	1,758	0,215	$< 0,001^{***}$
r/desabafos	1,381	0,027	$< 0,001^{***}$
r/saopaulo	1,053	0,277	$< 0,001^{***}$
r/conversas	1,019	0,085	$< 0,001^{***}$
r/investimentos	0,682	0,287	$0,017^{*}$
r/brasilivre	0,587	0,030	$< 0,001^{***}$
r/futebol	0,194	0,062	$0,002^{**}$
r/tiodopave	-0,321	0,287	$0,265$ (n.s.)
Legenda: *** p<0,001; ** p<0,01; * p<0,05; n.s. não significativo.			

incidência de toxicidade, identificamos uma minoria hiperativa (cerca de 5%) que gera mais de 50% de todo o volume de ofensas. Esse grupo se destaca pela onipresença: alto alcance com baixa densidade por *thread*, ampliando o desafio de moderação contextual.

Adicionalmente, a intensidade de produção do conteúdo tóxico tende a seguir eventos reais (com picos alinhados ao calendário político e esportivo) e é alimentada por incentivos da própria comunidade. Nossos resultados mostram que existe aceitação social do conteúdo tóxico compartilhado, indicando que, em muitos *subreddits*, comentários tóxicos recebem, em média, *score* superior ao de comentários não tóxicos, sugerindo um mecanismo de validação social associado ao engajamento, exceto em ambientes onde a agressividade já se encontra saturada ou normalizada.

Como limitações, parte do treino do modelo foi ampliada com rótulos por consenso entre LLMs sem validação especialista individual, embora o teste tenha permanecido humano. Além disso, o F1 de 0,65 para a classe Tóxico pode introduzir erros em casos ambíguos. A generalização é limitada pelo recorte de 2022 e pelos dez maiores *subreddits*; trabalhos futuros devem replicar a análise em outros períodos e comunidades.

Em última análise, este trabalho reforça que a moderação eficaz deve transcender a remoção de mensagens individuais, focando na identificação e mitigação dos padrões comportamentais dos usuários reincidentes e nos incentivos sistêmicos que sustentam a toxicidade na plataforma.[2]

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

Referências

Almeida, T. S. et al. (2024). Sabiá-2: A new generation of portuguese large language models.

[2]Este trabalho foi conduzido em conformidade com as diretrizes éticas do CNS (Res. 510/2016) no que concerne ao processo de anotação manual. Ferramentas de IA generativa foram utilizadas como auxílio na escrita e geração de código. Disponibilidade de código e dados mediante solicitação aos autores.

Baumgartner, J. et al. (2020). The pushshift reddit dataset. In *Proc. of the Int. AAAI Conf. on Web and Social Media (ICWSM)*, volume 14, pages 830–839.

Brady, W. J. et al. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

CNN Brasil (2022a). Monark é desligado do flow podcast após defender existência de partido nazista. Acesso em: dez. 2025.

CNN Brasil (2022b). Olavo de carvalho morre aos 74 anos nos estados unidos. Acesso em: dez. 2025.

Costa, P. B. et al. (2023). BERTabaporu: Assessing a genre-specific language model for Portuguese NLP. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 217–223, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1(11):769–771.

de Pelle, R. and Moreira, V. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, pages 510–519, Porto Alegre, RS, Brasil. SBC.

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., and Joty, S. (2024). Data augmentation using large language models: Data perspectives, learning paradigms and challenges.

Duggan, M. (2017). Online harassment 2017. Technical report, Pew Research Center.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In Roberts, S. T., Tetreault, J., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

G1 (2022a). Bolsonaro reúne embaixadores para repetir sem provas suspeitas já esclarecidas sobre urnas. Acesso em: dez. 2025.

G1 (2022b). Bolsonaro usa 7 de setembro para fazer campanha, puxa coro machista e reúne multidões em atos com faixas antidemocráticas. Acesso em: dez. 2025.

G1 (2022c). Em áudios, arthur do val disse que ucranianas são 'fáceis, porque são pobres'. Acesso em: dez. 2025.

G1 (2022d). Lula vence o segundo turno e volta para o terceiro mandato de presidente. Acesso em: dez. 2025.

Globo Esporte (2022). Brasil perde para a croácia nos pênaltis e dá adeus à copa do mundo. Acesso em: dez. 2025.

Jiang, A. Q. et al. (2023). Mistral 7b.

Kumar, D. et al. (2023). Understanding the behaviors of toxic accounts on reddit. In *Proc. of the ACM Web Conf. 2023 (WWW '23)*, pages 2797–2807. ACM.

Leavitt, A. (2015). "this is a throwaway account": Temporary technical identities and perceptions of anonymity in a massive online community. In *Proceedings of the 18th ACM*

Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15, page 317–327, New York, NY, USA. Association for Computing Machinery.

Leite, J. A. et al. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*.

Lima, L. H. Q., Pagano, A. S., and da Silva, A. P. C. (2024). Toxic content detection in online social networks: A new dataset from brazilian reddit communities. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 472–482, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Mathew, B. et al. (2020). Hate begets hate: A temporal study of hate speech. *Proc. ACM Hum.-Comput. Interact. (CSCW)*, 4(CSCW2):1–25.

Olteanu, A. et al. (2018). The effect of extremist violence on hateful speech online.

Park, J. S. et al. (2022). Measuring the prevalence of anti-social behavior in online communities.

Piorino, G., Machado, M. A. d. A., Lima, L. H. Q., Pagano, A., and Silva, A. P. C. d. (2026). Diálogos tóxicos: Gatilhos e padrões de interação no Reddit brasileiro. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 581–590, Salvador, Brazil. Association for Computational Linguistics.

Ribeiro, M. H. et al. (2018). Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Thomas, K. et al. (2021). Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 361–380.

Touvron, H. et al. (2023). Llama 2: Open foundation and fine-tuned chat models.

Uol (2022a). 1º de maio é marcado por atos pró-lula e pró-bolsonaro. Acesso em: dez. 2025.

Uol (2022b). O que se sabe sobre os protestos que bloqueiam rodovias. Acesso em: dez. 2025.

Uol (2022c). Prf sobre morte de homem negro por policiais do órgão em se: 'indignação'. Acesso em: dez. 2025.

Vargas, F. et al. (2022). Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.

Vogels, E. A. (2021). The state of online harassment.

Xia, Y., Taylor, J., Nabeshima, T., and Mutton, P. (2020). Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–23.