O Fenômeno da Toxicidade em Jogos Digitais: Uma Caracterização da Comunidade Brasileira na Steam

Title: The Toxicity Phenomenon in Digital Games: A Characterization of the Brazilian Community on Steam

Paula Gibrim[1], Marcus Ribeiro[1], Daniel Barbosa[1], Philipe Melo[1], Julio C. S. Reis[1]

[1]Universidade Federal de Viçosa (UFV) – MG – Brasil

{paula.gibrim, marcus.guerra, danielmendes, philipe.freitas, jreis}@ufv.br

Abstract. Introduction: *Steam is one of the largest digital game distribution platforms in the world, hosting millions of textual reviews, including instances of toxicity that evade automatic moderation.* ***Objective:*** *This work characterizes toxicity in reviews produced by the Brazilian gaming community and examines behavioral differences between toxic and non-toxic users.* ***Methodology:*** *A dataset of over* 600,000 *reviews in Portuguese was analyzed using two automatic detection models: Perspective API and Detoxify.* ***Results:*** *Toxic discourse is marked by offensive vocabulary and graphic intensification, but often coexists with positive recommendations and gameplay descriptions, challenging automatic classifiers. Competitive games have the highest proportions of toxicity, and toxic users are more engaged on the platform, yet are not penalized with significantly higher probability than other users.* ***Warning! This work and the referenced data contain examples of offensive and hateful language.***
Keywords *Toxicity, Hate Speech, Steam, Digital Games, Content Moderation, Natural Language Processing (NLP).*

Resumo. Introdução: *A Steam é uma das maiores plataformas de distribuição de jogos digitais do mundo, reunindo milhões de avaliações textuais, incluindo manifestações de toxicidade que escapam à moderação automática.* ***Objetivo:*** *Este trabalho caracteriza a toxicidade nas avaliações produzidas pela comunidade brasileira e as diferenças comportamentais entre usuários tóxicos e não tóxicos.* ***Metodologia****: Um conjunto de dados de mais de 600 mil avaliações em português foi analisado com dois modelos de detecção automática: Perspective API e Detoxify.* ***Resultados:*** *O discurso tóxico é marcado por vocabulário ofensivo e intensificação gráfica, mas frequentemente coexiste com recomendações positivas e descrições de gameplay, desafiando classificadores automáticos. Jogos competitivos concentram as maiores proporções de toxicidade, e usuários tóxicos são mais engajados na plataforma, mas não são punidos com probabilidade significativamente maior que os demais.* ***Atenção! Este trabalho e os dados referenciados contêm exemplos de linguagem ofensiva e odiosa.***
Palavras-Chave *Toxicidade, Discurso de Ódio, Steam, Jogos Digitais, Moderação de Conteúdo, Processamento de Linguagem Natural (PLN).*

1. Introdução

A Steam consolidou-se como um dos maiores ecossistemas digitais de distribuição de jogos do mundo [Wijman 2018], reunindo milhões de usuários ativos e um volume

massivo de avaliações textuais associadas aos jogos disponíveis [Backlino Team 2026]. Esse ambiente configura-se como um dos maiores acervos de jogos digitais, contendo não apenas opiniões sobre produtos, mas também manifestações sociais, culturais e comportamentais dos usuários em larga escala [Plunkett 2013, xPaw 2026].

Entre esses fenômenos, destaca-se a toxicidade, entendida aqui como o conjunto de manifestações linguísticas que expressam agressividade, hostilidade ou linguagem ofensiva capaz de degradar a experiência de outros usuários [Lees et al. 2022]. No contexto de jogos digitais, esse fenômeno apresenta características particulares, uma vez que termos potencialmente agressivos podem ser parte de descrições mecânicas da *gameplay* [Sun et al. 2025], e expressões ofensivas são frequentemente empregadas como recurso humorístico ou intensificador discursivo dentro desse ecossistema [Sun et al. 2025], o que impõe desafios semânticos a sistemas automáticos de moderação.

O Brasil ocupa uma posição de destaque nesse ecossistema, sendo um dos dez maiores mercados de jogos do mundo [NewZoo 2024], e a comunidade brasileira é reconhecida por sua expressividade e alto volume de interações na plataforma [Valve Corporation 2026b]. Além disso, o português brasileiro apresenta características lexicais e pragmáticas próprias – como o uso de gírias, palavrões e recursos humorísticos com alta carga semântica ambígua – que tornam a detecção automática de toxicidade particularmente desafiadora [Moreira et al. 2026]. Estudar a toxicidade no contexto brasileiro é, portanto, tanto uma contribuição linguística quanto uma demanda prática para o desenvolvimento de sistemas de moderação mais inclusivos e culturalmente situados.

Diante desse cenário, este trabalho propõe uma caracterização da toxicidade na Steam, através das avaliações de títulos produzidas pela comunidade *gamer* brasileira. A investigação é orientada pelas seguintes questões de pesquisa: **QP1**: *O que caracteriza uma avaliação tóxica e qual é sua composição lexical?* **QP2**: *Em quais contextos — jogos e marcadores populares — a toxicidade ocorre com maior frequência?* **QP3**: *Quais são as diferenças comportamentais entre usuários tóxicos e não tóxicos na plataforma?*

Para responder essas questões de pesquisa, realizou-se uma coleta e filtragem de avaliações em português publicadas em títulos da Loja Steam[1], resultando em um conjunto de dados de mais de 600 mil textos. Em seguida, cada avaliação foi classificada com base em dois modelos de detecção automática de toxicidade — a Perspective API [Google 2026] e o Detoxify [Unitary 2024]. Complementarmente, aplicou-se modelagem temática para identificar os tópicos recorrentes nas avaliações tóxicas e não tóxicas. Por fim, realizou-se uma análise do perfil comportamental dos usuários, cruzando métricas de atividade na plataforma com os rótulos de toxicidade atribuídos.

Em resumo, os resultados revelam que, do ponto de vista linguístico, o discurso tóxico é fortemente marcado por vocabulário ofensivo e afeto negativo, mas emerge também em contextos não hostis — como descrições de *gameplay* ou expressões de humor —, reforçando o desafio para classificadores automáticos. De uma perspectiva contextual, jogos competitivos e de conteúdo violento concentram as maiores proporções de toxicidade. Por fim, usuários tóxicos tendem a ser mais ativos na plataforma em múltiplas dimensões, mas não são punidos com probabilidade significativamente maior que os demais usuários.

[1] `https://store.steampowered.com/`

2. Trabalhos Relacionados

A toxicidade em ambientes digitais tem sido amplamente investigada na literatura, com especial atenção a plataformas de redes sociais [Waseem e Hovy 2016, Lima et al. 2020]. Nesses contextos, ferramentas automáticas de detecção — como a Perspective API [Google 2026] e o Detoxify [Unitary 2024] — tornaram-se instrumentos centrais para identificar discurso ofensivo em larga escala [Lees et al. 2022], embora avaliações críticas apontem limitações relevantes nesses sistemas, como a incapacidade de considerar o contexto conversacional e a presença de vieses contra dialetos minoritários devido à dependência excessiva de pistas lexicais [Pavlopoulos et al. 2020, Gehman et al. 2020].

A literatura tem refletido essa preocupação de forma crescente também no contexto de jogos digitais. Em uma revisão de escopo, [Gibrim et al. 2025] levantaram 49 estudos sobre toxicidade e discurso de ódio em jogos digitais, revelando que a área foca primariamente na dinâmica de títulos específicos — como DOTA 2 [Blackburn e Kwak 2014] e *League of Legends* [Kou 2020] —, com ênfase na comunicação *in-game*. Plataformas de distribuição e suas comunidades de avaliação permanecem, portanto, sub-exploradas.

Do ponto de vista do consumidor brasileiro, [Vasconcelos et al. 2025] investigaram o perfil de jogadores pagantes no Brasil, identificando narrativa, jogabilidade e gráficos como principais fatores de decisão de compra, enquanto [Peron et al. 2025] apresentam um ensaio teórico que desmistifica associações entre videogames, vício e violência, destacando seus benefícios cognitivos e emocionais. Nenhum desses trabalhos, contudo, é conduzido na Steam ou examina o comportamento tóxico nas interações públicas da comunidade brasileira. O presente trabalho endereça essa lacuna ao deslocar o olhar do interior dos jogos para as avaliações públicas dessa plataforma, caracterizando o discurso tóxico produzido pela comunidade brasileira nesse ambiente.

3. Metodologia

A metodologia foi estruturada em quatro etapas principais, ilustradas na Figura 1: *(i)* coleta e filtragem dos dados (Seção 3.1); *(ii)* aplicação e comparação das métricas de detecção selecionadas (Seção 3.2); *(iii)* rotulação de avaliações tóxicas (Seção 3.3); e *(iv)* aplicação de técnicas auxiliares para as análises (Seção 3.4).

3.1. Coleta e Filtragem dos Dados

Os dados foram coletados por meio da Steam API [Valve Corporation 2026c] entre setembro de 2024 e abril de 2025, seguindo um processo sequencial. Inicialmente, foram coletados 169.141 títulos disponíveis na loja, acompanhados de metadados como nome, marcadores populares (*tags*)[2] e avaliações de usuários. Em seguida, a partir desses títulos, foram coletados 89.713.199 comentários, incluindo texto, data de publicação e dados públicos de identificação do usuário. Por fim, a partir dos comentários, foi extraída uma base de 21.513.415 usuários únicos, contendo informações como número de jogos na biblioteca, nível na plataforma e localização geográfica autodeclarada.

3.2. Aplicação de Métricas de Detecção e Filtragem Adicional

Para identificar conteúdo tóxico, aplicamos, inicialmente, a Perspective API como métrica de detecção de toxicidade, em razão de sua ampla aceitação na literatura

[2]Marcadores populares são rótulos atribuídos pela própria comunidade para descrever características dos jogos, como estilo, mecânicas e temas. Mais informações: `https://store.steampowered.com/tag/`.

Figura 1. Etapas metodológicas adotadas no estudo.

[Lees et al. 2022]. A fim de tornar os resultados mais robustos, incorporou-se também o Detoxify como métrica complementar – reforçado pelo recente anúncio de descontinuação da primeira [Vadim Berman 2026]. Em ambas as métricas, cada comentário recebeu uma pontuação contínua no intervalo $[0, 1]$, além da indicação de idioma do texto fornecida pela Perspective API. A consistência entre as ferramentas foi avaliada pelas correlações entre suas pontuações sobre o conjunto de dados completo, revelando concordância moderada: coeficiente de Pearson de 0,7462 e de Spearman de 0,5755.

Realizou-se, então, uma etapa adicional de filtragem, composta por: seleção de perfis localizados no Brasil, retenção exclusiva das avaliações redigidas em português e remoção de usuários e jogos que, após esses filtros, não apresentavam nenhum comentário associado. O conjunto de dados final totaliza 605.938 avaliações de 246.311 usuários distintos, distribuídas entre 17.666 títulos que apresentam ao menos uma análise no idioma.

3.3. Rotulação de Mensagens Tóxicas

Para calibrar os limiares para classificação de mensagens tóxicas, conduzimos um processo de anotação humana com três avaliadores independentes. As pontuações de cada modelo foram segmentadas em *bins* de amplitude 0,1 — i.e., [0,0, 0,1), [0,1, 0,2), …, [0,9, 1,0] — e de cada intervalo foram amostradas avaliações para compor um conjunto de 200 instâncias por métrica. Os anotadores classificaram cada mensagem como TÓXICA ou NÃO TÓXICA, sem acesso ao jogo de origem nem às pontuações dos modelos, a fim de evitar viés de ancoragem.

A concordância entre anotadores foi mensurada pelo *Fleiss' Kappa* [Fleiss 1971], calculado globalmente sobre todas as amostras. O Detoxify alcançou $\kappa = 0{,}8533$ (concordância quase perfeita), ao passo que a Perspective API obteve $\kappa = 0{,}7533$ (concordância substancial), conforme a escala de [Landis e Koch 1977]. Esses resultados indicam que o Detoxify produz pontuações mais alinhadas ao julgamento humano no contexto investigado. Com base nessa avaliação, definiram-se limiares distintos para cada ferramenta: avaliações com pontuação $\geq 0{,}9$ no Detoxify **ou** $\geq 0{,}7$ na Perspective API foram classificadas como TÓXICAS, resultando em 8.640 ocorrências distribuídas entre 7.657 perfis únicos. Para os fins deste trabalho, denomina-se USUÁRIO TÓXICO todo perfil que produziu ao menos uma avaliação classificada como tóxica.

A Figura 2 apresenta a função de distribuição acumulada (CDF) das pontuações de ambos os modelos. O Detoxify cresce mais rapidamente nas faixas baixas de toxicidade, atribuindo pontuações consideravelmente mais elevadas do que a Perspective API para a maioria dos comentários tóxicos — diferença que reflete distinções de calibração entre as ferramentas. As linhas verticais tracejadas marcam os limiares adotados.

O diagrama de Venn da Figura 3 detalha a sobreposição entre os conjuntos sinalizados por cada modelo. Das avaliações identificadas por alguma das ferramentas, apenas 1.872 (21,7%) são detectadas por ambos; 911 (10,5%) são sinalizados

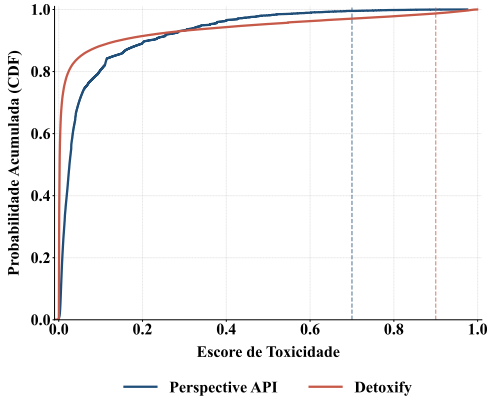


Figura 2. CDF das pontuações dos modelos.

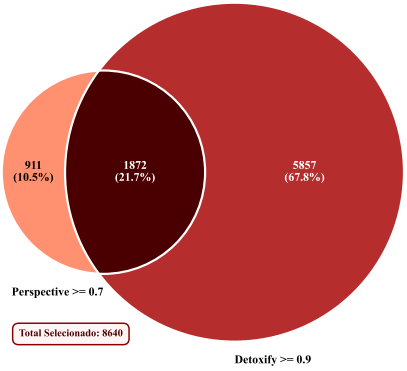


Figura 3. Diagrama de Venn das avaliações.

exclusivamente pela Perspective API e 5.857 (67,8%) exclusivamente pelo Detoxify. Essa assimetria evidencia que os modelos capturam perspectivas distintas, reforçando a adoção da união como critério de rotulação, de modo a maximizar a cobertura do fenômeno.

3.4. Técnicas Complementares

Para identificar e quantificar a relevância dos termos presentes nas avaliações, empregou-se *Term Frequency-Inverse Document Frequency* (TF-IDF) [Salton e Buckley 1988]. O método foi aplicado separadamente sobre o subconjunto tóxico e o conjunto de avaliações regulares, com o objetivo de isolar o vocabulário que caracteriza e diferencia o discurso tóxico na plataforma. Ao ponderar a frequência local de cada termo contra a sua ocorrência no conjunto de dados global, o TF-IDF penaliza palavras de uso generalizado e destaca aquelas com maior poder discriminativo em cada classe. A análise dos resultados obtidos é apresentada na Seção 4.1.

Para a modelagem temática das avaliações tóxicas, aplicou-se o BERTopic [Grootendorst 2022] com *embeddings* gerados pelo modelo multilíngue `paraphrase-multilingual-MiniLM-L12-v2` [Reimers e Gurevych 2019], redução dimensional via UMAP [McInnes et al. 2018] e clusterização via HDBSCAN [McInnes e Healy 2017]. Para garantir reprodutibilidade, fixou-se uma semente global (*seed* = 3) em todas as etapas estocásticas do pipeline — geração de *embeddings*, redução dimensional e inicialização dos geradores de números aleatórios.

A rotulação dos tópicos resultantes foi realizada com auxílio do modelo de linguagem *Claude Sonnet 4.6 (Anthropic)*. Para cada tópico, forneceram-se ao modelo as palavras-chave geradas automaticamente pelo BERTopic e uma amostra representativa das avaliações atribuídas ao respectivo *cluster*. O modelo foi, então, instruído a propor um rótulo sintético que capturasse o fenômeno semântico central do grupo. Os resultados dessa etapa são discutidos na Seção 4.1.

4. Resultados

4.1. QP1 – Caracterização de uma avaliação tóxica

A análise dos escores de TF-IDF (Tabela 1) permite delinear um perfil linguístico característico das avaliações classificadas como tóxicas na plataforma, distinguido-as, de forma consistente, das avaliações regulares. Foi observado, também, uma recorrência de sequências do símbolo ♥ no corpus tóxico, evidenciando uma tentativa de moderação automática da Steam [Valve Corporation 2026a] — na qual palavrões são substituídos

por símbolos — mas que se mostra ainda superficial, uma vez que o conteúdo ofensivo permanece identificável pelo contexto.

Do ponto de vista da composição lexical, as avaliações tóxicas são marcadas pela predominância de termos ofensivos e de baixo calão, com alta frequência relativa em seu grupo e ocorrência mínima nos textos regulares. Termos como *lixo*, *caralho*, *porcaria* e *porrada* apresentam escores TF-IDF cerca de 40 vezes mais elevados nas avaliações tóxicas do que nas regulares, evidenciando seu papel como marcadores quase exclusivos de discurso tóxico na plataforma. Casos extremos, como *fodase* e *maldito*, chegam a escores 100 vezes superiores, configurando fortes indicadores de toxicidade.

Tabela 1. Valores de TF-IDF para termos de tendência regular e tóxica.

Tendência Regular			Tendência Tóxica		
Termo	**TF-IDF Tóx.**	**TF-IDF Neu.**	**Termo**	**TF-IDF Tóx.**	**TF-IDF Neu.**
bom	0.034269	0.123506	lixo	0.046950	0.001007
bem	0.006324	0.029516	caralho	0.035835	0.001176
divertido	0.005045	0.025076	sexo	0.020197	0.000377
recomendo	0.012031	0.027735	matar	0.017742	0.001092
excelente	0.000968	0.015173	porcaria	0.015644	0.000198
história	0.002446	0.016404	porrada	0.014296	0.000257
melhor	0.010770	0.024339	pau	0.011858	0.000186
ótimo	0.002230	0.015724	bunda	0.009274	0.000077
top	0.002536	0.014075	fodase	0.008887	0.000016
otimo	0.002151	0.012382	tiro	0.009017	0.001339
boa	0.002891	0.012879	odeio	0.008231	0.000557
legal	0.006764	0.016045	mata	0.007412	0.000534
jogabilidade	0.002775	0.012038	caguei	0.007119	0.000266
pena	0.004415	0.013496	bomba	0.006864	0.000365
vale	0.004712	0.013733	maldito	0.006477	0.000053

Outra característica notável é o caráter de expressão emocional negativa do vocabulário tóxico. Enquanto as avaliações regulares têm maior escore em termos relacionados a atributos do jogo (e.g., *divertido*, *história*, *jogabilidade*) ou recomendações (e.g., *recomendo*, *bom*, *excelente*), as avaliações tóxicas empregam uma linguagem orientada ao afeto negativo e à agressividade, com verbos de ação violenta (e.g., *matar*) e substantivos de cunho depreciativo (e.g., *lixo*, *porcaria*). Isso sugere que a avaliação tóxica não apenas julga negativamente o produto, mas dirige sua carga semântica à provocação, ao insulto ou à expressão de raiva – em detrimento da crítica articulada.

Essa combinação entre vocabulário ofensivo e recursos gráficos de intensificação é exemplificada em dois casos representativos. No comentário *"Jogo lixo trava para caralho e ainda paguei 90 conto na* ♥♥♥♥♥ *que um jogo lixo lixo vai toma no no* ♥♥ *jogo lixo do caralho"* (avaliação do jogo **Automobilista 2**), a repetição exaustiva de *lixo* e *caralho* evidencia o uso de termos ofensivos como estratégia de ênfase. No segundo, *"JOGO LIXOOOOOOOOOO VERGONHOSO NEM PAGANDO DA PRA JOGAR ISSO"* (**Astra Nova**), o usuário combina caixa alta e alongamento vocálico (*LIXOOOOOO*) — recurso típico do discurso digital agressivo para simular ênfase prosódica na escrita.

No entanto, a presença de linguagem tóxica não implica necessariamente uma avaliação negativa do produto: 70,3% das avaliações tóxicas ainda recomendam o jogo analisado. Nesses casos, o vocabulário se manifesta de forma desvinculada do julgamento crítico, assumindo funções expressivas distintas e impondo desafios severos à classificação automática em NLP.

Esse desafio semântico se manifesta principalmente por meio de dois fenômenos. O primeiro é a contaminação do vocabulário pelas próprias mecânicas do jogo. Termos que indicariam ameaça ou violência em outros contextos muitas vezes descrevem apenas a *gameplay*, como em *"Muito legal matar o amiguinho com a bombinha"* (**Magicka**) e *"jogo muito brabo, pato pato tiro tiro porrada porrada, recomendo"* (**Duck Game**). Nesses casos, os termos provavelmente não carregam intenção de agredir outros usuários, mas refletem ações literais dentro do ambiente virtual. Já o segundo é a toxicidade performática. Em muitos casos, usuários se apropriam de termos ofensivos não para atacar um alvo específico, mas como um registro coloquial extremo, expressando humor absurdo ou hiperbolizando emoções dentro de uma subcultura *gamer*. Isso fica evidente em trechos como *"chorei igual uma vagabunda"* (**Tell Me Why**), visto que, embora os classificadores automáticos acionem alertas devido à carga semântica isolada do termo, a intenção do usuário é apenas demonstrar entusiasmo de forma informal, sem a intenção de atacar qualquer interlocutor.

Bertopic. Os resultados do BERTopic estão sumarizados na Tabela 2. Os dois maiores *clusters* agrupam formas distintas de crítica agressiva. Crítica Genérica ($n = 5.107$) reúne desde *spam* repetitivo e *trollagem* — incluindo avaliações em que nomes ou mecânicas de jogos como *Celeste* e *Cookie Clicker* são repetidamente utilizados como forma de provocação — até xingamentos soltos sem alvo definido. Frustração Técnica ($n = 1.431$), por sua vez, centrada em reclamações sobre qualidade, custo-benefício e tempo desperdiçado. A separação entre os dois evidencia que a toxicidade motivada por insatisfação com o produto não é uniforme, podendo emergir como uma crítica difusa e emocional ou uma crítica mais articulada, ainda que agressiva.

Tabela 2. Resumo dos tópicos gerados pelo BERTopic, com volume, rótulo sintetizado e palavras-chave representativas.

#Avaliações	Tópico	Palavras-chave
5107	Crítica Genérica	clica, bolas, odeio, bom, caralho, celeste, mata, matar, vai, vc
1431	Frustração Técnica	lixo, porcaria, vai, vc, tempo, horas, ter, dinheiro, desse, nao
361	Conteúdo Adulto	sexo, penis, melhor, sex, anal, bom, faltou, gostoso, drogas, ter
317	Vocabulário de Gameplay	bomba, tiro, nazistas, nazista, porrada, matar, bom, bala, nazi, racista
207	Ódio Explícito	pssimo, odeio, ruim, caralho, mal, otimizado, pior, horrvel, pattico, sucks
97	Discurso Homofóbico	gay, brasil, china, sinto, sexo, mae, garoto, porque, local, vergonho
61	Conteúdo Viral	sexo, fao, mulheres, bonito, porque, amigas, sexy, fazer, fiz, nua
56	Discurso Xenofóbico	macaco, argentino, macacos, argentinos, dardo, vai, xingar, pipa, dinossauros, passar
49	Crítica Localizada	brasil, brasileiro, brasileiros, aumentou, buga, servidor, empresa, servidores, tomar, pirateado
44	Gíria Sexual	pau, lendo, ta, duro, pega, bom, correndo, curto, escrevi, legal

Um tópico curioso é Vocabulário de Gameplay ($n = 317$), cujas palavras-chave — *bomba*, *tiro*, *nazistas*, *porrada*, *matar* — correspondem, em sua maioria, a descrições literais de mecânicas de jogos de ação. Os documentos representativos confirmam que a maioria das avaliações nesse *cluster* são positivas, como *"tiro, desarma bomba, morre, rage cabo"* (**Counter-Strike 2**) e *"Excelente jogo. Campanha extensa com história intrigante. E é muito divertido matar os nazistas! Vale cada centavo."* (**Wolfenstein: The New Order**). Trata-se, portanto, de falsos positivos nos quais o classificador reage ao vocabulário sem considerar o contexto dos gêneros dos jogos.

Identificam-se, ainda, alguns tópicos de conteúdo sexual semanticamente distintos, evidenciando que um vocabulário semelhante pode emergir de motivações opostas. O tópico Conteúdo Adulto ($n = 361$) agrupa vocabulário explícito característico

de jogos de nicho, onde esse registro é parte do discurso esperado. O tópico Conteúdo Viral ($n = 61$) isola um *copypaste* de virilidade exagerada — uma narrativa humorística em que o narrador afirma repetidamente fazer "muito sexo com mulheres" e compara explicitamente essa experiência ao jogo em questão — postada sistematicamente como avaliação de títulos não relacionados ao tema. Discurso Homofóbico ($n = 97$), por sua vez, agrupa usos do termo *gay* como xingamento ou marcador informal, com ocorrências que variam do insulto direto à apropriação coloquial ambígua.

Complementarmente, Ódio Explícito ($n = 207$), com linguagem de repúdio direto ao jogo ou ao desenvolvedor; Discurso Xenofóbico ($n = 56$) reúne dois padrões: avaliações como *"O macaco de gelo coça a bunda"* (**Bloons TD 6**) descrevem personagens do jogo, enquanto *"um bom tiroteio mas cuidado com os argentinos aqueles vagabundos a xenofobia é tão ruim que provavelmente foi inventada por um argentino fdp"* (**Counter-Strike 2**) expressa hostilidade genuína a jogadores estrangeiros em contexto competitivo; Crítica Localizada ($n = 49$) reúne reclamações sobre servidores e sobre a experiência da comunidade brasileira na plataforma; e Gíria Sexual ($n = 44$) apresenta um vocabulário sexual informal e desvinculado de ataques diretos.

Em conjunto, os tópicos revelam que a toxicidade observada na plataforma não é uniforme pois emerge de contextos distintos — competitivos, temáticos, culturais — e assume formas que vão do insulto deliberado ao falso positivo estrutural.

4.2. QP2 - Onde a toxicidade acontece

A fim de identificar os contextos em que a toxicidade se manifesta com maior frequência, a análise foi conduzida em relação às *tags* e jogos individuais — duas dimensões complementares. Em ambos os casos, adotou-se como critério de representatividade estatística o percentil de 90% (P90) por volume de avaliações, considerando apenas as entidades que ultrapassam esse limiar.

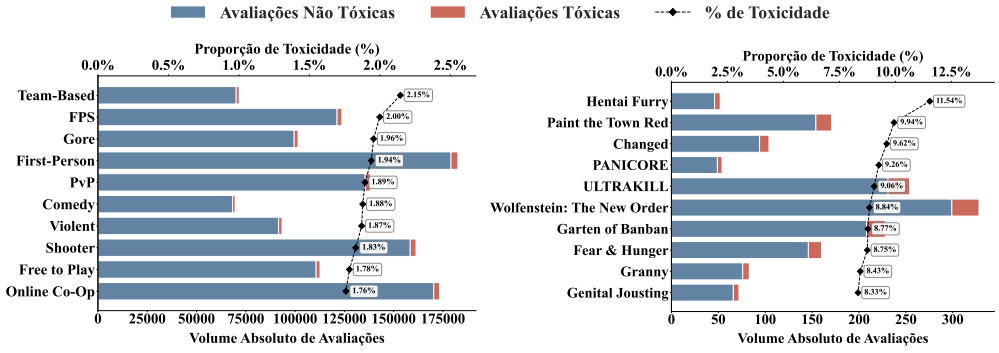
Entre as *tags* com maior proporção de toxicidade (Figura 4a), distinguem-se três padrões. O primeiro é a competição direta entre jogadores, representada por Team-Based (2,15%), PvP (1,89%) e Online Co-Op (1,76%). O exemplo (1) ilustra como a euforia de uma jogada é rapidamente convertida em uma agressividade dirigida aos adversários.

(1) *"vai passa a bola aqui no meio, belo passe, olha esse chute,* ♥♥♥♥ *que pariu que golaço do caralho,* ♥♥♥♥ ♥♥♥♥♥*, eu sou muito* ♥♥♥♥*, todos voces, abaixem e me chupem, seus lixos, eu sou literalmente Jesus Cristo desse jogo, seus pobres imundos".* (**Pro Soccer Online**)

O segundo é o conteúdo temático violento — Gore (1,96%), Shooter (1,83%) e Violent (1,87%) —, cujas comunidades tendem a fazer uso de um padrão linguístico mais agressivo. FPS (2,00%) e First-Person (1,94%) aparecem entre os marcadores mais tóxicos por serem mecânicas frequentemente associadas a ambos os contextos — competição e violência. O exemplo (2) ilustra essa sobreposição, combinando o vocabulário violento com a frustração dirigida ao desenvolvedor.

(2) *"Parabéns, conseguiram estragar um ótimo jogo, empresa incompetente do caralho. ERA SÓ FAZER O BÁSICO SEUS FILHO DA P(_)TA! NEM A* ♥♥♥♥♥ *DO cl_righthand 0 TEM NESSE LIXO AGORA,Só sabe trazer atualização que aumenta venda de skin, mas o jogo e o anti-cheat vcs deixam cagados... Vermes imundos!"* (**Counter-Strike 2**)

Por fim, o mais contraintuitivo é Comedy (1,88%), sugerindo que contextos humorísticos favorecem linguagem mais agressiva, onde os limites entre o jocoso e o ofensivo tendem a se tornar difusos. O comentário *"zumbis nazistas, aliens, poder anal, entrei no cool*



(a) Top 10 *tags* **(b) Top 10 jogos**

Figura 4. Proporção de toxicidade (corte P90) agrupada por (a) *tags* e (b) jogos.

de um gay, meus peidos têm poderes, isto é South Park. Amei!" (**South Park™: The Stick of Truth™**) é uma recomendação entusiasmada, mas emprega vocabulário que os classificadores identificam como tóxico — evidenciando como o contexto do jogo e o humor adulto podem gerar falsos positivos e desafiar sistemas de moderação automática.

Os jogos com maior proporção individual de toxicidade são apresentados na Figura 4b. Três dos dez títulos possuem comunidades que normalizam um vocabulário explícito em suas interações — *Hentai Furry* (11,54%), *Changed* (9,62%) e *Genital Jousting* (8,33%) —, nas quais a linguagem classificada como tóxica reflete convenções do grupo e do conteúdo do jogo, e não necessariamente uma intenção ofensiva.

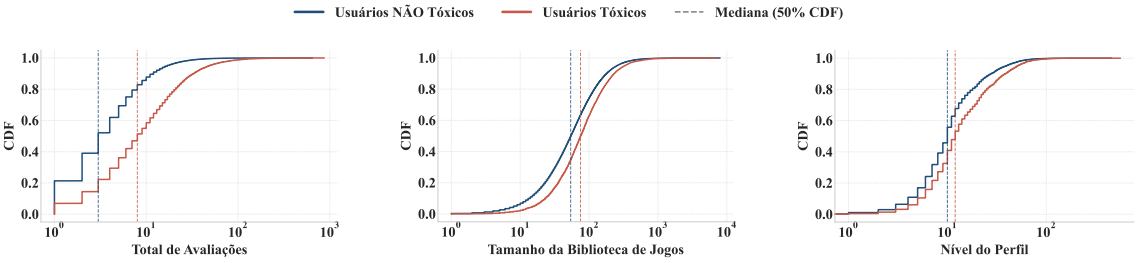
Um segundo grupo é composto por jogos de violência extrema — *Paint the Town Red* (9,94%), *ULTRAKILL* (9,06%), *Wolfenstein: The New Order* (8,84%) e *Fear & Hunger* (8,75%) —, nos quais a toxicidade tem origem dupla. De um lado, o vocabulário utilizado para descrever a *gameplay* — termos como *sarrafo*, *porrada* e *morte* — pode ser capturado pelo classificador como ofensivo mesmo quando descreve mecânicas do jogo. De outro, a imersão contínua em conteúdo violento parece normalizar um registro que ocasionalmente cruza para a agressividade genuína, como em *"Amo dar porrada em pessoas cúbicas, melhor que chutar mendigo na rua."* (**Paint the Town Red**), onde a violência do jogo é explicitamente comparada a violência contra pessoas reais.

Por fim, destacam-se *PANICORE* (9,26%), *Garten of Banban* (8,77%) e *Granny* (8,43%), jogos de horror indie com mecânicas simples e público casual. Nesses, a toxicidade aparece na reação emocional exagerada ao conteúdo, e na frustração com a qualidade do jogo expressa de forma agressiva.

4.3. QP3: Diferença entre os usuários

Com o objetivo de entender as diferenças comportamentais entre usuários regulares e usuários tóxicos na plataforma, analisou-se o total de comentários, número de jogos na conta e nível do usuário na plataforma.

A distribuição do total de avaliações por usuário (Figura 5a) revela que usuários tóxicos tendem a ser mais ativos: ao percentil de 50%, esse grupo acumula até 8 avaliações, enquanto usuários regulares apresentam até 3 para o mesmo ponto da distribuição. Padrão análogo emerge na distribuição do número de jogos por conta (Figura 5b). Ao percentil de 50%, usuários tóxicos possuem até 75 jogos, contra 54 dos usuários regulares. A distribuição de níveis (Figura 5c) é menos assimétrica, mas confirma a tendência, uma vez que, ao percentil de 50%, usuários atingem até o nível 12, enquanto os regulares chegam ao nível 10. O nível de um perfil na Steam reflete o



(a) Avaliações (b) Volume de jogos (c) Nível

Figura 5. Distribuições acumuladas (CDF) das métricas analisadas

engajamento geral do usuário, sendo determinado por atividades como aquisição de jogos, conquistas, participação em eventos e interações na comunidade.

Consolidados, os resultados indicam que usuários tóxicos são, em média, mais ativos na plataforma em múltiplas dimensões, ao passo que publicam mais avaliações, possuem mais jogos e apresentam maior progressão no sistema de níveis, o que sugere maior investimento e permanência na plataforma. Contudo, a análise da taxa de banimento revela um comportamento contraintuitivo. Dos 7.657 usuários classificados como tóxicos, 279 (3, 6%) já foram banidos, enquanto que os 238.613 usuários regulares, 7.044 (3, 0%) sofreram a mesma sanção. A diferença proporcional é pequena — especialmente considerando que os usuários tóxicos produziram ao menos uma avaliação classificada como tóxica. Em outras palavras, apesar de apresentarem maior atividade e, consequentemente, maior potencial de impacto negativo na comunidade, esses usuários não são punidos com probabilidade significativamente maior.

5. Conclusão

Este trabalho investigou a toxicidade nas avaliações da Steam produzidas pela comunidade brasileira, revelando um fenômeno heterogêneo e semanticamente complexo. A análise lexical evidenciou que a linguagem ofensiva frequentemente coexiste com recomendações positivas e descrições de *gameplay* (QP1), desafiando a premissa de que toxicidade e hostilidade são equivalentes. Jogos competitivos e de conteúdo violento concentram as maiores proporções desse comportamento (QP2), enquanto usuários tóxicos, apesar de mais engajados, escapam à punição na mesma medida que os demais (QP3) — sinalizando lacunas nos mecanismos de moderação da plataforma.

Do ponto de vista metodológico, a combinação de detecção automática e anotação humana revelou divergências relevantes entre a Perspective API e o Detoxify, reforçando a necessidade de abordagens multi-modelo em domínios linguisticamente específicos. Trabalhos futuros podem estender a análise a avaliações em outros idiomas e culturas, ampliando o escopo do dataset, além de incorporar modelos sensíveis ao contexto para mitigar os falsos positivos estruturais identificados. Espera-se que os resultados aqui apresentados sejam úteis tanto para pesquisadores da área de PLN e moderação de conteúdo quanto para desenvolvedores e plataformas de distribuição digital que buscam compreender e mitigar comportamentos tóxicos em comunidades de jogadores.

Declaração de uso de IA. O modelo *Claude Sonnet 4.6 (Anthropic)* foi utilizado para a geração de tópicos sintéticos, conforme a metodologia, e para revisão textual. Todo o conteúdo analítico e intelectual deste estudo é de autoria exclusiva dos autores.

Agradecimentos. CAPES, CNPq, FAPEMIG e INCT-TILD-IAR (#408490/2024-1).

Referências

Backlino Team (2026). Steam Usage and Catalog Stats. BackLinko. `https://backlinko.com/steam-users`. Acessado em: 27 mar. 2026.

Blackburn, J. e Kwak, H. (2014). Stfu noob! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the International Conference on World Wide Web (WWW)*, page 877–888.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., e Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Cohn, T., He, Y., e Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP*, pages 3356–3369.

Gibrim, P., Ribeiro, M., Moreira, L., Melo, P., e Reis, J. (2025). I'm not welcome here – uma revisão de escopo sobre toxicidade e discurso de Ódio em jogos digitais. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 646–660.

Google (2026). Perspective API | documentation. Acesso em: 25 abr. 2026.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. In *arXiv preprint arXiv:2203.05794*. arXiv.

Kou, Y. (2020). Toxic behaviors in team-based competitive gaming: The case of League of Legends. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play: CHI PLAY*, pages 81–92.

Landis, J. R. e Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., e Vasserman, L. (2022). A new generation of perspective API: Efficient multilingual character-level transformers. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 3197–3207.

Lima, L., Reis, J. C. S., Melo, P., Murai, F., e Benevenuto, F. (2020). Characterizing (un)moderated textual data in social systems. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 430–434.

McInnes, L. e Healy, J. (2017). Accelerated HDBSCAN* Clustering Hierarchical Density Based Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):228–241.

McInnes, L., Healy, J., Saul, N., e Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Moreira, L. S., Gibrim, P. T. M., Rocha, L., e Reis, J. C. S. (2026). Anatomy of data repositories for the analysis and detection of toxicity in Portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 456–466.

NewZoo (2024). 2024 Global Games Market Report | Free Version. Technical report, Newzoo. Acessado em: 08 de ago. 2025.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., e Androutsopoulos, I. (2020). Toxicity detection: Does context really matter? In Jurafsky, D., Chai, J., Schluter, N., e Tetreault, J., editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4296–4305.

Peron, D. M., Passos, L. P., Camargo, F. E. d., e Fornari, J. (2025). Videogames: transcendendo vício, violência e escapismo. In *Anais do XXIV Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*.

Plunkett, L. (2013). Steam is 10 today. remember when it sucked? https://kotaku.com/steam-is-10-today-remember-when-it-sucked-1297594444/. Acessado em: 4 fev. 2026.

Reimers, N. e Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks.

Salton, G. e Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Sun, X., Yu, V., e Chen, V. H. H. (2025). Toxic behavior in multiplayer online games: the role of witnessed verbal aggression, game engagement intensity, and social self-efficacy. *Chinese Journal of Communication*, 18(4):426–444.

Unitary (2024). Detoxify | documentation. Acesso em: 25 abr. 2026.

Vadim Berman (2026). Goodbye, perspective api. Medium. Acessado em: 26 abr. 2026.

Valve Corporation (2026a). Community moderation. Acessado em: 26 abr. 2026.

Valve Corporation (2026b). Steam: Download Stats. Steam Store. Acessado em: 26 abr. 2026.

Valve Corporation (2026c). Steam Web API. Acesso em: 20 abr. 2026.

Vasconcelos, F. d. F. M., Nascimento, A. M., e Souza, A. P. C. d. (2025). Explorando o comportamento do consumidor em jogos eletrônicos: uma análise dos perfis e valores dos gamers pagantes no brasil. In *Anais do Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*.

Waseem, Z. e Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In Andreas, J., Choi, E., e Lazaridou, A., editors, *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Wijman, T. (2018). Mobile revenues account for more than 50% of the global games market as it reaches $137.9 billion in 2018. Newzoo Blog. Acessado em: 8 ago. 2025.

xPaw (2026). Steam Charts. `https://steamdb.info/app/753/charts/`. Acesso em: 04 fev. 2026.