Incivilidade em Comunidades Brasileiras de Jogos Competitivos no Discord: Limitações da Moderação Automatizada

Title: Incivility in Brazilian Competitive Gaming Communities on Discord: Limitations of Automated Moderation

Thiago Lima Matos[1], Marina Capelo[1], Salete Fontenele Lopes[1], Diogo Saraiva[1], Alicia Paiva[1], Matheus Paixao[1]

[1]Curso de Ciências da Computação – Universidade Estadual do Ceará (UECE)
Fortaleza – CE – Brasil

{thiaguinho.lima, marina.capelo,salete.lopes}@aluno.uece.br,

{diogo.saraiva,alicia.paiva}@aluno.uece.br,matheus.paixao@uece.br

Abstract. Introduction: *Online gaming communities have grown in scale and relevance, but this expansion has been accompanied by persistent problems of toxicity and incivility, particularly in competitive settings.* ***Objective:*** *This paper investigates how the Brazilian competitive gaming community behaves on Discord and what limitations automated models face in this context.* ***Methodology:*** *Building on the Discord Unveiled corpus, we derive a large, filtered subset of Brazilian game-related channels in Portuguese and construct two complementary message samples, which are automatically classified using Detoxify and Perspective API and manually evaluated.* ***Results:*** *The models showed good overall agreement, but relevant biases emerged in handling colloquial vulgarities, acronyms, and code-mixed insults.*
Keywords *Online games, Discord, Toxicity detection, Incivility, Brazilian Portuguese.*

Resumo. Introdução: *As comunidades de jogos online cresceram em larga escala, mas essa expansão foi acompanhada por problemas persistentes de toxicidade e incivilidade, sobretudo em ambientes competitivos.* ***Objetivo:*** *Este artigo investiga como a comunidade brasileira de jogos competitivos se manifesta em servidores do Discord e quais limitações os modelos automáticos de classificação de incivilidades apresentam nesse contexto.* ***Metodologia:*** *A partir do corpus Discord Unveiled, é derivado um subconjunto filtrado de canais de jogos em português brasileiro e construídas duas amostras de mensagens, anotadas com as ferramentas Detoxify e Perspective e avaliadas manualmente.* ***Resultados:*** *Os modelos apresentaram boa concordância geral, mas mostraram vieses relevantes ao lidar com vulgaridades coloquiais, siglas e insultos em código misto.*
Palavras-Chave *Jogos online, Discord, Detecção de toxicidade, Incivilidade, Português brasileiro.*

1. Introdução

As comunidades de jogos digitais têm se consolidado como espaços centrais de sociabilidade, aprendizagem e interação entre jogadores [Brito e Ramos 2024]. No

entanto, essa expansão tem sido acompanhada por um fenômeno persistente e preocupante: a toxicidade e a incivilidade nas interações entre jogadores. O termo incivilidade é adotado para se referir de forma ampla a essas formas ofensivas, desrespeitosas e normativamente inadequadas de interação entre jogadores, enquanto toxicidade é utilizado para denotar expressões linguísticas que podem ser operacionalizadas e medidas por classificadores automáticos de linguagem natural.

Estudos revelam que aproximadamente 66% dos jogadores relatam terem sido vítimas de comportamentos tóxicos, sendo os mais jovens particularmente vulneráveis em ambientes competitivos [Zsila et al. 2022]. A gravidade dessa vivência é alarmante, visto que cerca de 80% dos usuários afirmam testemunhar comentários preconceituosos com frequência [Nexø et al. 2024]. Além disso, a exposição à toxicidade apresenta características de contágio social, aumentando causalmente a probabilidade de jogadores passivos exibirem comportamento agressivo semelhante em partidas futuras [Kou et al. 2025]. As consequências desse ecossistema hostil se estendem para além do ambiente virtual. A normalização da ciberviolência induz sintomas severos de depressão, hábitos problemáticos de consumo e consolidação do assédio desproporcional, principalmente contra minorias sociais [Gisbert-Pérez et al. 2024, Gan et al. 2024, Sun et al. 2024]. Entretanto, a literatura sobre toxicidade em jogos *online* continua sendo desenvolvida e testada majoritariamente em inglês, deixando uma lacuna profunda de ferramentas e análises contextualizadas para o contexto de jogadores no Brasil [Gibrim et al. 2025].

Nesse contexto, o Discord surge como uma plataforma particularmente relevante por delegar a governança de conteúdo quase que integralmente aos próprios usuários e administradores de servidores descentralizados [Seering 2020]. A plataforma se consolidou como um dos principais espaços de coordenação e sociabilidade de jogadores, e o conjunto de dados *Discord Unveiled* [Aquino et al. 2025], que reúne mais de 2,05 bilhões de mensagens públicas em 3.167 servidores entre 2015 e 2024, representa um marco infraestrutural para o estudo empírico desses ambientes.

Para preencher parcialmente essa carência empírica, este artigo concentra-se na construção de um subconjunto de mensagens em português brasileiro em servidores de jogos competitivos no Discord e na avaliação da adequação de classificadores de toxicidade existentes a esse cenário sociolinguístico. A partir do *Discord Unveiled*, é extraído um subconjunto de canais relacionados a jogos. Sobre esse subconjunto, aplica-se um *pipeline* de filtragem que remove *spam*, canais de divulgação e conteúdo promocional, e que classifica os canais de acordo com o idioma predominante de suas mensagens. Como resultado, obtém-se um *corpus* de mensagens orgânicas em português ou português misto, distribuídas em canais com forte participação da comunidade brasileira, que serve de base para as análises apresentadas neste trabalho.

Tomando esse *corpus* como ponto de partida, são definidas duas amostras complementares de mensagens. A primeira é uma amostra orgânica, selecionada de forma inteiramente aleatória, que busca refletir o fluxo cotidiano de conversas entre jogadores brasileiros. A segunda é uma amostra alvo, construída a partir de um léxico de palavras potencialmente incivís comuns no vernáculo *gamer* brasileiro. Essas mensagens são utilizadas para auditar o desempenho de dois classificadores automáticos amplamente adotados na literatura, o Detoxify [Hanu e Unitary team 2020] e a Perspective API, bem

como para explorar as nuances sociolinguísticas das interações.

Dessa forma, este trabalho contribui em três frentes principais. Primeiro, deriva-se do *Discord Unveiled* um *corpus* de grande escala focado em canais de jogos competitivos com predominância de português brasileiro, preservando metadados ricos e reduzindo substancialmente ruído de *spam* e conteúdo promocional. Segundo, constrói-se um conjunto anotado de mensagens orgânicas e potencialmente tóxicas em português brasileiro, permitindo uma análise qualitativa das fronteiras entre *trash talk*, humor agressivo e incivilidade em comunidades *gamer*. Terceiro, realiza-se uma avaliação comparativa de classificadores de toxicidade pré-existentes, evidenciando vieses e limitações desses modelos no contexto brasileiro e fundamentando a necessidade de futuras soluções de detecção de incivilidade sensíveis ao idioma e à cultura locais.

2. Trabalhos relacionados

A identificação e mitigação de comportamentos tóxicos em ambientes virtuais têm mobilizado esforços significativos nas áreas de Ciência da Computação e Ciências Sociais. No contexto dos jogos digitais, a incivilidade é frequentemente influenciada pelo efeito de desinibição *online* e pelo distanciamento moral inerente ao anonimato e à competitividade [Gan et al. 2024]. Para capturar empiricamente esses fenômenos, trabalhos recentes têm buscado isolar a comunicação entre jogadores em diferentes plataformas e modalidades de jogo.

No ecossistema do Discord, Fillies et al. [Fillies et al. 2024] compilaram um conjunto de dados focado em mensagens de ódio produzidas por adolescentes, evidenciando a complexidade de moderar comunidades segmentadas e altamente especializadas. Em ambientes *in-game*, Yang et al. [Yang et al. 2023] demonstraram as dificuldades de treinar modelos de detecção de toxicidade diretamente sobre *chats* de partidas, destacando problemas de viés e de generalização dos classificadores. Em escala semelhante, Naseem et al. [Naseem et al. 2025] apresentaram o GameTox, um *dataset* detalhadamente anotado com 53 mil interações de jogadores, voltado para a análise de intenção tóxica em jogos; contudo, o *corpus* é limitado ao idioma inglês e a interações restritas ao contexto de partidas, deixando de lado as interações que ocorrem em canais de comunidade e coordenação.

Apesar desses avanços, a pesquisa focada na realidade sociolinguística brasileira ainda é incipiente. Uma revisão de escopo conduzida por Gibrim et al. [Gibrim et al. 2025] sobre discurso de ódio em jogos constatou que a esmagadora maioria das publicações concentra-se nos contextos norte-americano e europeu, com pouca atenção dedicada a comunidades lusófonas. No domínio do português brasileiro, os principais esforços em moderação automatizada têm sido impulsionados por conjuntos de dados oriundos de redes sociais de *microblogging*, como o ToLD-BR [Leite et al. 2020], voltado à toxicidade em postagens do Twitter, e o OLID-BR [Trajano et al. 2023], focado em linguagem ofensiva. Esforços mais recentes buscaram ampliar o escopo, como o ToxSyn-PT [Brito et al. 2025], que gerou mais de 50 mil sentenças sintéticas para detecção de discurso de ódio em português; no entanto, trata-se de um *dataset* construído com modelos de linguagem, que não reflete plenamente a organicidade das interações informais entre jogadores.

O presente trabalho diferencia-se ao focar em mensagens orgânicas provenientes

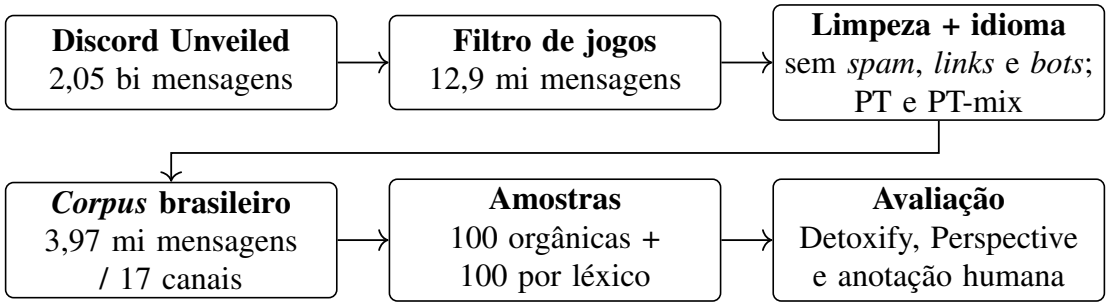

Figura 1. Síntese do fluxo metodológico de construção do *corpus* brasileiro e avaliação das amostras.

de interações de jogos competitivos no ecossistema do Discord, entendido como paratexto que circunda, complementa e prolonga as experiências *in-game*. Em vez de se concentrar exclusivamente em *chats* dentro do jogo ou em dados sintéticos, é construído um *corpus* de aproximadamente 3,97 milhões de mensagens em português ou português misto, extraído de 17 canais com forte participação da comunidade brasileira. A partir desse *corpus*, são definidas amostras para avaliar os classificadores de toxicidade amplamente utilizados, com o objetivo de compreender tanto os limites dessas ferramentas no contexto brasileiro quanto as nuances sociolinguísticas da incivilidade em comunidades *gamer*.

3. Metodologia

Esta seção descreve o fluxo metodológico adotado para construir um *corpus* brasileiro de mensagens de jogos competitivos no Discord, selecionar amostras para análise e avaliar as limitações de classificadores automáticos de toxicidade. A Figura 1 sintetiza as principais etapas do processo.

3.1. Preparação dos dados

O conjunto de dados *Discord Unveiled*[1] é disponibilizado publicamente comprimido no formato Zstandard[2]. Na etapa inicial, o arquivo comprimido foi convertido para o formato Apache Parquet[3] utilizando um sistema gerenciador de banco de dados analítico em modo colunar.[4] Essa transformação permitiu a execução de consultas analíticas eficientes sobre milhões de mensagens, reduzindo o custo de leitura e facilitando a filtragem incremental.

Em seguida, foi realizada uma filtragem lexical para isolar canais relacionados a jogos competitivos. Foram definidas expressões regulares (*regex*) *case-insensitive* projetadas para capturar variações morfológicas, abreviações e erros de digitação de títulos de *eSports*. Como exemplo, para *Counter-Strike: Global Offensive* foram incluídos padrões como `\bcounter[\s\-_]?strike\b`, que cobre as formas "counter strike" e "counter-strike", e `\bcs[\s\-_]?:?[\s\-_]?go\b`, que captura variantes de "cs go", "cs:go" e "cs-go". Mensagens geradas por *bots* foram excluídas por meio do campo lógico que indica autoria automatizada, de forma a evitar a contaminação da análise.

[1] https://huggingface.co/datasets/SaisExperiments/Discord-Unveiled-Compressed

[2] https://facebook.github.io/zstd/

[3] https://parquet.apache.org/

[4] https://duckdb.org/

3.2. Filtragem de canais brasileiros no Discord

A partir do subconjunto de canais de jogos competitivos, composto por aproximadamente 12,9 milhões de mensagens em 815 canais, foi conduzida uma filtragem voltada à identificação de espaços com forte presença da comunidade brasileira. O objetivo foi reduzir ruídos como *spam*, divulgação comercial, mensagens automatizadas e conteúdo em outros idiomas.

Inicialmente, foram excluídos canais estruturalmente não orgânicos, cujos nomes remetiam a divulgação, vendas, sorteios, lojas, trocas, recrutamento ou autopromoção. Também foram removidas mensagens individuais contendo convites de servidor, anúncios com *links*, ofertas explícitas de produtos ou serviços e mensagens muito curtas ou vazias. Após essa etapa de limpeza, restaram cerca de 7,3 milhões de mensagens orgânicas em 763 canais.

Na etapa seguinte, estimou-se o perfil linguístico de cada canal. Para isso, foram amostradas até 200 mensagens orgânicas por canal, com pelo menos oito caracteres, selecionadas por embaralhamento determinístico dos identificadores das mensagens, de modo a evitar a concentração em um único trecho do histórico. Cada mensagem foi submetida ao detector *lingua-language-detector*[5], complementado por uma heurística lexical para identificar alternância de códigos típica de comunidades *gamer* brasileiras. Um canal foi considerado brasileiro quando ao menos 40% das mensagens válidas foram classificadas como português ou aceitas pela heurística de `PT-mix`. Canais com menos de oito mensagens válidas foram descartados por insuficiência de evidência, enquanto canais com nomes fortemente associados ao público brasileiro, como “chat-br”, “brasil” e “brazil”, puderam ser preservados mesmo sem atingir o limiar principal.

Com esse critério, 741 canais foram avaliados: 16 foram classificados como majoritariamente em português e um canal adicional foi incorporado por forte evidência lexical de comunidade brasileira, enquanto os demais foram classificados como de outros idiomas ou permaneceram sem dados suficientes para decisão. Por fim, foi construído o *corpus* brasileiro retornando a esses 17 canais aprovados e recuperando todas as suas mensagens orgânicas que atendiam aos filtros de conteúdo e de tamanho mínimo. O resultado é um *corpus* de aproximadamente 3,97 milhões de mensagens distribuídas em 17 canais, que serve como base para as amostras avaliadas e para as análises de incivilidade apresentadas neste trabalho.

3.3. Estratégias de amostragem

A partir do *corpus* brasileiro filtrado, foram construídas duas amostras complementares, cada uma com 100 mensagens. A primeira, denominada amostra orgânica, teve como objetivo representar o fluxo cotidiano de conversas entre jogadores brasileiros. Para isso, foi extraído um conjunto intermediário de aproximadamente 15 mil mensagens candidatas por amostragem aleatória, priorizando mensagens curtas, conversacionais, com comprimento entre 8 e 120 caracteres e sem *links* ou convites de servidor. Essas mensagens foram submetidas novamente à detecção de idioma, mantendo-se apenas aquelas reconhecidas como português. A partir desse conjunto, foi selecionada aleatoriamente, de forma reprodutível, uma amostra final de 100 mensagens.

[5] `https://github.com/pemistahl/lingua-py`

A segunda, denominada amostra alvo, foi direcionada a mensagens com maior probabilidade de toxicidade lexical. Para sua construção, foi utilizado um dicionário curado de 44 palavras-chave associadas a xingamentos, palavrões e insultos comuns no vernáculo de jogadores brasileiros. O léxico foi dividido em dois grupos: (i) termos inequivocamente associados ao português brasileiro, como "porra", "caralho", "cuzão", "viado" e variantes; e (ii) termos ambíguos, que também podem ocorrer em outros idiomas ou contextos, como "puta", "lixo", "noob", "fdp" e "burro".

Com base nesse léxico, foi aplicada uma expressão regular sobre o *corpus* brasileiro para selecionar mensagens com pelo menos uma dessas palavras, comprimento entre 8 e 200 caracteres e ausência de *links*. Mensagens contendo termos inequivocamente brasileiros foram mantidas diretamente, enquanto mensagens contendo apenas termos ambíguos foram submetidas novamente à detecção de idioma. Do conjunto resultante, foi selecionada aleatoriamente, também de forma reprodutível, uma amostra final de 100 mensagens. Assim, a amostra orgânica permite observar o comportamento dos classificadores diante de mensagens comuns, enquanto a amostra alvo concentra casos lexicalmente mais próximos de toxicidade.

3.4. Classificação automática e avaliação manual

As duas amostras, totalizando 200 mensagens, foram organizadas em formato tabular e avaliadas com dois classificadores automáticos de toxicidade: Detoxify e Perspective API. Para cada mensagem, foram registrados os escores produzidos pelos modelos nas categorias analisadas, permitindo comparar convergências e divergências entre as ferramentas.

Além da classificação automática, foi realizada uma avaliação manual de caráter qualitativo, com o objetivo de auditar os resultados dos modelos e identificar casos de falso positivo, falso negativo e discordância entre classificadores. Nessa etapa, as mensagens foram analisadas considerando não apenas a presença de termos ofensivos, mas também sua função comunicativa, a existência de alvo explícito e a distinção entre incivilidade, *trash talk*, humor agressivo e vulgaridade expressiva.

A avaliação manual não teve como objetivo estimar de forma definitiva a prevalência de incivilidade no *corpus* completo, mas examinar as limitações dos classificadores diante de usos sociolinguisticamente marcados do português brasileiro em comunidades *gamer*. Assim, os resultados devem ser interpretados como uma auditoria qualitativa exploratória, voltada à identificação de limitações, inconsistências e padrões de erro apresentados pelos modelos.

4. Resultados

Esta seção apresenta os primeiros resultados obtidos a partir das amostras selecionadas, com foco na avaliação da viabilidade de classificação automatizada de incivilidade em português brasileiro e na identificação de padrões sociolinguísticos específicos da comunidade *gamer*. Embora sejam apresentados resultados quantitativos descritivos sobre a concordância entre os classificadores, o foco principal da análise é qualitativo. Essa escolha decorre do objetivo do estudo: compreender as limitações sociolinguísticas dos classificadores automáticos diante do português brasileiro *gamer*, e não estimar de forma definitiva a prevalência de incivilidade em todo o *corpus*.

4.1. Panorama quantitativo da amostra anotada

Tabela 1. Resumo dos resultados de classificação por ferramenta

Ferramenta	Total de Mensagens	Mensagens Tóxicas	Percentual (%)
Detoxify (Multi)	200	81	40,5%
Perspective API	200	82	41,0%

As duas estratégias de amostragem descritas na Seção 3.3 resultaram em um conjunto de 200 mensagens em português, provenientes de múltiplos canais de jogos competitivos. A amostra foi avaliada por dois classificadores de toxicidade: o modelo multilíngue Detoxify e a Perspective API, considerando o atributo *TOXICITY*. Para fins de análise binária, adotou-se o limiar $0{,}5$ para rotular uma mensagem como tóxica em cada modelo.

Os resultados agregados indicam que o Detoxify rotulou 81 das 200 mensagens como tóxicas (40,5%), enquanto a Perspective API rotulou 82 mensagens como tóxicas (41,0%). Ambos os modelos concordaram em rotular 64 mensagens como tóxicas e 101 como não tóxicas, o que corresponde a uma concordância de 82,5% entre os modelos. Em contrapartida, houve 35 mensagens (17,5%) em que apenas um dos modelos marcou toxicidade: 17 mensagens foram consideradas tóxicas apenas pelo Detoxify, e 18 apenas pela Perspective.

Ao observar os escores contínuos de toxicidade, a correlação de Pearson entre as predições dos dois modelos atinge 0,82, sugerindo forte alinhamento estrutural na maior parte do espectro de casos. No entanto, como discutido a seguir, essa convergência mascarou divergências sistemáticas em subgrupos específicos de mensagens, especialmente aquelas que exploram nuances do vernáculo brasileiro *gamer*.

4.2. Padrões de divergência entre Detoxify e Perspective

A análise qualitativa das 35 mensagens em que houve discordância entre os modelos revelou dois padrões principais de viés. O primeiro diz respeito à hipersensibilidade à vulgaridade coloquial por parte do Detoxify. O modelo atribuiu escores elevados de toxicidade a expressões vulgares utilizadas em contextos de desabafo ou ênfase emocional, sem alvo pessoal claro. Exemplos típicos incluem mensagens como “Brasil porra” ou comentários autoirônicos do tipo “meu português é merda” em contexto de brincadeira *in-game*, que foram classificados como severamente tóxicos pelo Detoxify, mas considerados não problemáticos pela Perspective API. Nesses casos, o uso de palavrões funciona mais como marcador de intensidade afetiva ou de identidade local do que como ataque.

O segundo padrão, em sentido inverso, está relacionado à não identificação de siglas e insultos compactos em português por parte do Detoxify, com melhor desempenho relativo da Perspective. Mensagens contendo abreviações como “fdp” ou insultos regionalizados, por exemplo “gringo corno” ou “bando de corno”, foram frequentemente subestimadas ou ignoradas pelo Detoxify, enquanto a Perspective API lhes atribuía escores altos de toxicidade. De forma semelhante, combinações típicas de xenofobia e rivalidade esportiva (por exemplo, “seus americano de merda, aqui é Brasil porra”)

foram mais bem capturadas pela Perspective, sugerindo maior sensibilidade a ataques direcionados, ainda que em código misto.

Esses padrões indicam que, enquanto o Detoxify tende a superestimar a toxicidade de palavrões descontextualizados e expressões de frustração com o jogo, a Perspective API é mais sensível a insultos pessoais compactos, siglas e ataques com alvo explícito. Em ambos os casos, observa-se que a distinção entre *trash talk* competitivo, humor agressivo e incivilidade propriamente dita é desafiadora para classificadores treinados majoritariamente fora do contexto linguístico e cultural brasileiro.

4.3. Incivilidade, *trash talk* e nuances sociolinguísticas

A inspeção conjunta das 200 mensagens evidenciou que o léxico utilizado por jogadores brasileiros em canais de jogos competitivos situa-se em uma fronteira tênue entre ofensa e ludicidade. Palavrões e insultos diretos aparecem tanto em contextos de ataque explícito a outros jogadores quanto em situações de comemoração, camaradagem e reforço de identidade nacional, como em "AQUI É VASCO PORRA" ou "TEM ALGUM BRASILEIRO AQUI PORRA" observados na amostra. Do ponto de vista social, essas expressões podem ser interpretadas como elementos de *trash talk* e de performatividade competitiva; entretanto, do ponto de vista de um modelo supervisionado, tendem a acionar detecções automáticas de toxicidade.

Ao mesmo tempo, foram encontrados casos em que ataques explícitos a grupos ou indivíduos escapam parcialmente à sensibilidade dos classificadores. Insultos que combinam ofensa, nacionalidade ou pertencimento regional (e.g., "seus americano de merda") e comentários desclassificadores ligados a gênero ou orientação sexual apresentam variação considerável nos escores atribuídos, a depender do padrão lexical exato e do idioma predominante da mensagem. Essas observações sugerem que, mesmo em um recorte relativamente pequeno, há forte dependência das escolhas lexicais locais, de siglas e de gírias, confirmando a necessidade de abordagens sensíveis ao contexto sociolinguístico brasileiro.

4.4. Implicações para a classificação em larga escala

Os resultados desta etapa preliminar não têm como objetivo fornecer uma estimativa exata da prevalência de incivilidade nos 3,97 milhões de mensagens brasileiras filtradas, mas avaliar a adequação de modelos existentes como alicerce para a rotulagem em larga escala. A alta correlação entre escores e concordância de 82,5% nas decisões binárias indicam que Detoxify e Perspective capturam um núcleo comum de casos claramente tóxicos ou não tóxicos. No entanto, a análise qualitativa mostra que uma fração não desprezível das mensagens mais sociolinguisticamente marcadas (código misto, gírias regionais, palavrões com função expressiva) é tratada de forma inconsistente.

Do ponto de vista deste trabalho, isso reforça a importância de empregar esses classificadores com cautela, combinando-os a amostras anotadas manualmente e a regras lexicais específicas para português brasileiro ao construir um futuro *dataset* rotulado de incivilidade. A auditoria das 200 mensagens fornece um primeiro mapeamento dos pontos cegos e dos vieses culturais desses sistemas, orientando a escolha de limiares de toxicidade, estratégias de combinação entre modelos e prioridades para a expansão da anotação humana sobre o *corpus* brasileiro.

5. Limitações do estudo

Apesar do esforço em construir um *pipeline* rigoroso de filtragem e análise, este estudo está sujeito a diferentes ameaças à validade que devem ser explicitadas.

A primeira limitação refere-se à forma como incivilidade e toxicidade são operacionalizadas. Neste trabalho, incivilidade é entendida como um fenômeno mais amplo de violação de normas de respeito em interações entre jogadores, enquanto toxicidade é medida indiretamente por modelos de linguagem supervisionados e por um léxico de 44 termos ofensivos. Essa aproximação pode não capturar dimensões pragmáticas importantes, como ironia ou humor autodepreciativo, além de depender fortemente da superfície lexical.

Adicionalmente, os classificadores automáticos empregados (Detoxify e Perspective API) foram treinados majoritariamente em contextos culturais e linguísticos distintos do cenário brasileiro, o que introduz riscos de vieses culturais e de sub ou superestimação de toxicidade em gírias, siglas e código misto. Expressões que funcionam como marcadores de identidade ou ênfase emocional podem ser confundidas com ataques hostis, enquanto insultos compactos e regionalismos podem ser subestimados.

Em termos de validade interna, o processo de detecção de idioma e de classificação de canais pode introduzir erros sistemáticos. O detector de idioma utilizado é adequado a textos curtos e informais, mas não é infalível, especialmente diante de gírias, abreviações e mensagens muito breves. Isso pode levar tanto à exclusão de mensagens genuinamente em português quanto à aceitação ocasional de conteúdo em outros idiomas.

A decisão de classificar um canal como brasileiro a partir de uma amostra de até 200 mensagens e de um limiar de 40% de mensagens em português ou português misto também implica incerteza. Canais altamente mistos ou com mudanças de perfil ao longo do tempo podem ser aprovados ou rejeitados de forma sensível à amostra específica considerada. A heurística adicional que resgata canais com marcadores explícitos de brasilidade no nome busca mitigar casos como o de canais “br-chat” com forte mistura de idiomas, mas permanece baseada em regras manuais e, portanto, sujeita a exceções.

A amostra anotada, composta por 100 mensagens orgânicas e 100 mensagens selecionadas por léxico tóxico, fornece um recorte útil para auditoria, mas é relativamente pequena frente ao universo de 3,97 milhões de mensagens.

Por fim, a confiabilidade do *pipeline* depende de decisões heurísticas que podem ser ajustadas em trabalhos futuros. As listas de palavras-chave tóxicas, os padrões de nomes de canal e de conteúdo promocional e os limiares empregados na detecção de idioma e na classificação de canais foram definidos com base em inspeções exploratórias e em boas práticas relatadas na literatura, mas não constituem parâmetros únicos ou definitivos. Pequenas alterações nessas regras podem levar a variações na composição exata do corpus brasileiro.

6. Conclusão e trabalhos futuros

Este artigo apresentou um *pipeline* metodológico para a construção de um subconjunto brasileiro de mensagens de jogos competitivos no Discord, a partir do corpus *Discord Unveiled*. Por meio de uma combinação de engenharia de dados, filtragem lexical e detecção automática de idioma, foi possível derivar um conjunto de aproximadamente

3,97 milhões de mensagens orgânicas em português ou português misto, distribuídas em 17 canais com forte evidência de comunidade brasileira. Esse corpus oferece uma base estruturalmente rica e tematicamente focada para o estudo de incivilidade em contextos gamer.

A partir desse universo, foram definidas duas amostras complementares de mensagens em português: uma amostra orgânica, selecionada aleatoriamente, e uma amostra alvo, composta por mensagens contendo termos de um léxico de 44 palavras potencialmente incivis. As 200 mensagens foram avaliadas manualmente para comparar o desempenho dos modelos Detoxify e Perspective API. Os resultados evidenciaram elevada correlação entre os escores e alta concordância nas classificações binárias, mas também revelaram diferenças sistemáticas na detecção de incivilidade, indicando maior sensibilidade do Detoxify a vulgaridades coloquiais e da Perspective a siglas ofensivas, insultos compactos e ataques direcionados.

Esses achados reforçam que a distinção entre *trash talk* competitivo, humor agressivo e incivilidade propriamente dita é particularmente desafiadora em comunidades gamer, sobretudo quando se consideram código misto, gírias locais e marcadores identitários. Os classificadores analisados capturam um núcleo comum de casos claramente hostis ou claramente neutros, mas apresentam dificuldade em lidar com zonas cinzentas sociolinguisticamente marcadas. Consequentemente, o uso direto de suas predições como rótulos definitivos de incivilidade pode levar tanto a falsos positivos quanto a falsos negativos em escala.

Como trabalhos futuros, pretende-se expandir a anotação para um subconjunto maior e mais diverso de mensagens brasileiras, de forma a construir um *dataset* anotado de incivilidade em português brasileiro com escala suficiente para treinar e avaliar modelos especializados. Essa expansão incluirá a definição de um esquema de rótulos mais refinado, contemplando diferentes tipos de incivilidade como insultos genéricos, ameaças, assédio sexual e *trash talk* competitivo.

Como continuidade, pretende-se também explorar métodos de classificação automatizada baseados em *large language models* (LLMs), com o objetivo de investigar se modelos mais recentes conseguem capturar melhor as nuances pragmáticas e sociolinguísticas da incivilidade em português brasileiro. Em particular, esses modelos podem ser avaliados tanto em cenários de classificação binária quanto em tarefas multirrótulo, nas quais diferentes tipos de incivilidade sejam identificados de forma mais fina. Essa linha de investigação pode contribuir para reduzir os erros observados nos classificadores tradicionais e ampliar a robustez da rotulagem automática em comunidades gamer.

Por fim, uma vez estabelecida essa infraestrutura de dados e modelos, abrem-se caminhos para estudos substanciais sobre o comportamento da comunidade *gamer* brasileira, abrangendo análises temporais, da relação entre tipos de canal e formas de agressão, dos vetores de toxicidade associados a gênero, região e minorias, além da recepção de personagens e jogadores LGBTQIAPN+. Essas investigações podem subsidiar estratégias de moderação e políticas mais adequadas ao contexto brasileiro.

Referências

Aquino, Y., Bento, P., Buzelin, A., Dayrell, L., Malaquias, S., Santana, C., Estanislau, V., Dutenhefner, P., Evangelista, G. H. G., Porfírio, L. G., Grossi, C. S., Rigueira, P. B., Almeida, V., Pappa, G. L., e Meira Jr., W. (2025). Discord unveiled: A comprehensive dataset of public communication (2015-2024). *arXiv preprint arXiv:2502.00627*.

Brito, C. R. d. e Ramos, D. K. (2024). Comunidades de jogos digitais: formas de habitar e aprender. *Revista e-Curriculum*, 22:e61666.

Brito, I. A., Dollis, J. S., Färber, F. B., Silva, D. F. C., e Galvão Filho, A. R. (2025). Toxsyn-pt: A large-scale synthetic dataset for hate speech detection in portuguese. *arXiv preprint arXiv:2506.10245*.

Fillies, J., Peikert, S., e Paschke, A. (2024). Hateful messages: A conversational data set of hate speech produced by adolescents on discord. In *Data Science—Analytics and Applications*, pages 37–44. Springer.

Gan, W. et al. (2024). Aggression in online gaming: The role of online disinhibition, social dominance orientation, moral disengagement, and gender traits. *Frontiers in Psychology*, 15:1480214.

Gibrim, P. T. M., Ribeiro, M. V. G., Moreira, L. S., Melo, P. d. F., e Reis, J. C. S. (2025). I'm not welcome here – uma revisão de escopo sobre toxicidade e discurso de ódio em jogos digitais. In *Anais do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, pages 646–660. SBC.

Gisbert-Pérez, J. et al. (2024). Gender differences in internet gaming among university students: A discriminant analysis. *Frontiers in Psychology*, 15:1412739.

Hanu, L. e Unitary team (2020). Detoxify. GitHub repository.

Kou, Y. et al. (2025). How do the effects of toxicity in competitive online video games vary by source and outcome? *PLOS ONE*, 20(6):e0325462.

Leite, J. A., Silva, D. F., Bontcheva, K., e Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.

Naseem, U., Shiwakoti, S., Shah, S. B., Thapa, S., e Zhang, Q. (2025). Gametox: A comprehensive dataset and analysis for enhanced toxicity detection in online gaming communities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 440–447.

Nexø, L. A. et al. (2024). Toxic behaviours in esport: A review of data-collection and toxicity detection methods. *International Journal of Esports*, 1(1).

Seering, J. (2020). Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28.

Sun, X. et al. (2024). Toxic behavior in multiplayer online games: The role of witnessed verbal aggression, game engagement and social self-efficacy. *Behaviour & Information Technology*.

Trajano, D., Bordini, R. H., e Vieira, R. (2023). Olid-br: offensive language identification dataset for brazilian portuguese. In *Language Resources and Evaluation*. Springer.

Yang, D. et al. (2023). Unveiling identity biases in toxicity detection: A game-focused dataset and methodological framework. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 271–281.

Zsila, Á., Shabahang, R., Aruguete, M. S., e Orosz, G. (2022). Toxic behaviors in online multiplayer games: Prevalence, perceived severity, and effects on gaming experience. *Telematics and Informatics*, 73:101856.