Toxicidade em Chats de Live Streaming: Um Estudo Comparativo com Coleta Síncrona em Twitch, YouTube e Kick

Rodrigo Cardoso[1], **Carlos A. Silva**[1]

[1]Departamento de Informática – Instituto Federal de Minas Gerais (IFMG)
CEP 34.590-390 – Sabará, MG – Brasil

rodrigo.cardoso1007@gmail.com,
carlos.silva@ifmg.edu.br

Abstract. *This paper investigates toxicity patterns in live streaming chats by mining 6.8 million messages collected simultaneously from Twitch, YouTube, and Kick over 90 days. Brazilian Portuguese messages were processed through a two-stage NLP pipeline using BERT models, combining semantic polarity classification and toxicity categorization. Kick presented the highest proportion of harmful messages (21.4%), followed by YouTube (19.6%) and Twitch (16.3%), with insults and obscene language predominating across platforms. Temporal analysis identified recurring peaks between 11 a.m. and 4 p.m. (UTC-3), evidencing consistent differences in toxicity profiles across platforms and reinforcing the viability of synchronous collection for large-scale multi-platform comparisons.*

Resumo. *Este artigo analisa padrões de toxicidade em chats de live streaming por meio da mineração de 6,8 milhões de mensagens coletadas simultaneamente na Twitch, YouTube e Kick durante 90 dias. As mensagens em português brasileiro foram processadas por um pipeline de PLN em duas etapas com modelos BERT, combinando polaridade semântica e categorização de toxicidade. A Kick apresentou a maior proporção de mensagens nocivas (21,4%), seguida pelo YouTube (19,6%) e pela Twitch (16,3%), com predominância de insultos e linguagem obscena. A análise temporal identificou picos recorrentes entre 11h e 16h (UTC-3), mostrando diferenças consistentes no perfil de toxicidade entre plataformas e indicando que a coleta síncrona é adequada para comparações multiplataforma em larga escala.*

1. Introdução

Plataformas de streaming ao vivo se tornaram grandes espaços de interatividade social, onde milhões de usuários trocam mensagens em tempo real durante transmissões de jogos, eventos esportivos e de entretenimento. Twitch, YouTube e Kick atraem grandes públicos e geram imensos volumes de dados textuais, refletindo dinâmicas complexas do comportamento coletivo. Segundo [Li et al. 2020], a Twitch acumulou mais de 434 bilhões de minutos assistidos em 2019, mostrando a dimensão desse ecossistema, que segue em expansão com o surgimento de novas plataformas e formatos de transmissão. Esse cenário torna os chats de *live streaming* um objeto relevante de estudo nas áreas de computação, mineração de dados e análise de interações online.

Contudo, a dinâmica veloz e síncrona desses ambientes também favorece a ocorrência e a rápida disseminação de comportamentos abusivos. Estudos apontam

que plataformas digitais interativas têm se consolidado como espaços recorrentes para a manifestação de mensagens nocivas, incluindo insultos, assédio, discurso de ódio e outras formas de toxicidade [Sheth et al. 2022]. Em transmissões de jogos, como *League of Legends*, *Counter-Strike 2*, *Valorant* e *Grand Theft Auto 5*, esses comportamentos podem ser intensificados por frustrações, rivalidades e tensões associadas à dinâmica das partidas, ampliando a agressividade nas interações do chat [Morrier et al. 2025].

Embora comunidades online frequentemente se distribuam por múltiplas plataformas, ainda não está claro em que medida os usuários ajustam seu comportamento linguístico às características de cada ambiente. Políticas de controle, mecanismos de engajamento e regras da comunidade podem ter um impacto considerável na produção e na aceitação de conteúdos tóxicos [Dreier and Pirker 2023]. No entanto, análises comparativas sistemáticas entre plataformas ainda são escassas, e poucos estudos adotam estratégias de coleta simultânea. Essa limitação metodológica compromete a validade das comparações, pois diferenças observadas podem refletir períodos distintos de coleta, e não variações estruturais reais entre plataformas [Singh et al. 2024].

Este estudo investiga, então, padrões de mensagens nocivas para ajudar a preencher essa lacuna em chats de serviços como Twitch, YouTube e Kick, extraindo dados de 6,8 milhões de mensagens coletadas ao mesmo tempo durante 90 dias em streams de jogos, com o intuito de avaliar as variações no perfil de toxicidade entre diferentes plataformas que estão sob o mesmo intervalo de tempo. As mensagens passaram por um pipeline de Processamento de Linguagem Natural (PLN) em duas fases, fundamentado em modelos da arquitetura BERT, possibilitando reconhecer a polaridade semântica e classificar vários tipos de toxicidade presentes nas interações.

As principais contribuições deste trabalho são:

- uma análise comparativa de padrões de toxicidade em três plataformas de *live streaming*, baseada em coleta síncrona em larga escala sob condições temporais equivalentes;
- a aplicação de um *pipeline* de PLN em duas etapas para detecção e categorização de mensagens nocivas em chats de *live streaming* em português brasileiro, considerando textos curtos, informais e ruidosos;
- a identificação de padrões temporais de toxicidade ao longo do dia, com picos horários recorrentes e diferenças consistentes entre plataformas.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta a revisão da literatura relacionada ao tema. A Seção 3 descreve como os dados foram coletados e pré-processados, além da arquitetura de análise. A Seção 4 apresenta os resultados obtidos. Por fim, a Seção 5 conclui o trabalho e aponta direções para pesquisas futuras.

2. Revisão Bibliográfica

A análise automatizada de mensagens em chats de *live streaming* apresenta desafios que a diferenciam de outros domínios da mineração de texto. Ao contrário de ambientes assíncronos, como fóruns e redes sociais tradicionais, os chats síncronos caracterizam-se por um fluxo intenso de mensagens curtas, uso massivo de gírias, abreviações e referências contextuais imediatas. Essas características limitam a aplicação direta de

técnicas convencionais de Processamento de Linguagem Natural (PLN) e exigem abordagens adaptadas a dados ruidosos e de alta densidade temporal. [Ringer et al. 2020] demonstraram que a distribuição lexical de chats da Twitch não obedece a pressupostos clássicos como a Lei de Zipf — segundo a qual a frequência de uma palavra é inversamente proporcional à sua posição no ranking de popularidade do vocabulário — sugerindo que esse ecossistema possui particularidades linguísticas próprias. Essa característica reforça a necessidade de modelos robustos à linguagem informal de redes sociais, motivando o uso de modelos pré-treinados em dados de redes sociais em português brasileiro, como os adotados neste trabalho.

No campo da detecção automática de toxicidade, as técnicas variam de métodos estatísticos a arquiteturas de aprendizado profundo. [Tereshchenko and Hämäläinen 2025] realizaram um estudo comparativo com modelos tradicionais, *transformers* ajustados e grandes modelos de linguagem, avaliando acurácia, tempo de resposta e custo computacional em chats de jogos. Os resultados indicam que a eficiência computacional é crítica em ambientes síncronos, onde o alto fluxo de mensagens exige processamento analítico em tempo quase real. Essa restrição torna importante a abordagens em cascata, nas quais uma triagem inicial reduz o custo de etapas posteriores de análise mais detalhada. O estudo, contudo, não cruza dados entre plataformas distintas nem emprega coleta simultânea, o que limita comparações multiplataforma sob o mesmo recorte temporal.

Focando no português brasileiro, [Vargas et al. 2021] introduziram o HateBR, um *corpus* anotado por especialistas para a identificação de linguagem ofensiva e discurso de ódio em comentários do Instagram. No contexto de *live streaming*, [Lobo et al. 2025] analisaram a disseminação de toxicidade em cascatas de chat da Twitch, revelando padrões de contágio entre mensagens em português brasileiro. No entanto, o estudo se limita a uma única plataforma e não realiza coleta simultânea em diferentes ambientes. Para analisar toxicidade em domínios síncronos e informais, a literatura tem recorrido a bibliotecas de análise de sentimento, como o *pysentimiento* [Pérez et al. 2021], e a classificadores granulares pré-treinados disponíveis no repositório Hugging Face,[1] como o modelo *toxicity-type-detection* [Trajano 2023]. Esses trabalhos e recursos representam avanços importantes para o português brasileiro, mas não abrangem de forma satisfatória o cenário de chats síncronos, ruidosos e comparados entre múltiplas plataformas de *live streaming*.

Além das ferramentas, fatores contextuais e características específicas de cada plataforma — como sistemas de moderação automatizada, mecanismos de engajamento (*emotes*, *raids*, assinaturas) e a arquitetura do chat — afetam a disseminação de discursos prejudiciais [Dreier and Pirker 2023]. [Sheth et al. 2022] defendem que a identificação de toxicidade em redes sociais depende do contexto das mensagens, tornando comparações entre plataformas distintas um desafio metodológico. [Dreier and Pirker 2023] descobriram relações entre o aumento da toxicidade na Twitch e fatores como gênero do *streamer*, tamanho da comunidade e nicho do jogo, mostrando que mesmo em uma única plataforma o comportamento nocivo é heterogêneo, embora o estudo tenha se limitado a cerca de 100 mil mensagens. Em transmissões de jogos competitivos, [Morrier et al. 2025] mostraram que a toxicidade se propaga de forma viral, amplificada por frustrações e ri-

[1] `https://huggingface.co`

validades, padrão relevante para o corpus deste trabalho. [Singh et al. 2024] compararam linguagem tóxica no Reddit e no Discord, mostrando que arquitetura e normas modulam a interação; nenhuma dessas redes opera sob dinâmica de *live streaming*, e a ausência de simultaneidade reduz a robustez de comparações temporais.

Comparações empíricas entre plataformas exigem condições estritas de coleta. Quando os dados são obtidos em janelas temporais distintas, fica difícil determinar se as variações detectadas decorrem de diferenças estruturais das plataformas ou de flutuações sazonais no comportamento dos usuários [Singh et al. 2024]. Sem controle temporal, afirmações comparativas sobre o comportamento de usuários entre plataformas ficam menos robustas, pois podem refletir oscilações do período analisado e não diferenças estruturais entre os ambientes. A Tabela 1 sintetiza as principais referências mapeadas e mostra a lacuna estrutural que o presente estudo busca preencher.

Tabela 1. Comparação de trabalhos relacionados sobre análise e detecção de toxicidade em comunicação online.

Trabalho	Plataforma(s)	Síncrono	Idioma	Método principal	Escala / Foco
[Ringer et al. 2020]	Twitch	Sim	EN	Análise linguística/estatística	Padrões linguísticos em chats
[Dreier and Pirker 2023]	Twitch	Sim	EN	VADER + Detoxify	∼100 mil msgs; fatores contextuais
[Tereshchenko and Hämäläinen 2025]	Chats de jogos	Sim	Var.	Embeddings / Transformers / LLMs	Custo e tempo de processamento
[Vargas et al. 2021]	Instagram	Não	PT-BR	*Corpus* anotado	∼7 mil comentários
[Lobo et al. 2025]	Twitch	Sim	PT-BR	Análise de cascatas	Propagação de toxicidade
[Singh et al. 2024]	Reddit + Discord	Parcial	EN	Análise semântica e temporal	Comparação entre plataformas
Este trabalho	Twitch, YouTube, Kick	Sim	PT-BR	Classificação em cascata com BERT	Coleta simultânea (6,8M msgs)

Os trabalhos discutidos até aqui contribuíram para a compreensão de diferentes aspectos da toxicidade online, mas nenhum deles combina coleta síncrona multiplataforma, mineração em larga escala e categorização detalhada de discursos nocivos em ambientes de *live streaming* com foco em português brasileiro. Este estudo reúne esses três elementos e permite comparar o comportamento de diferentes plataformas sob as mesmas condições de observação.

3. Metodologia

Esta pesquisa adota uma abordagem empírica baseada na mineração e análise de mensagens provenientes de chats de transmissões ao vivo em múltiplas plataformas. A metodologia, apresentada na Figura 1, foi organizada em quatro etapas: coleta das mensagens, pré-processamento, processamento de linguagem natural e análise estatística.

A fase de **coleta de dados** ocorreu entre 25 de novembro de 2025 e 23 de fevereiro de 2026, totalizando 90 dias contínuos de monitoramento de transmissões de jogos competitivos de grande audiência, incluindo *League of Legends*, *Counter-Strike 2*, *Valorant* e *Grand Theft Auto 5*. Foram monitorados seis canais selecionados com base em dois critérios principais: (i) ocorrência simultânea em pelo menos duas das três plataformas analisadas e (ii) maior média de espectadores nas dez transmissões mais recentes do canal, com frequência mínima de três transmissões por semana. O primeiro critério permitiu observar as plataformas sob o mesmo recorte temporal, reduzindo vieses associados

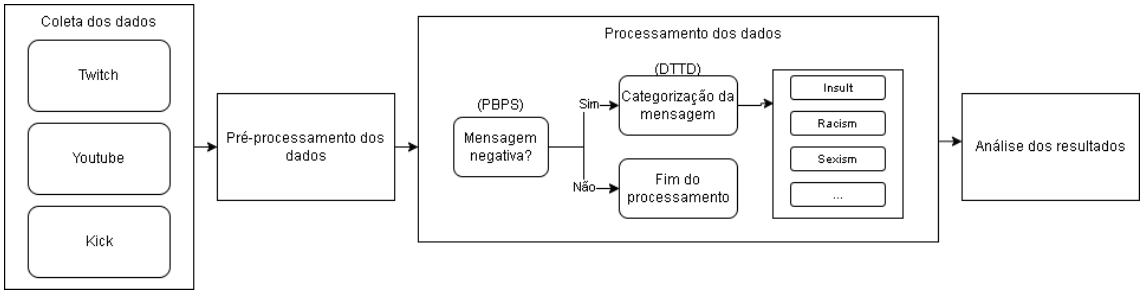

Figura 1. Fluxo metodológico da pesquisa.

a diferenças no período de coleta. O segundo buscou assegurar um *corpus* representativo em termos de engajamento e regularidade das transmissões.

Os canais selecionados eram predominantemente brasileiros, de modo que o português brasileiro foi o idioma predominante no *corpus*, embora também tenham sido registradas mensagens em inglês. Todas as mensagens foram armazenadas em formato bruto, sem filtros prévios, de modo a preservar a dinâmica original das interações síncronas.

Para a captura das mensagens, foram utilizadas soluções técnicas distintas de acordo com as características de cada plataforma. Na Twitch, utilizou-se a biblioteca *tmi.js*, que fornece acesso nativo ao protocolo IRC da plataforma. Para Kick e YouTube, desenvolveu-se um *web crawler* com a biblioteca *Puppeteer*, diante da ausência de APIs públicas que permitam a coleta em tempo real de mensagens de chat. Cada mensagem foi armazenada como um objeto estruturado contendo os metadados `platform`, `channel`, `message`, `author` e `timestamp`, correspondentes, respectivamente, à plataforma de origem, ao canal, ao conteúdo textual, ao identificador do usuário e ao momento de envio. Devido à elevada taxa de inserção concorrente nos horários de pico, adotou-se o MongoDB como repositório de armazenamento, por sua capacidade de suportar altas taxas de escrita simultânea com baixa latência. Ao término do período de monitoramento, foram registradas mais de 6,8 milhões de mensagens textuais brutas.

A etapa de **pré-processamento** teve como objetivo garantir que apenas mensagens representativas de interações humanas fossem submetidas à análise. Foram removidas mensagens automatizadas de usuários identificados como bots, comandos nativos das plataformas, URLs isoladas e emotes sem conteúdo textual associado. Essa filtragem utilizou regras baseadas em padrões recorrentes de mensagens automatizadas, identificação de usuários associados a bots, comandos iniciados por prefixos típicos das plataformas e remoção de mensagens compostas apenas por URLs ou emotes sem conteúdo textual analisável. Após essa etapa, o *corpus* passou de 6.811.720 mensagens brutas para 6.158.647 mensagens válidas, correspondendo a uma taxa de retenção de 90,41%. Em seguida, aplicou-se a anonimização dos dados por meio da biblioteca *anonymizedf*, substituindo nomes de usuários e menções diretas por identificadores genéricos, com o objetivo de preservar a privacidade dos participantes e evitar a identificação individual nas etapas subsequentes.

Como medida ética adicional, os dados foram analisados apenas de forma agregada, sem investigação individual de usuários. Devido à presença original de identifica-

dores e ao caráter sensível de mensagens potencialmente ofensivas, os dados brutos não serão disponibilizados publicamente. Para favorecer a replicabilidade, serão disponibilizados os scripts de coleta[2], pré-processamento e análise[3], juntamente com estatísticas agregadas dos experimentos.

A etapa de **processamento textual** foi realizada por meio de um fluxo de classificação em cascata. Na primeira fase, cada mensagem foi submetida ao modelo *pysentimiento/bertweet-pt-sentiment* (PBPS) [Pérez et al. 2021], pré-treinado em dados de redes sociais em português brasileiro. Esse modelo classificou a polaridade semântica das mensagens em três categorias: positiva, neutra ou negativa. Mensagens classificadas como positivas ou neutras tiveram o processamento encerrado nessa etapa, por apresentarem menor probabilidade inicial de conteúdo nocivo. Entretanto, mensagens irônicas, sarcásticas ou ambíguas podem ser classificadas como positivas ou neutras mesmo contendo conteúdo potencialmente tóxico, o que representa uma limitação da abordagem em cascata.

Algoritmo 1: Processamento e categorização em cascata

Dados: Mensagem de texto $mensagem$
Resultado: Categoria final associada à mensagem
$resultadoPBPS \leftarrow$ ClassificarPBPS($mensagem$)
if $resultadoPBPS =$ *Negativa* **then**
  $categoria \leftarrow$ ClassificarDTTD($mensagem$)
else
  $categoria \leftarrow resultadoPBPS$
return $categoria$

Mensagens classificadas como negativas foram encaminhadas para a segunda fase do *pipeline*, na qual o modelo *dougtrajano/toxicity-type-detection* (DTTD) — uma arquitetura BERT com fine-tuning — classificou o tipo de toxicidade presente. Como mostra a Tabela 2, o modelo identifica dez categorias que incluem diferentes formas de agressão, discriminação e linguagem ofensiva, apresentando uma precisão de 0,8180 e uma acurácia de 0,4214. Ambos os modelos foram aplicados sem validação específica no domínio deste *corpus*, o que deve ser considerado na interpretação dos resultados. Além disso, o modelo DTTD atribui uma categoria principal a cada mensagem, o que pode simplificar casos em que uma mesma mensagem contém simultaneamente insultos, linguagem obscena e conteúdo discriminatório.

Neste estudo, o limite operacional entre mensagens não nocivas e nocivas foi definido pela combinação entre a triagem de polaridade negativa e a posterior classificação em uma das categorias do DTTD: mensagens positivas ou neutras foram tratadas como não nocivas, enquanto mensagens negativas categorizadas pelo DTTD foram consideradas nocivas. Essa decisão reduz o custo computacional ao concentrar a inferência mais custosa apenas no subconjunto com indícios de negatividade semântica, tornando viável a análise de grandes volumes de dados, mas pode gerar falsos positivos em mensagens negativas não abusivas e falsos negativos em mensagens irônicas ou ambíguas.

[2]`https://github.com/rodrigocardoso-rc/web-crawler`
[3]`https://github.com/rodrigocardoso-rc/message-analyzer`

Tabela 2. Categorias de toxicidade identificadas pelo modelo DTTD.

Categoria	Descrição
health	Ataques relacionados a condições de saúde ou doenças
ideology	Hostilidade baseada em crenças políticas ou ideológicas
insult	Insultos diretos e ofensas pessoais gerais
lgbtqphobia	Conteúdo tóxico contra pessoas LGBTQ+
other_lifestyle	Ataques direcionados a escolhas de estilo de vida
physical_aspects	Ofensas relacionadas à aparência física
profanity_obscene	Palavrões, linguagem vulgar ou obscena
racism	Discriminação racial ou étnica
sexism	Discriminação baseada em gênero
xenophobia	Hostilidade contra estrangeiros ou imigrantes

4. Resultados e Discussão

Esta seção apresenta os resultados da análise do *corpus* final, composto por 6.158.647 mensagens válidas, obtidas a partir de um total de aproximadamente 6,8 milhões de mensagens coletadas durante 90 dias em transmissões simultâneas nas plataformas Twitch, YouTube e Kick. Como os volumes absolutos de mensagens diferem entre plataformas, as comparações foram realizadas com base em proporções normalizadas em relação ao total de mensagens de cada ambiente. A seção está organizada em três eixos: (i) proporção global de mensagens nocivas por plataforma, (ii) distribuição das categorias de toxicidade e (iii) variação temporal dessas ocorrências ao longo do dia. Ao final, os três eixos são reunidos em uma síntese dos padrões observados.

4.1. Proporção de mensagens nocivas por plataforma

A Figura 2 apresenta a proporção de mensagens classificadas como nocivas em cada plataforma, após normalização pelo total de mensagens coletadas. A Tabela 3 consolida os principais indicadores quantitativos por plataforma.

Tabela 3. Resumo quantitativo por plataforma.

Plataforma	Total de msgs	Msgs nocivas	Proporção (%)	Categoria principal
Kick	1.787.798	383.305	21,44	*insult*
YouTube	510.138	99.936	19,59	*insult*
Twitch	3.860.711	628.910	16,29	*insult*
Total	**6.158.647**	**1.112.151**	–	–

Embora seja tentador associar essas diferenças diretamente às políticas de moderação de cada plataforma, essa relação deve ser tratada com cautela. Este estudo não mediu variáveis como intensidade de moderação automatizada, taxa de banimento, atuação de moderadores voluntários ou volume de mensagens removidas em tempo real.

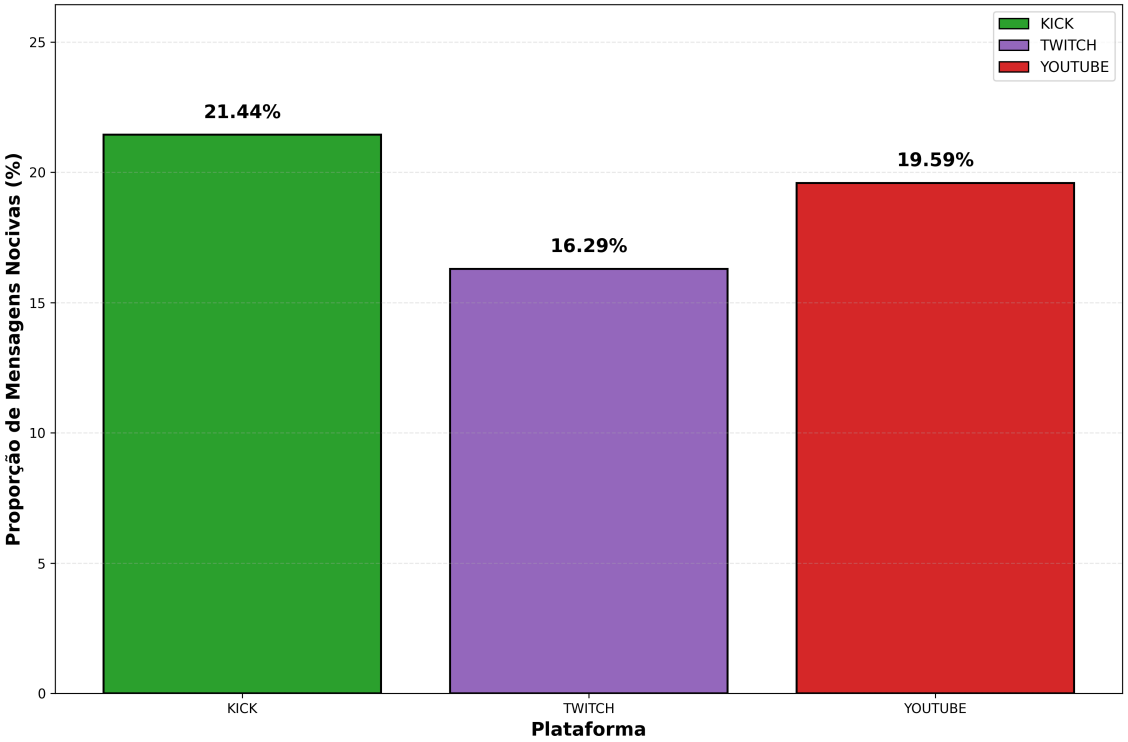

Figura 2. Proporção de mensagens nocivas por plataforma.

Assim, os dados mostram que as três plataformas apresentam taxas de toxicidade sistematicamente distintas sob o mesmo recorte temporal. A literatura sugere que políticas de moderação, mecanismos automáticos de filtragem e normas comunitárias bem estabelecidas geralmente ajudam a reduzir comportamentos abusivos em comunidades online [Dreier and Pirker 2023]. As variações observadas, portanto, são compatíveis com a hipótese de que aspectos estruturais das plataformas contribuam para os padrões de toxicidade registrados, embora relações causais diretas não possam ser estabelecidas a partir dos dados deste estudo.

4.2. Distribuição das categorias de toxicidade

A Figura 3 apresenta a distribuição percentual das categorias de discurso nocivo identificadas pelo modelo DTTD em cada plataforma.

Independentemente da plataforma, as categorias *insult* e *profanity_obscene* concentraram mais de 90% das mensagens classificadas como nocivas. Ainda assim, há diferenças relevantes na composição desse conteúdo. A Twitch apresentou a maior proporção de *insult* (71,32%), acima do YouTube (68,61%) e da Kick (66,39%), enquanto a Kick registrou a maior participação de *profanity_obscene* (27,51%). Esse padrão sugere que, embora a agressividade verbal predomine em todas as plataformas, sua manifestação não ocorre da mesma maneira entre os ambientes observados.

O predomínio dessas duas categorias é compatível com a dinâmica de chats síncronos, nos quais o ritmo acelerado das interações tende a reduzir o tempo de elaboração das mensagens e a favorecer respostas impulsivas [Singh et al. 2024]. Em transmissões de jogos competitivos, frustração, rivalidade entre equipes e reações imedia-

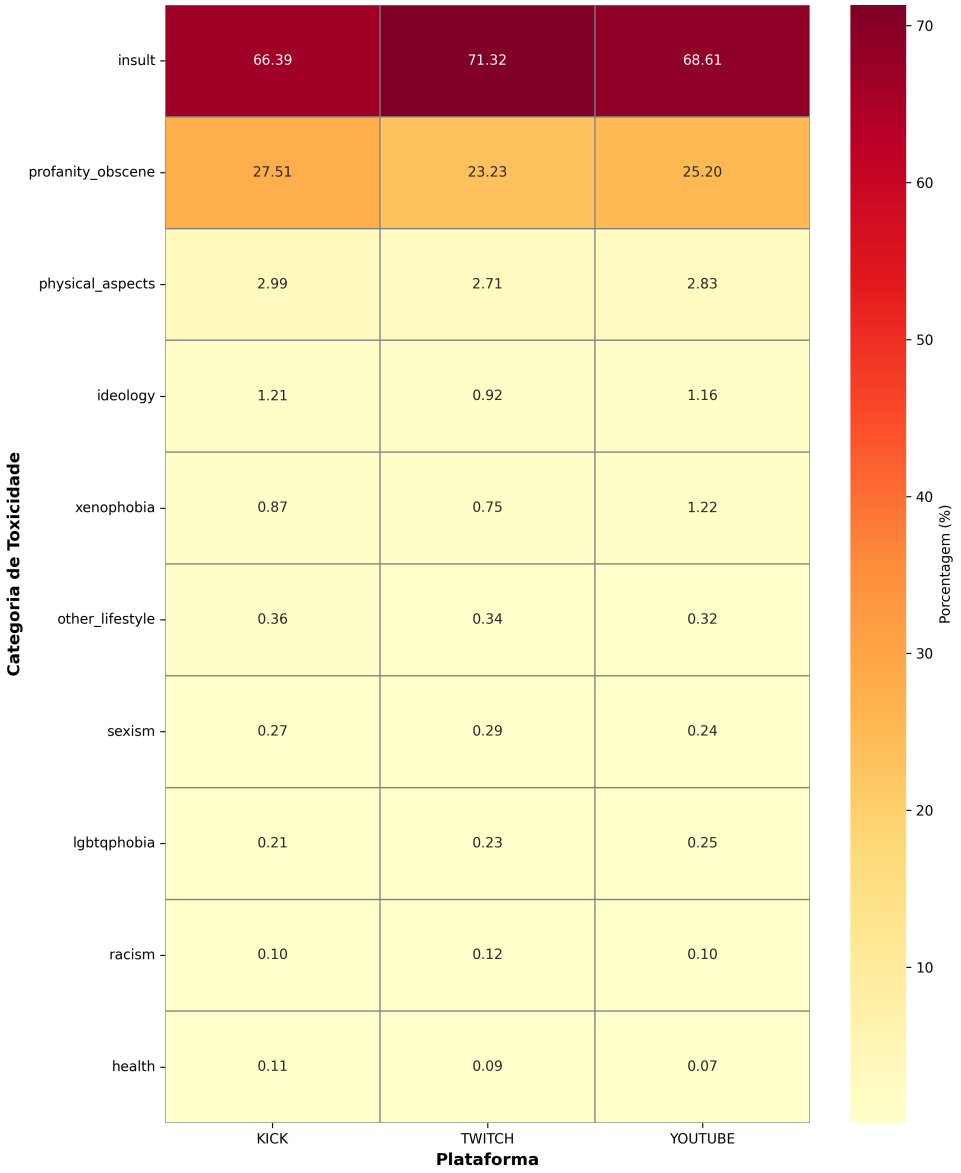

Figura 3. Distribuição das categorias de discursos nocivos por plataforma.

tas a eventos da partida podem intensificar esse comportamento [Dreier and Pirker 2023]. Ainda assim, os resultados apresentados descrevem o *tipo* de linguagem classificada pelo modelo DTTD, e não a intenção dos usuários ou os motivos por trás das mensagens, o que limita inferências sobre seu comportamento.

As categorias associadas a formas mais explícitas de discurso de ódio apresentaram frequências proporcionais menores em todas as plataformas. *Racism* variou entre 0,10% e 0,12%, *sexism* entre 0,24% e 0,29%, e *lgbtqphobia* entre 0,21% e 0,25%. A Twitch registrou, de forma consistente, os menores percentuais também nessas categorias. Embora esses valores sejam baixos em termos relativos, eles correspondem a um número absoluto de ocorrências que não pode ser desconsiderado do ponto de vista prático, especialmente em contextos de moderação e política de plataforma.

4.3. Padrões temporais de toxicidade

A Figura 4 apresenta a variação horária da proporção de mensagens nocivas ao longo das 24 horas do dia, considerando o fuso horário de Brasília (UTC-3).

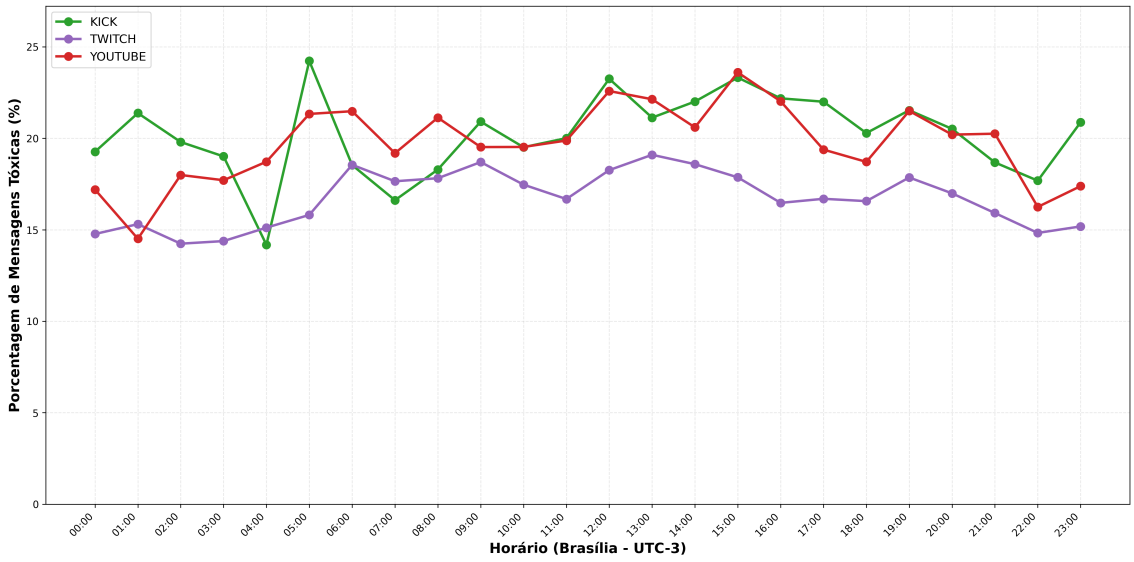

Figura 4. Proporção de mensagens nocivas ao longo do dia.

A Twitch manteve as menores taxas de toxicidade ao longo de quase todo o período analisado, oscilando entre 14% e 19%. Kick e YouTube, por sua vez, apresentaram valores consistentes acima de 20%. As duas plataformas também exibiram variações mais acentuadas ao longo do dia, com picos concentrados principalmente entre 11h e 16h, além de elevações registradas na madrugada, quando ambas superaram 22% de mensagens nocivas.

Esses intervalos são compatíveis com períodos de maior atividade no ecossistema brasileiro de *e-sports*, especialmente em faixas horárias associadas a transmissões competitivas regionais e internacionais. No entanto, essa interpretação deve ser vista como hipótese explicativa, e não como evidência direta, pois o estudo não coletou dados de audiência por janela horária. Outros fatores não controlados, como tipo de partida, composição da audiência em cada faixa de horário e presença ativa de moderadores, podem igualmente contribuir para as variações observadas.

4.4. Síntese dos resultados

Considerados em conjunto, os três eixos de análise apontam para um padrão consistente. A Twitch apresentou as menores taxas de toxicidade em todos os recortes considerados, tanto na proporção global quanto na distribuição por categoria e na variação ao longo do dia. Kick e YouTube, por sua vez, exibiram proporções mais elevadas e maior oscilação temporal.

A predominância de *insult* e *profanity_obscene* em todas as plataformas indica que a toxicidade nesses ambientes se manifesta principalmente como agressividade reativa e linguagem vulgar, mais do que como formas explícitas de discurso de ódio. As categorias menos frequentes não devem ser tratadas como irrelevantes, dado o volume absoluto de ocorrências acumuladas ao longo do período de monitoramento.

Em conjunto, os resultados revelam elementos convergentes e divergentes entre as plataformas. A convergência reside no predomínio universal de *insult* e *profanity_obscene* como categorias majoritárias, sugerindo que a agressividade reativa é um padrão estrutural desses ambientes síncronos. As divergências manifestam-se na proporção global de toxicidade — Kick (21,44%) e YouTube (19,59%) sistematicamente acima da Twitch (16,29%) —, na composição interna das categorias e na amplitude das variações temporais ao longo do dia. Essas diferenças são compatíveis com a hipótese de que normas comunitárias, mecanismos de moderação e a arquitetura de interação influenciam o comportamento linguístico dos usuários, embora relações causais diretas não possam ser estabelecidas com os dados deste estudo.

5. Conclusão

Este trabalho analisou padrões de toxicidade em chats de transmissões ao vivo a partir da mineração de aproximadamente 6,8 milhões de mensagens coletadas simultaneamente nas plataformas Twitch, YouTube e Kick durante 90 dias. A adoção de coleta síncrona multiplataforma mitiga uma limitação metodológica recorrente em estudos baseados em coletas assíncronas [Singh et al. 2024], nos quais diferenças temporais entre amostras dificultam atribuir variações observadas a características das plataformas analisadas. A combinação com o *pipeline* em cascata torna viável o processamento em larga escala sem sacrificar a granularidade da categorização, constituindo, em conjunto, o principal avanço metodológico deste trabalho em relação à literatura.

Os resultados mostraram diferenças consistentes entre as plataformas quanto à incidência de mensagens nocivas, com maior proporção observada na Kick, seguida pelo YouTube e pela Twitch. Em todas as plataformas, a maior parte das mensagens classificadas como nocivas concentrou-se nas categorias *insult* e *profanity_obscene*, enquanto categorias associadas a discurso de ódio explícito apareceram em proporções menores, embora ainda representem um volume absoluto relevante de ocorrências. Também foram identificados padrões temporais recorrentes ao longo do dia, com picos concentrados em faixas horárias compatíveis com maior atividade no ecossistema de *e-sports*, ainda que a relação entre audiência e toxicidade não possa ser verificada diretamente com os dados disponíveis.

Em conjunto, os resultados mostram que as plataformas analisadas diferem não apenas no volume de interação, mas também no perfil proporcional de toxicidade observado em seus chats. Essa contribuição é particularmente relevante por combinar coleta síncrona, larga escala e categorização detalhada de discursos nocivos em português brasileiro no contexto de *live streaming*.

Entre as limitações do estudo destacam-se a seleção de canais focados em transmissões de jogos competitivos, o que restringe a generalização dos resultados para outros domínios de *live streaming* — em particular, não é possível afirmar se o maior perfil de toxicidade observado na Kick se manteria em nichos como streams de música ou eventos ao vivo não competitivos. Outras limitações incluem a estrutura em cascata do *pipeline* de classificação, que pode deixar de identificar mensagens tóxicas expressas por meio de ironia, sarcasmo ou ambiguidade semântica. Além disso, os modelos empregados não foram validados especificamente no domínio deste *corpus*, e o DTTD atribui uma categoria principal por mensagem, o que pode simplificar casos com múltiplas formas simultâneas

de toxicidade.

Trabalhos futuros podem explorar modelos mais robustos para detecção de sarcasmo, ampliar o *corpus* para outros nichos de *live streaming* e outras plataformas como Discord, além de integrar dados de audiência em tempo real para investigar de forma mais direta a relação entre volume de espectadores e incidência de toxicidade.

Referências

Dreier, L. and Pirker, J. (2023). Toxicity in twitch live stream chats: Towards understanding the impact of gender, size of community and game genre. In *2023 IEEE Conference on Games (CoG)*, pages 1–4. IEEE.

Li, Y., Wang, C., and Liu, J. (2020). A systematic review of literature on user behavior in video game live streaming. *International Journal of Environmental Research and Public Health*, 17(9):3328.

Lobo, J. V. C., Barbosa, D. M., and de Freitas Melo, P. (2025). Investigando a dinâmica da propagação do ódio em cascatas de toxicidade nos chats ao vivo da twitch. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 331–339. SBC.

Morrier, J., Mahmassani, A., and Alvarez, R. M. (2025). Uncovering the viral nature of toxicity in competitive online video games. *IEEE Transactions on Games*.

Pérez, J. M., Giudici, J. C., and Luque, F. (2021). pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Ringer, C., Nicolaou, M., and Walker, J. (2020). Twitchchat: A dataset for exploring livestream chat. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, pages 259–265.

Sheth, A., Shalin, V. L., and Kursuncu, U. (2022). Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Singh, A. K., Ghafouri, V., Such, J., and Suarez-Tangil, G. (2024). Differences in the toxic language of cross-platform communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1463–1476.

Tereshchenko, Y. and Hämäläinen, M. K. (2025). Efficient toxicity detection in gaming chats: A comparative study of embeddings, fine-tuned transformers and llms. *Journal of Data Mining & Digital Humanities*.

Trajano, D. (2023). toxicity-type-detection: A model for detecting types of toxicity in portuguese. `https://huggingface.co/dougtrajano/toxicity-type-detection`. Hugging Face model repository, accessed: 2026-02-22.

Vargas, F. A., Carvalho, I., de Góes, F. R., Benevenuto, F., and Pardo, T. A. S. (2021). Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. *arXiv preprint arXiv:2103.14972*.