Investigando a Dinâmica da Propagação do Ódio em Cascatas de Toxicidade nos Chats Ao Vivo da Twitch

João Vitor Chagas Lobo
Universidade Federal de Viçosa
Florestal, MG
joao.lobo@ufv.br

Daniel Mendes Barbosa
Universidade Federal de Viçosa
Florestal, MG
danielmendes@ufv.br

Philipe de Freitas Melo
Universidade Federal de Viçosa
Florestal, MG
philipe.freitas@ufv.br

ABSTRACT

The popularization of live streaming on platforms such as Twitch has transformed how users interact in real time, creating dynamic environments that are also susceptible to the spread of toxic discourse. This poses a new challenge for the automated moderation of these online communities, especially in the face of coordinated hate actions. This study investigates how toxicity spreads on Twitch through the identification and analysis of cascades of toxic messages. To this end, we developed an automated data collection system that monitored, in real time, messages from streamings of popular Twitch channels. These messages were processed and classified using natural language processing tools, including sentiment analysis and hate speech detection via lexical dictionaries from *Hatebase* and *WeaponizedWord*, as well as deep learning models from *PerspectiveAPI* and *pysentimiento*. Based on this classification, offensive content in chats was identified, and 472 toxicity cascades were mapped, enabling the analysis of their frequency, intensity, and dissemination patterns. Our results show that although most hate speech is sporadic, some channels exhibit recurrent and intense episodes of collective toxicity. This study contributes to a better understanding of the dynamics of hate speech propagation in live environments and provides insights for future moderation and intervention strategies.

KEYWORDS

discurso de ódio, toxicidade, Twitch, transmissão ao vivo, cascatas de informação, hate speech, live streamings

1 INTRODUÇÃO

A forma de interação online tem mudado significativamente nas últimas décadas. Inicialmente, usuários da Web eram meros espectadores de conteúdo, com um consumo passivo das informações contidas em websites e blogs. Porém, com a evolução da Internet por meio de fóruns e, mais recentemente, com redes sociais como o Twitter, Instagram e Youtube, a experiência digital transformou-se radicalmente, permitindo que usuários comuns se tornem também produtores de seu próprio conteúdo online. Essa transformação atingiu um novo patamar com a chegada de plataformas de transmissão ao vivo, como a *Twitch*, um serviço voltado principalmente à exibição de vídeos ao vivo, com destaque para conteúdos relacionados a jogos eletrônicos, cultura pop e debates, nas quais a comunicação ocorre em tempo real. Nessas plataformas, o chat funciona como um canal direto entre o público (*viewers*) e o criador de conteúdo (*streamer*), estabelecendo uma dinâmica interativa que sustenta o modelo de comunicação dessas comunidades digitais.

Em 2024, a *Twitch* teve cerca de 21 bilhões de horas assistidas, das quais 2,8 bilhões foram de "Só na conversa" [22], categoria onde o *streamer* conversa com o chat sobre algum assunto qualquer. Nas diretrizes da comunidade da plataforma, estão descritas normas que não permitem conduta de ódio voltada à raça, etnia, orientação sexual, gênero; atos e ameaças de violência, dentre outros. Porém, nestes chats, temos visto recentemente um aumento na ocorrência de casos de misoginia [7], incitação ao ódio [24] e xenofobia [2], o que nos leva a pensar que a *Twitch* pode se tornar um ambiente onde esse tipo de discurso ganha espaço. Devido à natureza dinâmica deste tipo de comunicação e à falta de ferramentas para combater o avanço do ódio dentro da plataforma, este tipo de conteúdo tem-se espalhado livremente sem a devida moderação. Além disso, pouco se sabe sobre os mecanismos que podem impulsionar ou conter a propagação da toxicidade dentro de chats ao vivo.

Por meio de uma coleta em larga escala de mensagens enviadas no chat ao vivo da *Twitch*, do uso de ferramentas automatizadas de identificação de toxicidade, de uma análise de sentimentos e da identificação de termos ofensivos, este trabalho investiga o ódio e a toxicidade nos chats e como eles podem ser agrupados em cascatas. Através dos dados levantados, podemos apontar canais que disseminam esta cultura e que não têm sido devidamente monitorados pela plataforma. Ao fim, esta pesquisa também contribui para a criação de um banco de dados de mensagens de grandes canais da *Twitch* com a identificação de diversas mensagens tóxicas.

O restante deste trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos e estudos relacionados à temática desta pesquisa; a Seção 3 descreve a metodologia adotada, incluindo os critérios de seleção de canais, a coleta e o processamento das mensagens, a forma adotada para identificação das cascatas de toxicidade bem como as limitações técnicas e éticas do estudo; a Seção 4 traz os resultados da caracterização dos dados e da análise de sentimentos; a Seção 5 discute os achados relacionados ao discurso de ódio identificado nos chats; a Seção 6 aprofunda a análise sobre a propagação da toxicidade por meio das cascatas identificadas; e, por fim, a Seção 7 apresenta as conclusões, as principais descobertas e sugestões para trabalhos futuros.

2 TRABALHOS RELACIONADOS

A literatura sobre conteúdo de redes sociais e plataformas de *live stream* vem ganhando grande relevância nos últimos anos [12]. Nestes estudos, são abordados principalmente os eixos de desenvolvimento/investigação de ferramentas para moderação [4, 20, 21], análises de interações entre usuários, culturas de suas comunidades

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

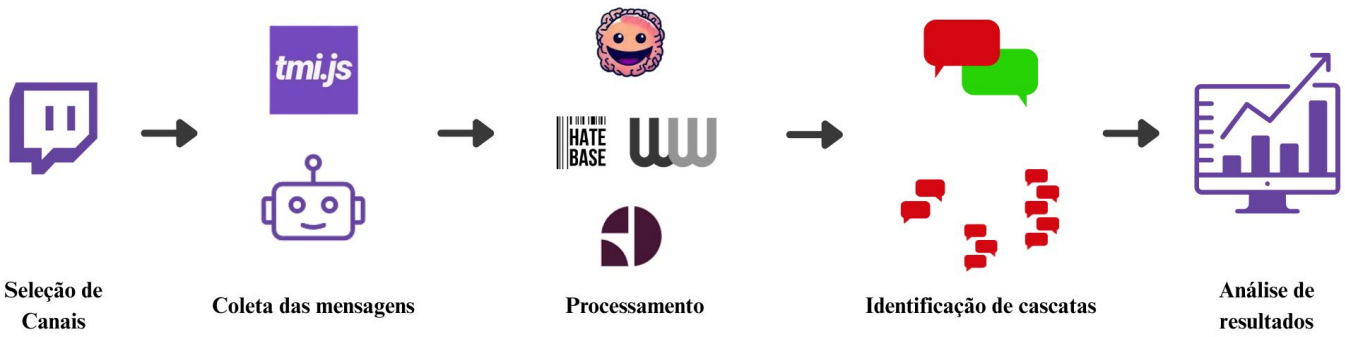


Figura 1: Diagrama esquemático da metodologia aplicada no trabalho

[9, 16] e a identificação de padrões de comportamento [10]. Este trabalho inicia sua investigação dos conteúdos e chats da *Twitch* utilizando uma coleta de mensagens ao utilizar uma metodologia semelhante à implementada por [1]. Em ambas as pesquisas, são realizadas coletas em tempo real do chat das *streams* e são propostas abordagens para análise de sentimentos.

Apesar da *Twitch* contar com usuários moderadores e ferramentas para auto-moderação do chat, alguns estudos mostram que ambos são falhos. É o que exemplifica a pesquisa de [11] que mostra como *emotes* são utilizados para burlar a auto-moderação da plataforma ao substituir letras e representar visualmente um comentário ofensivo. De forma complementar, [13] avaliam como formas de moderação e a ação do *streamer* são utilizadas para inibir a conduta inadequada dos usuários. Essas pesquisas evidenciam que o comportamento abusivo do chat é um fenômeno difícil de monitorar e controlar. Ademais, estudos como o de [18] mostram como a *Twitch* pode se tornar um espaço emergente para discurso político, favorecendo o aparecimento de comportamento hostil entre seus usuários. Este cenário se torna propício à ocorrência de manifestações de discurso de ódio e sua disseminação em cascatas, como o que ocorre no cenário do *Twitter* descrito por [14]. Nossa pesquisa propõe uma maneira mais eficaz de identificar a ocorrência de cascatas de toxicidade na *Twitch* e fornece uma análise de seu comportamento para facilitar o monitoramento de tais eventos.

Mais na direção deste trabalho, alguns estudos têm aprofundado a compreensão da toxicidade e discurso de ódio na *Twitch*, revelando que esse fenômeno vai além do uso isolado de linguagem ofensiva, manifestando-se também em formas coordenadas de assédio. [19] identificaram que chats de *streamers* homens e de canais voltados a jogos tendem a concentrar maior incidência de mensagens tóxicas, com destaque para xingamentos, insultos diretos e sarcasmos ofensivos entre usuários. Já os trabalhos de [3] e [8] exploram as chamadas *hate raids*, ataques coordenados, muitas vezes com uso de bots, que visam sobrecarregar os chats com mensagens de ódio, especialmente contra *streamers* pertencentes a grupos marginalizados, como pessoas negras e LGBTQ+. Complementarmente, [21] realizaram uma auditoria da ferramenta de moderação automática da *Twitch*, o AutoMod, e constataram falhas significativas: até 94% das mensagens de ódio não foram detectadas, enquanto conteúdos pedagógicos ou de empoderamento com termos sensíveis foram erroneamente bloqueados, destacando a dificuldade desses sistemas em interpretar o contexto. Já o estudo de [15] mostra como, diante da omissão da plataforma, os próprios criadores passaram a desenvolver redes de apoio e ferramentas comunitárias para reagir aos ataques, ainda que essas respostas sejam fragmentadas e limitadas.

Esses estudos evidenciam como a toxicidade é um problema atual na *Twitch* e como ela pode ser sistemática e estratégica, explorando as funcionalidades da própria plataforma para evitar detecção ou punição. Além disso, revelam brechas nas ferramentas de moderação existentes e apontam a necessidade de abordagens estruturais capazes de prevenir o abuso e proteger comunidades vulneráveis. Este trabalho se insere nesse contexto ao investigar como essas manifestações tóxicas se organizam em cascatas e como podem ser monitoradas e caracterizadas automaticamente a partir de grandes volumes de dados coletados em tempo real.

3 METODOLOGIA

Nesta seção, apresentamos a metodologia adotada para a realização deste estudo, que abrange desde a coleta e o processamento dos dados até as análises realizadas. Pelo fato das mensagens objetos deste estudo se tratarem de uma forma muito específica e dinâmica (transmissões ao vivo realizadas na plataforma *Twitch*), a obtenção dos dados representa um desafio particular, uma vez que exige o monitoramento contínuo e em tempo real dos chats durante as *live-streamings*. Para contornar essa limitação, foi elaborado um processo estruturado e sistemático que permitisse identificar canais relevantes, capturar as mensagens emitidas durante as transmissões e, posteriormente, processá-las de forma adequada. A Figura 1 apresenta uma visão geral das etapas envolvidas neste processo, que serão detalhadas individualmente nas subseções a seguir.

3.1 Seleção de Canais

A primeira etapa para a construção de um conjunto de dados de mensagens da *Twitch* é a escolha de potenciais canais/*streamers* para coleta das mensagens de chat das suas "*lives*". Para segmentar melhor nossa amostra, de forma que abrangesse canais mais relevantes e aqueles em que haveria uma maior probabilidade de ocorrer manifestações de ódio, os canais foram identificados e coletados a partir de três fontes distintas: dois conjuntos baseados em canais de grande popularidade em inglês e português, respectivamente, e um terceiro conjunto formado por canais atribuídos com marcadores de conteúdo tóxico pela própria Twitch.

Para fazer o levantamento de canais mais populares da Twitch para os dois primeiros grupos, utilizamos o site *TwitchTracker*[1]. Além de conter diversas estatísticas sobre os canais, ele também os classifica periodicamente baseado no seu número médio de espectadores simultâneos, seguidores, visualizações e tempo de transmissão. Utilizando a listagem disponível dias antes do início da coleta, selecionamos os 50 maiores em inglês e, para entender melhor a comunidade brasileira, também foram escolhidos os 50 maiores canais em português para um segundo conjunto.

Para o terceiro conjunto, utilizamos o sistema de marcadores da plataforma. A *Twitch* fornece aos *streamers* a possibilidade de atribuírem *tags* às suas transmissões para categorizar e identificar o próprio conteúdo do seu canal, sendo um recurso muito utilizado para buscas e filtros dentro da plataforma. Entre os marcadores disponíveis existe a *tag* "***toxic***". Combinando este filtro com o filtro de linguagem em inglês e a ordenação de canais com mais para menos espectadores, pôde-se obter uma listagem dos canais autointitulados como tóxicos [2]. Assim, os 30 primeiros canais resultantes desta busca eram adicionados à listagem da coleta, constituindo, desta forma, o terceiro grupo de canais coletados.

3.2 Coleta de Dados na *Twitch*

A coleta de dados numa plataforma de transmissão ao vivo impõe desafios particulares devido à arquitetura dinâmica e efêmera das conversas. Na *Twitch*, o conteúdo é postado ao vivo e não possui ferramentas de busca específicas para o conteúdo do chat que está sendo enviado durante a live, sendo praticamente impossível recuperar as mensagens após a finalização da transmissão. Portanto, precisamos desenvolver estratégias para tentar identificar e coletar o conteúdo tóxico expresso na plataforma no momento em que ele ocorre. Para isto, foi desenvolvido um *bot* em JavaScript que utiliza a biblioteca *tmi.js*[3] para coletar as mensagens. Com a habilitação deste *bot*, é possível monitorar o chat de diversos canais ao vivo simultaneamente, e coletar os dados daquele canal como: (1) `message` (conteúdo da mensagem), (2) `channel` (canal da transmissão), (3) `user-id` (ID único do usuário) (4) `timestamp` (hora de envio da mensagem), dentre outros metadados dos canais, tais como (5) `followers` (número de seguidores), e (6) `viewers` (número de usuários visualizando aquela transmissão).

Para a execução da coleta, foram selecionados alguns dias entre o período de 08/03/2025 a 30/03/2025, em horários que compreenderam os períodos da manhã, tarde e noite. No momento da coleta, o *bot* era habilitado e monitorava um dos três grupos de canais. Ao fim do intervalo de aproximadamente uma hora, o monitoramento era interrompido e o grupo de canais era alterado; esse processo se repetia até que todos os três grupos fossem coletados ao menos uma vez naquele dia. Vale ressaltar que, para os dois primeiros grupos, que foram feitos com base na popularidade dos canais, não havia uma forma de garantir se todos estariam com a transmissão ao vivo durante a coleta, sendo coletados apenas aqueles que, de fato, estivessem ativos, ou seja, fazendo "live" naquele dia. Já para o terceiro grupo, sempre eram coletados os canais com a tag *toxic* ativos para aquele dia. Os dados resultantes estão descritos na Tabela 1.

[1]Disponível em: https://Twitchtracker.com/
[2]Disponível em: https://www.Twitch.tv/directory/all/tags/toxic
[3]Disponível em: https://tmijs.com/

Tabela 1: Resumo dos dados coletados da *Twitch*

Número total de Mensagens	#Mensagens em inglês	#Mensagens em português	Canais coletados	Usuários únicos
627.237	387.423	239.814	165	120.311

Tabela 2: Normalização do nível de toxicidades entre os dicionários léxicos

Valor HateBase	Valor WeaponizeWord	Valor adaptado
null / 0 a 29	Mildly offensive or inoffensive	1
30 a 50	Significantly offensive	2
51 a 69	Moderately offensive	3
70 a 90	Very offensive	4
91 a 100	Extremely offensive	5

3.3 Armazenamento e Processamento dos Dados

A partir do dataset final, com todas as mensagens coletadas devidamente armazenadas em arquivos JSON, o passo seguinte foi fazer o processamento do texto utilizando-se de ferramentas de processamento de linguagem natural (PLN), extraindo assim as informações para basear nossa análise. Nas subseções seguintes, estão detalhados os processos realizados:

Análise de Sentimento: A etapa de análise de sentimentos foi realizada utilizando a biblioteca *pysentimiento* [17], que fornece ferramentas para avaliação de sentimentos em textos provenientes de redes sociais em diversos idiomas. Esse modelo avalia uma sentença e retorna o sentimento predominante, além de fornecer os valores normalizados para os sentimentos positivo, neutro e negativo. Dados que as mensagens obtidas estavam em dois idiomas: inglês e português, foi necessário utilizar dois modelos para a análise das mensagens em ambas as línguas. Com o término do processamento, foram adicionados quatro novos atributos a cada instância dos dados: o "SENT" representando a categoria de sentimento predominante da mensagem, além de possuir os valores individuais de **POS** (Positivo), **NEU** (Neutro) e **NEG** (Negativo) de pontuação fornecida para cada um desses sentimentos.

Análise de Ocorrência de Palavras Ofensivas: Outra análise feita foi a identificação de termos tóxicos e ofensivos utilizando como base duas fontes de dados de expressões de ódio: a *Hatebase* [23] e o *WeaponizedWord*[4]. Ambas as fontes se tratam de dicionários léxicos multilíngues conhecidos de termos considerados potencialmente ofensivos a etnias, gêneros, religiões, nacionalidades, entre outros grupos. Estes dicionários são formas amplamente consolidadas a disponíveis na literatura para análise de ódio em métodos léxicos, (em detrimento daqueles puramente baseados em I.A.) como o de [6]. Embora ferramentas automatizadas de aprendizado de máquina auxiliem no processamento de grandes volumes de dados para identificação de discurso de ódio, esses dicionários são construídos de forma colaborativa (*crowdsource*) por humanos, acrescentando um fator essencial para tratar a subjetividade e dificuldade inerente à tarefa.

Além dos termos, cada dicionário possui uma taxonomia de classificação própria das ofensas ligadas a cada expressão. Um deles é o nível de ofensividade, que para a lista fornecida pela *HateBase*

[4]Lexicographic data courtesy of The Weaponized Word (weaponizedword.org)

se trata de um valor de 0 a 100, enquanto os da *WeaponizedWord* são definidos em 5 categorias distintas. Para que os dados de ambos fossem utilizados no mesmo processo de análise, foi preciso normalizar estes índices para valores que variam de 1 (mais baixo) a 5 (mais alto), conforme a Tabela 2. Desta forma, foi desenvolvido um algoritmo que identifica e contabiliza em cada mensagem a ocorrência dos termos ou sentenças de ódio contidos nela e o seu nível de ofensividade.

Identificação automatizada de conteúdo tóxico: A metodologia principal usada para a identificação e separação de mensagens tóxicas foi o resultado do processamento pela *PerspectiveAPI*[5]. A *PerspectiveAPI* se trata de uma ferramenta em nuvem com um modelo multilíngue pré-treinado de aprendizado profundo ao qual podem ser feitas requisições contendo sentenças e, como retorno, obtém-se valores de 0 a 1 indicando a probabilidade daquela mensagem ser considerada tóxica. Assim, aplicamos o modelo a cada uma das mensagens separadamente para identificar automaticamente o valor de toxicidade delas. Em seguida, selecionamos todas as mensagens com valor de toxicidade maior que 0.8 para serem classificadas como tóxicas. Este valor de limiar foi determinado com base na definição de outros estudos semelhantes, como o de [5] e observando empiricamente o resultado das mensagens encontradas.

3.4 Identificação das Cascatas de Toxicidade

Por último, vários trabalhos têm identificado uma característica particular da *Twitch*, em que a toxicidade do chat vai além do uso isolado de linguagem ofensiva, manifestando-se também em formas coordenadas de assédio. Também observamos em nossos dados que uma mensagem tóxica muitas vezes pode desencadear uma série de outras no chat em sequência; enquanto outras são isoladas dentro da *stream*. Desta forma, também propusemos uma forma de identificar e analisar essas ações mais coletivas, em que várias mensagens são enviadas em sequência, por meio do que chamamos neste trabalho de "***cascatas de toxicidade***". Estas podem então ser definidas como as manifestações conjuntas de cinco ou mais mensagens tóxicas enviadas por um ou vários usuários num curto período de tempo no chat durante as transmissões ao vivo.

O próximo passo para medir a forma como a toxicidade influencia o chat, portanto, consiste na identificação destas cascatas. Antes de mapear as mensagens encadeadas propriamente ditas, filtramos somente o conteúdo tóxico presente em cada canal, considerando o limiar de mensagens com valor de toxicidade maior que 0.8 (descrito anteriormente) para serem salvas em um conjunto separado para análise. E somente após esse processo, observamos e agrupamos então as ocorrências de mensagens relevantes para levantamento das cascatas. Para isso, foi desenvolvido um algoritmo que, para cada canal, processa cada mensagem tóxica e seu horário de envio, daí, conta quantas outras mensagens tóxicas foram enviadas numa janela de 10 minutos. Este intervalo de tempo foi determinado empiricamente de forma suficientemente grande para incluir todas as manifestações de toxicidade relacionadas no chat, mas não tão grande para que o foco do ódio fosse trocado. Além disso, também foi definido que a ocorrência de 10 mensagens tóxicas ou menos durante esta janela de tempo não representa a ocorrência de uma cascata. Observamos que valores muito menores quebravam a discussão, gerando várias cascatas menores que abordavam ainda o mesmo tema. Enquanto ampliar o tempo englobaria mensagens já desconexas da mensagem iniciadora.

[5]Disponível em: https://www.perspectiveapi.com/

3.5 Limitações e Questões Éticas

A coleta de dados em tempo real na *Twitch* impõe desafios específicos, dado que a plataforma não oferece acesso retroativo ao conteúdo dos chats. Embora o uso de um *bot* tenha permitido capturar uma amostra significativa de interações, a distribuição das mensagens entre canais e horários não pôde ser totalmente controlada, pois depende da disponibilidade e duração das transmissões ao vivo; portanto, não temos como verificar se a amostra é representativa da plataforma de forma geral, mas sim que representa uma parcela do público geral da *Twitch*, limitada aos canais e horários coletados. Além disso, os valores associados às definições de cascata foram definidos empiricamente e podem ter um impacto nos resultados obtidos, porém indicam fatores iniciais importantes a serem abordados para aprofundamento do tema. Apesar destas limitações, os resultados representam *insights* interessantes sobre o ecossistema de *live-streaming* e evidenciam características importantes da propagação de ódio online.

Do ponto de vista ético, os dados analisados foram extraídos exclusivamente de fontes públicas, incluindo chats abertos e por meio de APIs oficiais da plataforma. Apesar de seu caráter público, adotamos ainda o cuidado de não expor informações sensíveis ou identificadores diretos de usuários. As mensagens citadas ao longo do texto têm finalidade ilustrativa e foram selecionadas para representar padrões gerais observados no estudo, sem intenção de ofensa ou julgamento. Reconhecer essas limitações contribui para a construção de abordagens metodológicas mais responsáveis e reforça a importância de práticas éticas no estudo de fenômenos sociais em ambientes digitais.

4 CARACTERIZAÇÃO DOS DADOS

Após consolidar nosso dataset, foi realizada uma caracterização geral dos dados obtidos. Na Figura 2 vemos que a maioria dos canais teve apenas uma transmissão coletada durante o período e que, no máximo, tivemos um único canal com 5 transmissões coletadas (canal do *ziggylata*). Ao fim da coleta, foi possível obter dados

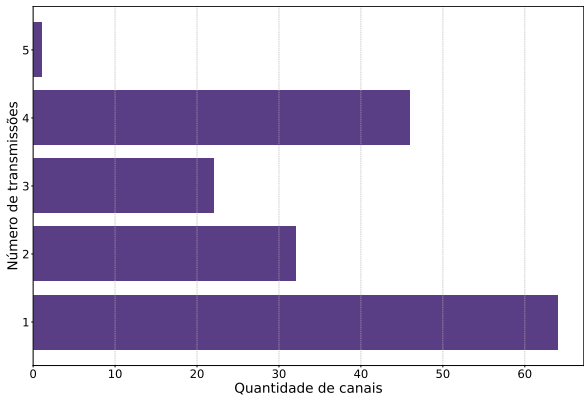
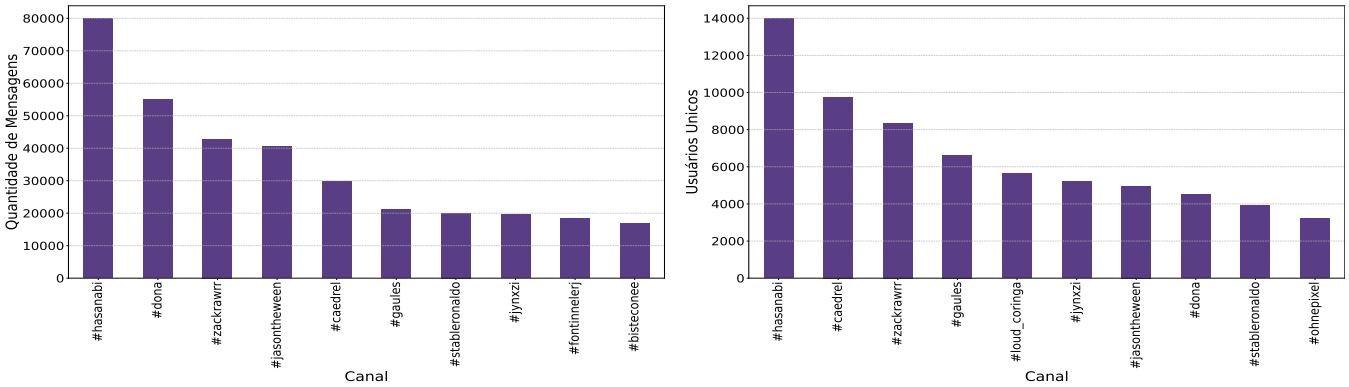


Figura 2: Volume de dados da *Twitch* sobre transmissões coletadas por cada canal monitorado



(a) Total de mensagens por canal **(b) Usuários únicos por canal**

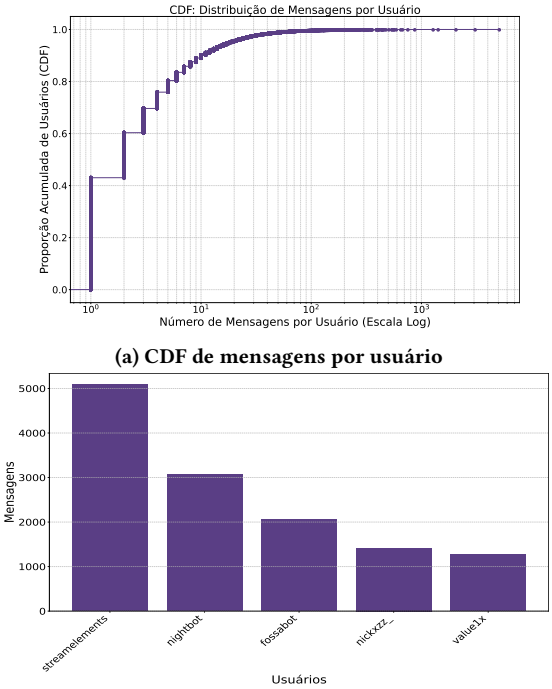
Figura 3: Volume de mensagens enviadas por canal e usuário

de **165** dos 210 canais selecionados. Essa distribuição evidencia a dificuldade de obter dados de forma constante e consistente da plataforma, com alta dependência da transmissão ao vivo estar acontecendo no momento da coleta.

O gráfico da Figura 3a ilustra os 20 canais com maior quantidade de mensagens. Nele pode-se observar que os canais que são exibidos não são necessariamente os primeiros canais do ranking do *TwitchTracker*. Quatro, dentre os cinco canais com mais mensagens coletadas (sendo o segundo canal pertencente ao *streamer* brasileiro *dona*) fizeram, na maior parte do tempo, transmissões em categorias que não são jogos [6]. Durante esse tipo de transmissão, há mais interações entre *streamer* e o chat, e é onde podemos observar um volume maior de mensagens enviadas coletadas. Ao observar a distribuição de usuários únicos por canal mostrada na Figura 3b, observa-se a ocorrência de mais canais brasileiros (*gaules*, *loud_coringa* e *dona*) mostrando que, apesar de não liderarem o ranking do *TwitchTracker*, merecem atenção por movimentarem um grande público. Considerando os gráficos da Figura 3, vemos que os canais dos usuários *hasanabi* e *zackrawrr* (também conhecido como *asmongold*) estão entre os três primeiros em ambos, destacando-se como uma quantidade expressiva dos dados coletados.

Dos dados obtidos, se observarmos a quantidade de mensagens enviadas por cada usuário (Figura 4a), destaca-se que a maioria dos usuários enviou ou apenas **1 mensagem (43,01%)** ou **de 2 a 10 mensagens (47,21%)**, sendo que a minoria **(9,78%) enviou mais de 100 mensagens**. Também temos cinco pontos mais extremos representando usuários que enviaram mais do que 1000 mensagens. Na Figura 4b, temos quem são estes usuários e quantas mensagens cada um enviou. Ao analisar estes usuários e o conteúdo enviado por eles, vemos que os três primeiros se tratam de *bots* que são utilizados pelos próprios canais para moderação e envio de informações relacionadas à própria transmissão. Enquanto o quarto e quinto tratam-se de usuários *spammers*, ou seja, que enviam muitas mensagens em sequência (em sua maioria repetidas); este é um comportamento mal visto por alguns canais e geralmente é contido pela própria moderação.

[6]Dados obtidos do *TwitchTracker*: https://Twitchtracker.com/zackrawrr/games



(a) CDF de mensagens por usuário

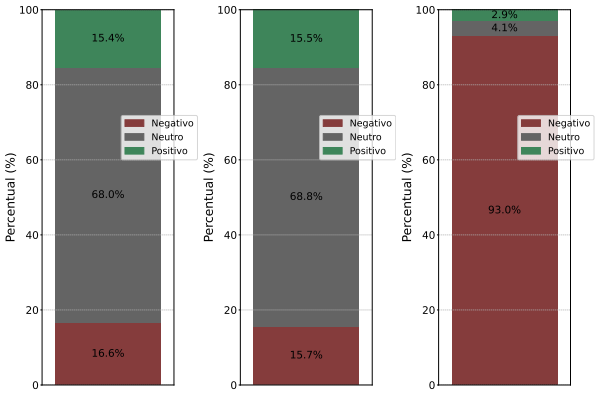
(b) Usuários com mais mensagens

Figura 4: Volume de mensagens enviadas por usuário

Em relação ao sentimento das mensagens, exibido na Figura 5, observa-se que a maior parte dele é neutro, e que os sentimentos positivo e negativo ocorrem em proporção semelhante, reforçando essa tendência de neutralidade. Ao analisarmos isoladamente o sentimento das mensagens não-tóxicas (Figura 5b), temos o mesmo padrão identificado nas mensagens gerais, por se tratar da maior parte dos dados. Por outro lado, há uma predominância do sentimento negativo nas mensagens tóxicas, o que já era esperado e reforça como essas mensagens afetam negativamente as *lives*.

5 O DISCURSO DE ÓDIO NA *TWITCH*

Um dos parâmetros para identificar conteúdo de ódio foi a ocorrência de termos discriminatórios listados pelos dicionários. Para a



(a) Todas as mensagens (b) Não-tóxicas (c) Mensagens tóxicas

Figura 5: Percentual dos Sentimentos das mensagens

Tabela 3: Exemplos de mensagens de discurso de ódio

Termos	Canal	Mensagem
mongoloide	dona	tira esse mongoloide da tela porra
mongol	fontinnelerj	koeeee esse mlq é mongol de maissss
halalification	punccc	the halalification of paul
macaco	dona	DONA MACACO PRETO, CRIOLO
nigga	jasontheween	ban this nigga
monkey, gay	jasontheween	gay little monkey ass pose
gay	dona	ODEIO GAY
puta	dona	Bia puta
puta	ayellol	puta n pensa n rapaz
whore	zackrawrr	she's probably a hoe. we live in a whore culture we almost elected a whore for President

separação de mensagens que continham termos ou sentenças ofensivas, foi elaborada uma filtragem das mensagens que possuíam ao menos um termo com nível de ofensividade 4 ou 5, com base nos léxicos da *Hatebase* e *WeaponizedWord*. Em seguida, foi realizada uma análise manual de algumas mensagens destes arquivos, tendo como resultado o reconhecimento de diversas manifestações de discurso de ódio, caracterizadas por ataques direcionados ao *streamer* ou a outras pessoas que aparecem na *stream*.

Na Tabela 3 estão alguns exemplos de mensagens com discurso de ódio encontradas, os termos do dicionário e o canal em que foram enviadas. Estas são apenas uma seleção das diversas mensagens encontradas nessa etapa, e estão sendo usadas aqui para exemplificar manifestações de xenofobia, racismo, homofobia e misoginia que foram encontradas. Muitas das mensagens que foram encontradas pertencem ao canal do *streamer dona*, onde pode-se observar a ocorrência destes quatro tipos de ofensas. Além de mensagens com ódio, foram encontrados alguns casos de xingamentos e uso de termos chulos. Durante o levantamento, constatou-se que parte destas mensagens, que supostamente continham termos muito ofensivos, não se tratava necessariamente de discurso de ódio. Houve três casos recorrentes de falsos positivos:

- Mensagens em português, com ocorrência de termos em inglês, porém para falantes de português este termo não caracteriza uma ofensa, muito menos discurso de ódio;
- Termos que devido ao contexto da frase não caracterizam discurso de ódio;

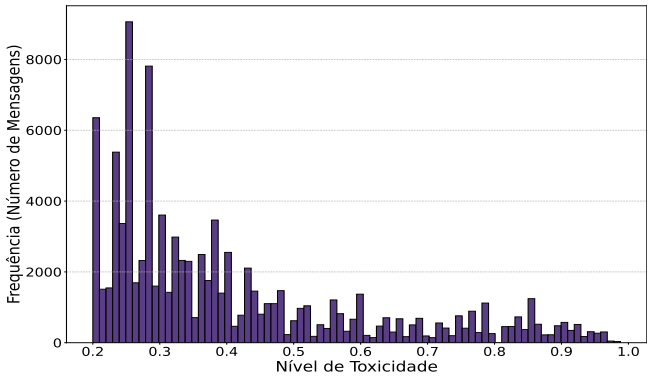


Figura 6: Frequência do índice de toxicidade acima de 0.2

- Frases sem contexto em que apenas a aparição do termo não é suficiente para se concluir se há ou não discurso de ódio.

Com os dados produzidos na etapa de análise de toxicidade utilizando a *PerspectiveAPI* também foi possível avaliar o comportamento tóxico na *Twitch* de cada mensagem. Deste conjunto, observamos que a grande maioria do conteúdo enviado não é tóxico (mais de 98%), com inclusive **84,10%** dos dados possuindo índice de toxicidade abaixo de 0,2. Por outro lado, ao utilizar o índice de toxicidade e filtrar somente as mensagens com valor maior que 0,8, foi obtido um conjunto com **7.294** mensagens, cerca de **1,16%** do total, demonstrando também a presença, ainda que em pequena quantidade, de conteúdo altamente tóxico e potencialmente danoso nas transmissões da *Twitch*, sendo alguns canais com volume muito maior de toxicidade do que outros. A distribuição do índice de toxicidade está exibida na Figura 6 (dados menores que 0,2 foram removidos para melhor visualização).

Na Figura 7 vemos a relação entre a quantidade de mensagens tóxicas e o total de mensagens coletadas, separada pelos top 10 canais com maior porcentagem de toxicidade. Interessantemente, todos os canais mostrados pertencem ao **grupo 3**, formado por canais que atribuíram o marcador como tóxico e cuja atividade justifica essa classificação. Isto revela que esses canais, apesar de não serem tão populares na plataforma, contribuem para criar uma comunidade tóxica na *Twitch*. Neste gráfico, podemos destacar a aparição do *streamer 6arakin* (quinta posição), que logo após a coleta, foi banido

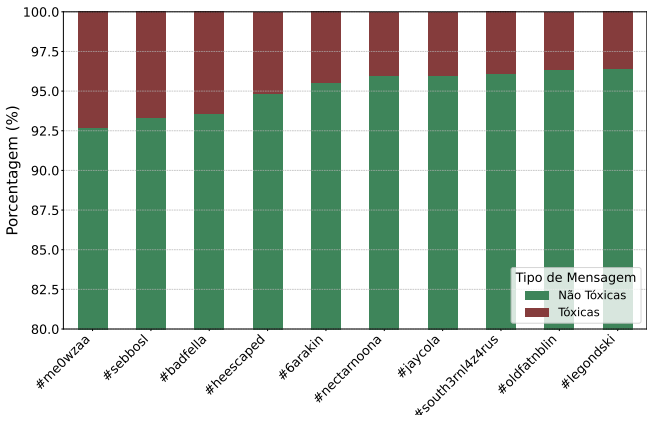


Figura 7: Mensagens Tóxicas x Não-Tóxicas por canal

(a) Tóxicas (EN) (b) Tóxicas (PT) (c) Não-tóxicas (EN) (d) Não-tóxicas (PT)

Figura 8: Nuvens de palavras por tipo de mensagem e idioma

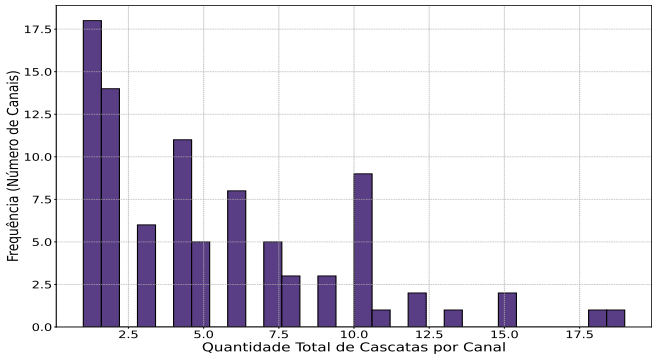
duas vezes da *Twitch* devido ao seu comportamento extremamente tóxico [7]. Outra análise foi realizada ao observar as nuvens de palavras geradas pelas mensagens tóxicas e das não-tóxicas na Figura 8, criadas com a biblioteca *WordCloud* [8]. Nas duas nuvens sem toxicidade, muitos dos termos tratam-se de *emotes* – que são enviados no *chat* como texto, mas são exibidos para os usuários como imagens – ao exemplo do "DinoDance", "KEKW", "Cheergi", "FBBlock", etc. É possível reconhecer algumas palavras que foram muito repetidas durante as *streams*, além de algumas palavras usualmente enviadas por *bots*, como links de outras redes.

Já nas nuvens das mensagens tóxicas, destacam-se ocorrências de muitos palavrões e xingamentos, característicos de uma comunicação mais ofensiva. Entre estas palavras, vemos alguns termos que podem estar diretamente relacionados a discurso de ódio, como "*viado*", "*puta*", "*hate*", "*gay*", entre outras. Palavras que, sem contexto, não se encaixam como ofensivas também ocorreram em menor proporção, como o nome do presidente dos EUA, "Trump" e o nome dos *streamers dona* e *hasanabi* que, em outros momentos desta pesquisa, foram identificados como tendo chats tóxicos.

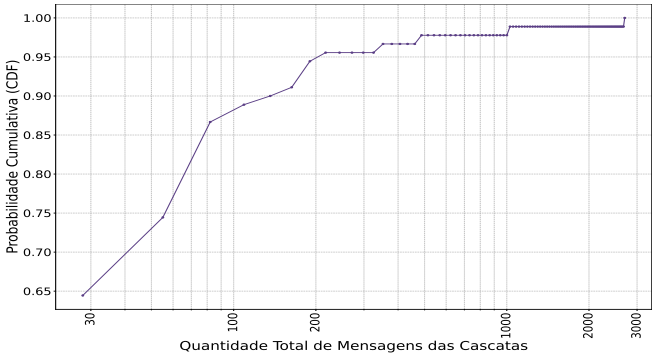
6 ANÁLISE DAS CASCATAS DE TOXICIDADE

A partir da identificação das cascatas de toxicidade, tornou-se possível observar com mais clareza como mensagens tóxicas se propagam nos diferentes canais analisados. Esta etapa da pesquisa concentrou-se na análise das cascatas agrupadas por canal, permitindo a comparação do comportamento coletivo da audiência em cada contexto. Ao todo, foram identificadas **472 cascatas de toxicidade**, que variam em volume, duração e número de participantes, oferecendo indícios sobre padrões de disseminação do discurso tóxico dentro da plataforma.

A Figura 9a apresenta a distribuição da quantidade de cascatas por canal. Observa-se que a maioria dos canais registrou poucas ocorrências com cerca de 18% tendo apenas uma única cascata, e mais de 60% tiveram menos de cinco no total. Ainda assim, é possível identificar alguns canais com comportamento atípico, com mais de 10 cascatas registradas ao longo do período analisado. Já a Figura 9b exibe a função de distribuição acumulada (CDF) da quantidade de mensagens por cascata, evidenciando que aproximadamente 65% das cascatas envolvem menos de 30 mensagens. No entanto, cerca de 5% delas ultrapassam 200 mensagens, indicando episódios



(a) Quantidade de Cascatas por canal



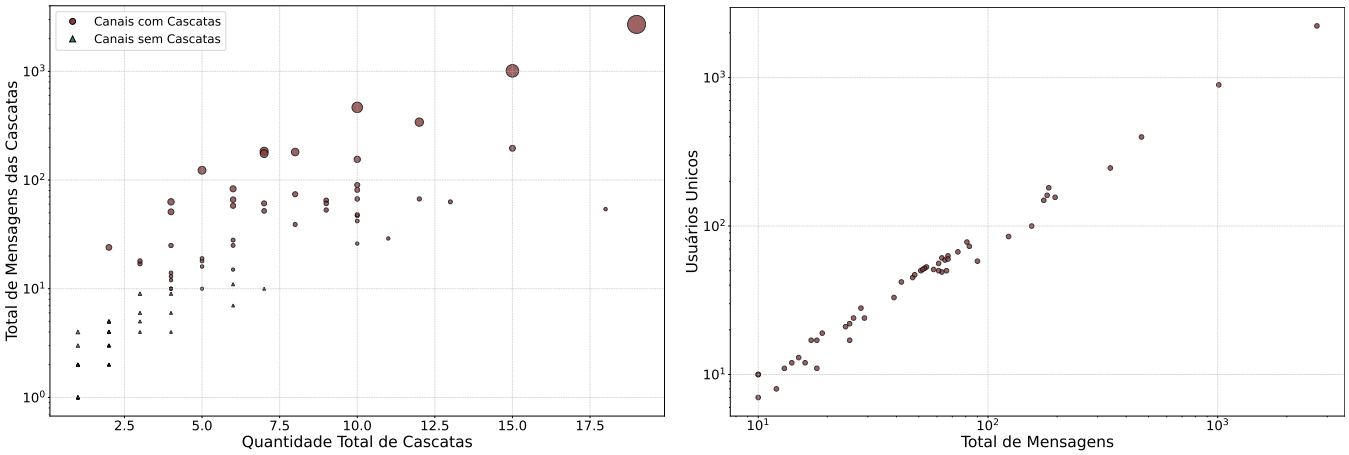
(b) Número de mensagens por cascata

Figura 9: Análise das cascatas de toxicidade na *Twitch*

intensos de toxicidade concentrada em curtos intervalos de tempo. Esses casos extremos sugerem ambientes altamente propícios à propagação de discurso tóxico, e serão aprofundados na análise dos canais mais críticos.

De acordo com os resultados observados nas figuras anteriores, e uma investigação manual das cascatas, foi possível definir valores para definir as cascatas encontradas. Desta forma, foram desconsiderados os canais com menos de 5 cascatas e menos de 10 mensagens, ou cuja média de mensagens por cascata fosse menor que 3, representados pela Figura 10a pela cor verde. Portanto, foi observado que em **45,56%** dos canais com mensagens tóxicas, as cascatas não se propagavam e/ou a ocorrência dessas mensagens foi rara e isolada. Isto ocorre devido à própria cultura do canal, impede

[7] Disponível em: https://streamerbans.com/user/6arakin
[8] Disponível em: https://amueller.github.io/word_cloud/



(a) Relação de canais com e sem cascatas

(b) Cascatas por total de mensagens e usuários únicos

Figura 10: Relação entre canais e ocorrência de cascatas

que o comportamento de um usuário afete os demais, e quando esses casos isolados ocorrem, são controlados pelo *streamer* ou pela moderação ativa de bots ou pelos próprios usuários.

No gráfico da Figura 10b, da relação entre número de mensagens por cascata e número de usuários, destacam-se os dois pontos mais extremos, ambos com mais de **1.000 mensagens tóxicas** cada e com os maiores números de usuários únicos. São os canais dos *streamers hasanabi* e *zackrawrr* que já apareceram anteriormente em outros indicativos de toxicidade, e agora, mostram novamente que seus chats são notoriamente tóxicos. Ambos compartilham a característica de já serem banidos da Twtich por diversas vezes e ambos fazem majoritariamente transmissões de "*Só na conversa*". Em suas transmissões, estes usuários comentam sobre notícias e acontecimentos políticos, onde discutem, muitas vezes em tom mais ofensivo sobre o assunto. O canal do usuário *hasanabi* se destaca como o mais tóxico entre as cascatas e com os maiores números de cascatas identificadas (19), média de 142 mensagens por grupo e cascata com maior número de mensagens (634).

7 CONCLUSÃO

O objetivo deste artigo foi investigar como é o comportamento das mensagens tóxicas e como elas podem se manifestar em cascatas dentro da *Twitch*. Através da análise de 627,237 mensagens de 165 canais diferentes, foi possível identificar este padrão dentro da plataforma, onde havia cascatas com até 634 mensagens tóxicas numa janela de 10 minutos. Com esses agrupamentos, foi possível identificar que, em alguns canais, o comportamento tóxico não desencadeava uma cascata, tratando-se de casos isolados onde os demais usuários não aderiam ao comportamento tóxico dos demais. Outro padrão encontrado foi que alguns canais são mais tendenciosos a conter discurso de ódio e, assim como a toxicidade, estão mais associados a canais de "*Só na conversa*" e transmitem assuntos polêmicos como **jogos de azar** e **política**. Nestes canais, a toxicidade é mais presente e são canais que devem receber uma atenção maior da moderação da plataforma.

Este estudo tem contribuições importantes para o desenvolvimento de estratégias de monitoramento e moderação de conteúdo dentro da *Twitch*. Dentre eles, a definição de parâmetros para a identificação de cascatas utilizada neste artigo pode ser implementada por *bots* e modelos computacionais que identificam toxicidade na plataforma. Outra implicação desta pesquisa é a identificação de canais que sustentam uma cultura tóxica e que ela está presente independentemente da popularidade do canal. Canais que utilizam a *tag* "toxic"têm comunidades menores e culturas mais tóxicas, enquanto canais mais populares, como os dos *streamers hasanabi*, *zackrawrr* e *dona* mostram um volume maior e consistente de toxicidade.

Em trabalhos futuros, sugere-se expandir a seleção de canais coletados e setorizá-los de acordo com sua popularidade, idioma e/ou categorias de transmissão e jogos para melhor entendimento como essas cascatas de ódio se comportam e se diferenciam nestes diferentes cenários dentro da plataforma. Uma outra direção é a realização de uma coleta de dados num período e frequência maior para aumentar a quantidade de dados e, assim, refinar os parâmetros de detecção das cascatas, assim como estudos focados na identificação dos limites de início e final destas cascatas, buscando como separá-las de forma a agrupar somente o conteúdo associado ao mesmo tempo que evitar fragmenta-las em pedaços desconexos.

Além disso, nossos resultados mostram como o conteúdo tóxico tem se espalhado na *Twitch*, em especial na influência de algumas mensagens em desencadear uma série de outros conteúdos de alta toxicidade, evidenciando a necessidade de criação de métodos direcionados para conter sua propagação e barrar conteúdos ofensivos no caso específico das plataformas de transmissão ao vivo. Adicionalmente, propõe-se um estudo semelhante aplicando-se a mesma metodologia, porém, analisando outras plataformas de conteúdo ao vivo, como *TikTok* e *YouTube* para melhor entendimento de como a toxicidade se manifesta em outros ambientes online. Estas direções podem ajudar a promover métodos para controle de conteúdo tóxico nos chats das plataformas de *live stream*.

AGRADECIMENTOS

O presente trabalho foi realizado graças ao apoio da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), projeto APQ-02174-25.

REFERÊNCIAS

[1] Kegan Michael Akinsanmi Ruiz. 2023. *Aplicación de una herramienta de análisis emocional en una plataforma de medios sociales.* B.S. thesis. Universitat Politècnica de Catalunya.

[2] Rhiannon Bevan. 2024. *Asmongold: Twitch streamer suspended after Palestinian rant.* https://www.thegamer.com/asmongold-banned-from-twitch-palestine-racism-controversy/ Notícia sobre suspensão de Asmongold após discurso sobre palestinos.

[3] Jie Cai, Sagnik Chowdhury, Hongyang Zhou, and Donghee Yvette Wohn. 2023. Hate Raids on Twitch: Understanding Real-Time Human-Bot Coordinated Attacks in Live Streaming Communities. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 342 (Oct. 2023), 28 pages. doi:10.1145/3610191

[4] Jie Cai and Donghee Yvette Wohn. 2019. Categorizing live streaming moderation tools: An analysis of twitch. *International Journal of Interactive Communication Systems and Technologies (IJICST)* 9, 2 (2019), 36–50.

[5] Lana Cuthbertson, Alex Kearney, Riley Dawson, Ashia Zawaduk, Eve Cuthbertson, Ann Gordon-Tighe, and Kory W Mathewson. 2019. Women, politics and Twitter: using machine learning to change the discourse. *arXiv preprint arXiv:1911.11025* (2019).

[6] Samuel Guimarães, Gabriel Kakizaki, Philipe Melo, Márcio Silva, Fabricio Murai, Julio C. S. Reis, and Fabrício Benevenuto. 2023. Anatomy of Hate Speech Datasets: Composition Analysis and Cross-dataset Classification. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media* (Rome, Italy) *(HT '23)*. Association for Computing Machinery, New York, NY, USA, Article 33, 11 pages. doi:10.1145/3603163.3609158

[7] Rachel Hall. 2025. *Beyond Andrew Tate: the imitators who help promote misogyny online.* https://www.theguardian.com/media/2025/mar/19/beyond-andrew-tate-the-imitators-who-help-promote-misogyny-online Acesso em 25 de junho de 2025..

[8] Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T. Hancock, and Zakir Durumeric. 2023. Hate Raids on Twitch: Echoes of the Past, New Modalities, and Implications for Platform Governance. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 133 (April 2023), 28 pages. doi:10.1145/3579609

[9] Mohamad Hoseini, Philipe Melo, Manoel Júnior, Fabrício Benevenuto, Balakrishnan Chandrasekaran, Anja Feldmann, and Savvas Zannettou. 2020. Demystifying the Messaging Platforms' Ecosystem Through the Lens of Twitter. In *Proceedings of the ACM Internet Measurement Conference* (Virtual Event, USA) *(IMC '20)*. ACM, NY, USA, 345–359. doi:10.1145/3419394.3423651

[10] Mehdi Kaytoue, Arlei Silva, Loïc Cerf, Wagner Meira, and Chedy Raïssi. 2012. Watch me playing, i am a professional: a first study on video game live streaming. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12 Companion)*. ACM, NY, USA, 1181–1188. doi:10.1145/2187980.2188259

[11] Jaeheon Kim, Donghee Yvette Wohn, and Meeyoung Cha. 2022. Understanding and identifying the use of emotes in toxic chat on Twitch. *Online Social Networks and Media* 27 (2022), 100180. doi:10.1016/j.osnem.2021.100180

[12] Yi Li, Chongli Wang, and Jing Liu. 2020. A systematic review of literature on user behavior in video game live streaming. *International journal of environmental research and public health* 17, 9 (2020), 3328.

[13] Tabitha M London, Joey Crundwell, Marcy Bock Eastley, Natalie Santiago, and Jennifer Jenkins. 2019. Finding effective moderation practices on Twitch. In *Digital ethics*. Routledge, 51–68.

[14] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*. 173–182.

[15] Colten Meisner. 2023. Networked Responses to Networked Harassment? Creators' Coordinated Management of "Hate Raids" on Twitch. *Social Media + Society* 9, 2 (2023), 20563051231179696. doi:10.1177/20563051231179696

[16] Gustavo Nascimento, Manoel Ribeiro, Loïc Cerf, Natália Cesário, Mehdi Kaytoue, Chedy Raïssi, Thiago Vasconcelos, and Wagner Meira. 2014. Modeling and Analyzing the Video Game Live-Streaming Community. In *2014 9th Latin American Web Congress*. 1–9. doi:10.1109/LAWeb.2014.9

[17] Juan Manuel Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2023. pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks. arXiv:2106.09462 [cs.CL]

[18] Nadia Ruiz-Bravo, Lisen Selander, and Maryam Roshan. 2022. The political turn of twitch–understanding live chat as an emergent political space. (2022).

[19] Erik Schleef. [n. d.]. Toxicity in Twitch Chat: Putting Offensive Language and Personal Attacks into Perspective. ([n. d.]).

[20] Joseph Seering and Sanjay R. Kairam. 2022. Who Moderates on Twitch and What Do They Do? Quantifying Practices in Community Moderation on Twitch. *Proc. ACM Hum.-Comput. Interact.* 7, GROUP, Article 18 (Dec. 2022), 18 pages. doi:10.1145/3567568

[21] Prarabdh Shukla, Wei Yin Chong, Yash Patel, Brennan Schaffner, Danish Pruthi, and Arjun Bhagoji. 2025. Silencing Empowerment, Allowing Bigotry: Auditing the Moderation of Hate Speech on Twitch. arXiv:2506.07667 [cs.CL] https://arxiv.org/abs/2506.07667

[22] Statista. 2025. *Twitch – Total de horas assistidas em 2024.* https://www.statista.com/topics/7946/twitch Estatística referente ao total de horas assistidas em 2024.

[23] Christopher Tuckwood. 2017. Hatebase: Online database of hate speech. *The Sentinal Project. Available at: https://www. hatebase. org* (2017).

[24] Jackson Walker. 2025. *Turkish Twitch streamer banned after urging viewers to "kill" Florida senator.* https://thenationaldesk.com/news/americas-news-now/turkish-twitch-streamer-banned-after-urging-viewers-to-kill-florida-senator Reportagem sobre banimento de Hasan Piker na Twitch após incitar espectadores a matarem o senador Rick Scott.