Narrativas de Jogos de Azar em Plataformas de Vídeo: Um Estudo Linguístico-Temático sobre o jogo do Tigrinho no YouTube

Gabriel Prenassi[1], Ana Machado[1], Ester Souza[1], Mateus Brito[1], Davi Reis[1], Jessica Costa[2], Geovana Oliveira[2], Carlos Ferreira[2], Leonardo Rocha[1]
[1]Universidade Federal de São João del-Rei, [2]Universidade Federal de Ouro Preto
(prenassigabriel,anaclaudiamachado211,estermariasouza2005,mateusdeoliveirabritoo,davireisjesus)@aluno.ufsj.edu.br;
(jessica.sc,geovana.so)@aluno.ufop.edu.br;chgferreira@ufop.edu.br;lcrocha@ufsj.edu.br

ABSTRACT

The dissemination of online gambling, particularly when amplified by digital influencers, presents a growing social concern. In Brazil, the game Fortune Tiger has gained popularity through YouTube videos that often promote quick and effortless earnings. While there is existing research on gambling in social media, our study is among the first to delve into how influencers on video platforms construct and disseminate such narratives. We examine the structure and perception of gambling-related content on YouTube, using a large-scale, linguistically informed thematic analysis. Through transcription, topic modeling, and linguistic feature extraction, we identify two dominant narrative types: critical videos that address legal and social harms, and promotional videos that use simplified and repetitive language to encourage engagement. Promotional content overwhelmingly dominates the platform, while critical content exhibits greater linguistic richness and higher per-video engagement. These findings underscore the structural imbalance in online gambling discourse and offer significant implications for future research, platform moderation, and regulatory frameworks.

KEYWORDS

Jogos de azar *online*, *Fortune Tiger*, YouTube, Análise linguística

1 INTRODUÇÃO

Os jogos de azar têm sido historicamente associados a riscos significativos para a saúde pública, com impactos que transcendem o comportamento individual e afetam também núcleos familiares, comunidades e sistemas de apoio social [28, 29]. Em sua vertente digital, esses riscos são potencializados pela acessibilidade e ubiquidade dos dispositivos conectados à *internet* [15, 25]. Estudos recentes indicam que os danos associados ao jogo patológico incluem desde o comprometimento da saúde mental até o endividamento crônico e a exclusão econômica [11, 13, 19]. Um relatório da *Lancet Public Health Commission* (2024) estima que 46,2% dos adultos e 17,9% dos adolescentes globalmente participaram de alguma forma de jogo de azar no último ano [46], evidenciando a escala do fenômeno.

No Brasil, um caso emblemático dessa problemática é o *Fortune Tiger*, informalmente denominado *Jogo do Tigrinho*, um caça-níqueis *online* desenvolvido pela empresa Pocket Games Soft[1]. Com mecânica simplificada, estética chamativa e promessas de ganhos rápidos, o jogo é amplamente promovido por influenciadores digitais em plataformas como YouTube, Facebook e X (antigo Twitter) [5, 10, 24]. Estimativas apontam que brasileiros apostam aproximadamente 20 bilhões por mês em plataformas de jogos de azar digitais[2], gerando preocupações crescentes sobre dependência, impacto econômico e uso indevido de benefícios sociais. O Banco Central do Brasil alertou que beneficiários do principal programa assistencial estariam comprometendo até 20% desses recursos com apostas *online*[3].

A literatura científica já examinou vários temas, incluindo a promoção de jogos de azar em diversas mídias sociais, como Facebook, X (antigo Twitter), Twitch, Reddit, Telegram e YouTube, com foco em temas como estratégias promocionais, engajamento emocional e formação de comunidades [1, 5, 7, 12, 14, 20, 21, 24, 27, 30, 32, 40, 44]. No entanto, poucos estudos investigaram sistematicamente o YouTube, uma das plataformas de vídeo mais acessadas no Brasil [3, 22, 23], cuja arquitetura centrada em criadores de conteúdo, monetização algorítmica e apelo visual favorece a difusão de conteúdos de alto impacto emocional, como jogos de azar. A maioria das análises existentes adota abordagens qualitativas [7, 27]. Já o trabalho de Costa et al. [10] investigou o posicionamento de usuários frente ao *Jogo do Tigrinho* a partir de comentários textuais. Embora relevantes, tais estudos negligenciam o conteúdo audiovisual dos vídeos, componente central na construção e legitimação das narrativas. Uma lacuna ainda pouco explorada reside na análise estrutural, temática e linguística desses vídeos. Especificamente, faltam estudos que explorem o conteúdo do vídeo para compreender como essas narrativas são organizadas por influenciadores no contexto brasileiro. Tal compreensão é fundamental para subsidiar estratégias mais eficazes de mitigação e regulação do conteúdo relacionado a jogos de azar nas plataformas digitais. Com esse propósito, este estudo tem como objetivo investigar como influenciadores constroem e difundem narrativas sobre o *Jogo do Tigrinho* no YouTube, por meio de uma abordagem linguístico-temática que combina técnicas de transcrição automática, modelagem de tópicos e extração de atributos textuais. Assim, o trabalho busca responder às seguintes perguntas de pesquisa:

RQ1: *Como vídeos do YouTube sobre o Fortune Tiger se distribuem tematicamente entre conteúdos promocionais e críticos?*

RQ2: *Quais as diferenças estruturais, linguísticas e de engajamento entre vídeos críticos e promocionais sobre o Jogo do Tigrinho?*

Para responder a **RQ1**, aplicamos técnicas de transcrição automática (*Whisper*) a um conjunto de 1.068 vídeos de alta qualidade,

[1]https://www.pgsoft.com/

[2]https://g1.globo.com/economia/noticia/2024/09/24/brasileiros-gastaram-cerca-de-r-20-bilhoes-por-mes-em-apostas-online-estima-revela-bc.ghtml

[3]https://www.bcb.gov.br/conteudo/relatorioinflacao/EstudosEspeciais/EE119_Analise_tecnica_sobre_o_mercado_de_apostas_online_no_Brasil_e_o_perfil_dos_apostadores.pdf

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

previamente selecionados com base em métricas acústicas e avaliados manualmente. Em seguida, utilizamos o modelo BERTopic para agrupar semanticamente os vídeos, complementando a análise com sumarização automática e extração de palavras-chave. Nossos resultados mostram que os vídeos se organizam majoritariamente em dois grandes tópicos, um com viés promocional, centrado em estratégias, experiências e incentivos à prática, representando 90% do total de vídeos analisados; e outro com viés crítico, abordando denúncias, investigações policiais e consequências sociais negativas do jogo. Para **RQ2**, a partir da separação temática dos vídeos, conduzimos uma análise quantitativa das transcrições com base em métricas de diversidade lexical, densidade de vocabulário, duração do vídeo, visualizações, curtidas e comentários. Extraímos mais de 200 atributos linguísticos com o uso do LFTK (*Linguistic Feature Toolkit*), selecionando os mais discriminativos via *Gini Gain*. Observamos que os vídeos promocionais tendem a ser mais curtos, com vocabulário menos diverso, estrutura repetitiva e linguagem mais simples, características compatíveis com estratégias de engajamento rápido e massivo. Em contraste, vídeos críticos são mais ricos linguisticamente, mencionam entidades específicas (como "polícia", "influenciador", "crime") e apresentam, em média, maior engajamento por vídeo em métricas como curtidas e comentários.

Os resultados de nosso estudo reforçam a necessidade de investigações que possam integrar análises multimodais (áudio, imagem e texto), bem como estudos sobre a eficácia de mecanismos de moderação e classificação automatizada de conteúdo sensível.

2 TRABALHOS RELACIONADOS

A literatura sobre jogos de azar *online* tem crescido significativamente nos últimos anos, impulsionada pela digitalização acelerada do setor e por eventos como a pandemia de COVID-19 [26, 39]. Esse cenário ampliou os desafios relacionados à regulação, à publicidade e aos riscos sociais associados à atividade [41, 43]. Uma primeira vertente de estudos investiga o comportamento dos usuários e a comunicação nas plataformas digitais. Van Der Maas e Samuel [43] analisaram milhares de comentários no Reddit por meio de técnicas de Processamento de Linguagem Natural (NLP), identificando padrões distintos entre postagens com conteúdos positivos (associados a vitórias) e negativos (relacionados a operadores considerados ilegítimos). Em um estudo de escopo nacional, Smith et al. [39] aplicaram *web scraping* em um fórum alemão especializado, revelando aumento de novos usuários e migração para jogos digitais durante o *lockdown* da pandemia. Já Singer et al. [37] examinaram mais de 30.000 tuítes de contas de operadores de jogos na Alemanha, destacando o uso limitado do Twitter como canal de *marketing* e a ausência de mensagens relacionadas ao jogo responsável.

Outra linha de pesquisa foca no desenvolvimento de sistemas para identificação de práticas ilícitas. Teppap et al. [41] propuseram um sistema para detectar anúncios de jogos de azar ocultos em sites universitários tailandeses, utilizando a biblioteca *BeautifulSoup* com acurácia de 89%. De forma semelhante, Harahap e Ridho [18] aplicaram técnicas de detecção de intrusão em domínios universitários, combinando regras baseadas em TF-IDF e teste qui-quadrado, com precisão de 97%. No campo da mineração de dados, Chen et al. [8] usaram o algoritmo DBSCAN para agrupar mais de 23.000 sites de apostas e analisar fluxos financeiros associados a essas plataformas.

Aspectos éticos e regulatórios também têm sido abordados. Cemiloglu et al. [6] discutem requisitos éticos para o *design* de tecnologias viciantes, tendo os jogos de azar como estudo de caso. Outros trabalhos exploram os desafios regulatórios da operação de serviços *offshore* [26] e o uso de dados rotineiramente coletados como ferramenta para políticas públicas baseadas em evidência [4].

No que se refere ao YouTube, poucos estudos abordam de forma sistemática sua relação com a promoção de jogos de azar. Chamil et al. [7] realizaram uma análise qualitativa dos impactos dos jogos *online* sobre a saúde mental durante a pandemia, utilizando vídeos da plataforma. Kroon [27], por sua vez, examinou anúncios no YouTube, destacando o uso de técnicas multimodais na construção de narrativas socioculturais sobre jogos. No contexto brasileiro, Costa et al. [10] analisaram vídeos relacionados ao *Fortune Tiger*, focando em comentários textuais para modelar a postura dos usuários. Seus resultados indicaram predominância de comentários favoráveis, com forte presença de padrões repetitivos e potencial automação.

Apesar dessas contribuições, a maioria dos estudos permanece centrada em comentários textuais ou campanhas pontuais, com foco limitado a recortes discursivos ou períodos curtos. Análises que considerem o conteúdo audiovisual de forma estruturada, quantitativa e em larga escala ainda são raras, especialmente no contexto brasileiro. Faltam, em particular, investigações que combinem transcrição automática, modelagem de tópicos e extração de atributos linguísticos para compreender como influenciadores constroem e difundem narrativas, promocionais ou críticas, sobre jogos de azar em plataformas de vídeo. Este trabalho busca preencher essa lacuna por meio de uma *análise linguístico-temática* de vídeos sobre o *Fortune Tiger* no YouTube, com foco na caracterização estrutural, discursiva e de engajamento do conteúdo produzido por criadores brasileiros.

3 METODOLOGIA

Esta seção descreve as etapas metodológicas empregadas neste estudo, estruturadas em **três etapas principais**. **Primeiro**, descrevemos o processo de coleta e pré-processamento de dados, detalhando como vídeos, canais e comentários do YouTube foram recuperados e preparados para análise. **Segundo**, descrevemos os processos de classificação automática dos vídeos coletados, de acordo com sua qualidade e transcrição textual. **Terceiro**, descrevemos o processo de Modelagem de Tópicos utilizado para separar os vídeos conforme seus temas, bem como os detalhes relacionados à caracterização estrutural e linguística dos vídeos de acordo com os tópicos abordados.

3.1 Coleta e Pré-processamento de Dados

Este trabalho se baseia na base de dados utilizada no trabalho [10]. Em resumo, a coleta de dados foi conduzida por meio da API do YouTube V3[4], abrangendo o período de janeiro de 2023 a julho de 2024. O foco esteve em vídeos relacionados ao jogo *Fortune Tiger*, abrangendo vídeos tradicionais e vídeos do formato *Shorts*. Para delimitar o escopo geográfico e linguístico, restringiu-se a busca à região do Brasil (código ISO 3166-1: BR) e ao idioma português, utilizando consultas específicas que capturam diferentes contextos associados ao jogo. Para cada mês do intervalo analisado, foram selecionados os 50 vídeos mais relevantes com base no algoritmo de ranqueamento do YouTube, visando balancear abrangência temporal e viabilidade computacional, permitindo capturar tendências mensais

[4] https://developers.google.com/youtube/v3/getting-started

sem sobrecarregar as etapas analíticas subsequentes. A ordenação por relevância segue o *ranking* nativo do YouTube, que prioriza vídeos com maior visibilidade e engajamento, oferecendo uma aproximação prática da exposição efetiva dos usuários àquele conteúdo.

Devido à ambiguidade de termos como "Tigrinho", que frequentemente remete a conteúdos sobre animais, entretenimento infantil ou esportes, foram definidos critérios para restringir a coleta a vídeos com títulos contendo expressões específicas: "Fortune Tiger", "Jogo do Tigrinho" e "Tigrinho". Ainda assim, foi necessária uma filtragem manual para eliminar vídeos irrelevantes, com base de termos nos títulos associados a conteúdos não relacionados ao jogo, como: *filhote*, *animal*, *selva*, *moto*, *desenho*, *king*, *leão*, *esportes* e *infantil*.

Realizamos o *download* dos áudios dos vídeos coletados, utilizando a ferramenta *open-source yt-dlp*, um *fork* do *youtube-dl*, o qual extrai conteúdos via protocolos *DASH* e *HLS* [9]. Durante esse processo, alguns vídeos não foram baixados por dois motivos: (i) foram publicados com restrição de idade, exigindo *login* para acesso; ou (ii) foram privados ou removidos da plataforma após a coleta inicial dos dados. Com isso, a base final analisada compreende 3.983 canais, totalizando 7.587 vídeos distintos e 60.086 usuários únicos, dos quais foi possível coletar um total de 1.956 áudios. Além dessas quantidades, também foram coletadas métricas de engajamento, como visualizações e curtidas, que foram utilizadas nas análises descritas na seção de resultados.

3.2 Filtragem pela Qualidade dos Áudios

Um de nossos objetivos é realizar uma análise de vídeos relacionados à promoção do *Fortune Tiger* no Brasil, mais especificamente, caracterizar as mensagens discutidas nesses vídeos, extraídas por meio de estratégias de transcrição dos seus respectivos áudios. A qualidade dessa transcrição depende da qualidade dos áudios (e.g. áudios muito ruidosos tendem a ser transcritos de forma equivocada), que, por sua vez, impacta diretamente na qualidade das análises que objetivamos realizar neste trabalho. Assim, elaboramos uma estratégia para realizar uma filtragem dos áudios coletados, mantendo aqueles que permitissem boas transcrições.

Primeiramente, utilizamos a biblioteca *Librosa* [5] para extrair uma série de características dos áudios coletados relacionadas à qualidade acústica dos mesmos: (i) duração total do áudio; (ii) frequência de amostragem, que indica o nível de preservação dos detalhes sonoros; (iii) amplitude média, usada para avaliar a captação e possíveis problemas de volume; (iv) amplitude máxima, útil para identificar picos sonoros; (v) razão entre a amplitude média e a máxima, que reflete a consistência do sinal ao longo do tempo; (vi) relação sinal-ruído (SNR), que quantifica a presença de ruído em relação ao conteúdo falado; (vii) taxa de cruzamento por zero (*Zero Crossing Rate*), que diferencia fala, ruído e silêncio com base na polaridade do sinal; e (viii) detecção de regiões de silêncio, que permite identificar pausas naturais, hesitações ou falhas na gravação.

Removemos todas as amostras com valores nulos em qualquer métrica, resultando em 1.930 áudios válidos para as análises de classificação. As ausências ocorreram apenas nas métricas *Razão entre Amplitude Média e Máxima* e *Relação Sinal-Ruído (SNR)*, sempre em amostras com sinal nulo (vetores com zeros), o que inviabiliza seus cálculos. Nessas situações, a amplitude máxima é zero, tornando a razão indefinida; da mesma forma, a SNR não pode ser calculada sem energia nas regiões de sinal ou silêncio. A exclusão dessas instâncias visou assegurar a consistência e comparabilidade das análises.

Após a extração das características acústicas, os áudios foram avaliados quanto ao seu potencial de gerar transcrições de qualidade. Inicialmente, os áudios foram agrupados com base em suas características, utilizando o *K-Means*. Os dados foram previamente normalizados com *StandardScaler* para garantir que todas as variáveis contribuíssem de forma equitativa para a formação dos grupos. Ambos os algoritmos foram utilizados com seus parâmetros padrão. A determinação do número ótimo de *clusters* (k) foi realizada por meio de uma variação do método do cotovelo, utilizando o índice *BetaCV* [33], definido como a razão entre a distância média intra-*cluster* e a distância média inter-*cluster*. Foram testados valores de k de 5 a 100, com incrementos de 5, e o ponto de inflexão da curva do *BetaCV* foi identificado como o valor ideal, ilustrado na Figura 1. A curvatura da função *BetaCV* foi analisada por derivadas de primeira e segunda ordem. O ponto de máxima curvatura, identificado pelo mínimo da segunda derivada, resultou em $k = 35$ como valor ótimo.

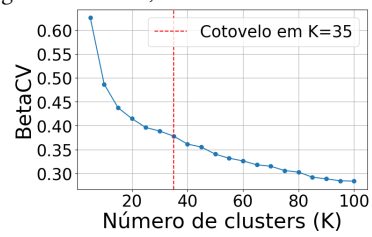

Figura 1: Determinação do número ideal de *clusters* (k).

Foi selecionada uma amostra aleatória de 100 áudios, proporcional ao tamanho dos *clusters* e limitada a áudios com até 2 minutos, visando manter o engajamento dos avaliadores, conforme estudos sobre atenção em formulários *online* [36]. As amostras foram transcritas automaticamente com o modelo *Whisper*[6], versão *Small*, escolhida por sua alta acurácia em português e baixo custo computacional [16]. As transcrições já vêm segmentadas e pontuadas, facilitando análises posteriores. A fidelidade das transcrições foi avaliada manualmente por três avaliadores independentes (cada um analisou no máximo 10 transcrições [36]), que classificaram cada uma delas como "boa" ou "ruim". O rótulo final foi definido pela moda das respostas. A Figura 2 mostra a distribuição dessas avaliações.

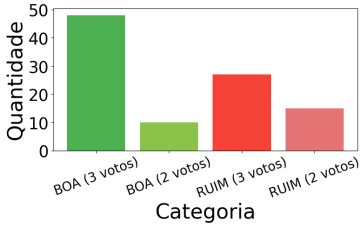

Figura 2: Distribuição das classificações manuais.

Utilizando essa base manualmente anotada, treinamos um classificador de árvore de decisão (ID3) [38], usando entropia como critério de divisão. O modelo foi ajustado com *Grid Search*, validação cruzada estratificada (5 *folds*) e avaliado com *Macro-F1*. Os hiperparâmetros ótimos foram: `class_weight=None`, `max_depth=None`, `max_features=None`, `min_samples_leaf=10`, `min_samples_split=2`,

[5] https://librosa.org/doc/latest/index.html

[6] https://openai.com/pt-BR/index/whisper/

`splitter=best`. Aplicamos o modelo aos 1.930 áudios da coleção, classificando 55% como adequados (bom) e 45% como inadequados (ruim) para transcrição. Para as etapas seguintes, utilizamos apenas as transcrições de áudios adequadas, incluindo as 58 amostras previamente rotuladas como "boas", totalizando 1.068 amostras.

3.3 Análise estrutural e Linguística dos Áudios

Para analisar os textos extraídos, propomos o uso e aplicação de uma série de estratégias relacionadas à área de Processamento de Linguagem Natural (PLN). Na primeira delas, propomos um agrupamento semântico desses textos utilizando estratégias de Modelagem de Tópicos [45]. Para compreender melhor os assuntos abordados em cada tópico, propomos o uso de abordagens de sumarização baseadas em Grandes Modelos de Linguagem (*Large Language Models - LLMs*. Por fim, cada conjunto de textos, relacionado aos tópicos encontrados, extraímos suas características linguísticas que permitem contrastá-los. A seguir, apresentamos um detalhamento dessas etapas.

3.3.1 ***Pré-processamento.*** Essa etapa tem como finalidade eliminar elementos linguísticos que não contribuem semanticamente para a análise e aprimorar a representação dos dados textuais [42]. Utilizamos os mesmos passos propostos e analisados em [34]: 1) lematização; 2) conversão para letras minúsculas e remoção de acentuação; 3) normalização de repetições de caracteres (e.g., *vaaaamos* → *vaamos*); 4) remoção de URLs, pontuação e números; 5) exclusão de palavras com até três letras; 6) remoção de *stopwords*.

3.3.2 ***Modelagem de Tópicos.*** Durante a execução do BERTopic, observou-se empiricamente a formação de diversos agrupamentos pequenos (com menos de cinco transcrições) ou classificados como *outliers* pelo próprio HDBSCAN. De acordo com as diretrizes do BERTopic [17], esses tópicos de baixa representatividade tendem a ser semanticamente frágeis e instáveis. Além disso, a inspeção qualitativa e hierárquica revelou que a maioria desses grupos convergia para dois núcleos temáticos dominantes. Com base nesses achados, optou-se por manter apenas os dois tópicos principais, que concentravam 998 vídeos (93% do total com transcrição adequada).

3.3.3 ***Sumarização dos Tópicos.*** Para obter uma visão representativa de cada grupo, para cada tópico, foram selecionadas as 10 palavras que melhor o descrevem, definidas pelo BERTopic. Além disso, foram escolhidas 10 transições do grupo: três consideradas as mais representativas do tópico, segundo o próprio BERTopic, e outras sete selecionadas aleatoriamente, garantindo que não fossem iguais às três principais. Utilizamos essas informações para sumarizar os tópicos utilizando o ChatGPT (*Prompt* da Figura 3).

3.3.4 ***Extração de Características Linguísticas.*** Esta etapa está ligada diretamente à **RQ2**, ao buscar caracterizar os grupos temáticos da **RQ1** em termos estruturais, linguísticos e de engajamento. O objetivo é aprofundar a compreensão das narrativas dos vídeos com base em atributos textuais e psicométricos. Para isso, extraímos um amplo conjunto de características linguísticas por meio do *Linguistic Feature Toolkit* (LFTK), ferramenta escolhida por sua eficiência na obtenção de mais de 200 atributos e por centralizar um processo analítico robusto, padronizado e escalável. O LFTK fornece métricas que vão desde estatísticas simples (e.g., contagem de palavras, tamanho médio de sentenças, frequência de substantivos) até índices mais complexos, como densidade lexical. Também foram consideradas propriedades psicométricas, como a idade média de aquisição do vocabulário e o nível educacional necessário para compreender o texto. A documentação clara e a organização dos atributos reforçam a confiabilidade e a replicabilidade do processo.

Prompt

```
The input is discussion topic derived from clustered YouTube video transcriptions.
Input:
- top_10_words: The 10 most representative keywords for the topic, extracted from the cluster.
- top_3_transcriptions: The 3 most representative transcriptions (messages) in the cluster.
- sample_transcriptions: 7 additional randomly selected transcriptions from the same topic cluster.
Instructions:
- Carefully analyze the provided content to infer the main theme of the topic cluster.
- Focus on identifying recurring patterns, themes, or subject matter across the transcriptions.
- Do not mention or reference specific videos, speakers, or quotes.
- Use clear, neutral, and concise language.
- Do not invent information not present in the inputs.
_
I will send you the three pieces of information in parts. Please wait to generate the final output until you
have received all three parts.
top_10_words: [top_10_words]
top_3_transcriptions: [top_3_transcriptions]
sample_transcriptions: [sample_transcriptions]
Considering all the provided information, generate the summary following this format:
Título do Tópico: <título do tópico>
Resumo: <resumo do tópico>
The response must be strictly in Portuguese.
```

Figura 3: *Prompt* para sumarizações dos tópicos. Os colchetes indicam onde as informações são inseridas.

Analisar manualmente as 220 características extraídas é excessivamente custoso. Por isso, selecionamos de forma criteriosa os atributos com maior potencial de diferenciar os tópicos identificados nas etapas anteriores. Para isso, ranqueamos os atributos com base na métrica de *Gini Gain* [35] (Equação 1), que avalia diversos limiares de separação e, para cada um, calcula o *Gini Impurity* [47] (Equação 2), que mede a "pureza" dos grupos formados. Quanto maior o *Gini Gain*, melhor é a separação entre tópicos.

$$\text{GiniGain}(A, t) = \text{Gini}(P) - \left(\frac{n_L}{n} \cdot \text{Gini}(L) + \frac{n_R}{n} \cdot \text{Gini}(R)\right) \quad (1)$$

em que A é a característica analisada, t o limiar de separação, P o conjunto completo de instâncias, L e R os subconjuntos com valores menores (ou iguais) e maiores que t, respectivamente, e n, n_L, n_R suas respectivas cardinalidades. A *Gini Impurity* é definida como:

$$\text{GiniImpurity}(S) = 1 - \sum_{i=1}^{k} p_i^2 \quad (2)$$

Com o ranqueamento realizado, selecionamos, ordenadamente, as 10 características com maior valor de *Gini Gain*. Entretanto, com tantos atributos extraídos, pode haver uma similaridade de significado e interpretação desses atributos. Assim, visando mitigar essas questões e obter a maior quantidade possível de diferentes análises sobre os dados, realizamos algumas etapas de filtragem dessas características. Primeiro, calculamos a correlação de *Spearman* [2], par a par, entre essas 10 características. Essa correlação foi escolhida devido ao fato de não ser necessário haver uma distribuição normal dos dados e, além disso, permitir captar correlações não-lineares. Em seguida, partindo da característica com maior valor de *Gini Gain*, eliminamos os atributos que apresentassem correlação maior que 0.7 (considerada alta de acordo com a literatura [2]). Selecionamos o próximo atributo da lista com maior valor de *Gini Gain* que ainda não tivesse sido excluído, e realizamos o mesmo processo de eliminação, sequencialmente, até não haver mais atributos.

Com base nas características mais discriminativas, segmentamos as mensagens pelos tópicos identificados, visando analisar o comportamento dessas variáveis em cada grupo. Calculamos as médias por tópico e, para verificar a significância das diferenças observadas, aplicamos o teste não paramétrico de *Mann-Whitney U* [31]. Também analisamos o engajamento dos vídeos (visualizações, curtidas e comentários), explorando aspectos estruturais adicionais relacionados à **RQ2**. Esses indicadores ajudam a entender o alcance e o impacto dos discursos, evidenciando como conteúdos críticos ou promocionais mobilizam a audiência.

4 RESULTADOS E DISCUSSÕES

Esta seção apresenta os resultados obtidos a partir das análises linguísticas e estruturais das transcrições dos vídeos selecionados. As evidências são discutidas sob duas perspectivas: (i) uma visão global, considerando o conjunto completo de transcrições; e (ii) uma análise segmentada por tópicos, permitindo responder às questões de pesquisa propostas, contrastando padrões linguísticos, estruturais e de engajamento entre os grupos de vídeos.

4.1 Caracterização Global das Transcrições

Nesta etapa, as análises consideraram o conjunto completo de transcrições com qualidade adequada, desconsiderando apenas os textos excluídos no pré-processamento. Foram conduzidas análises quantitativas para compreender as propriedades estruturais dos textos. A primeira avaliação investigou a distribuição do número de palavras por vídeo, ilustrada na Figura 4. Com base nas transcrições originais (sem pré-processamento), a quantidade de palavras por documento variou de 1 a 20.334, com média de 1.182 palavras e desvio padrão de aproximadamente 1.307, evidenciando alta dispersão. A mediana foi de 864 palavras, e 75% das transcrições possuem até 1.611 palavras, o que indica a presença de vídeos consideravelmente mais longos, justificando a diferença entre os valores médio e máximo.

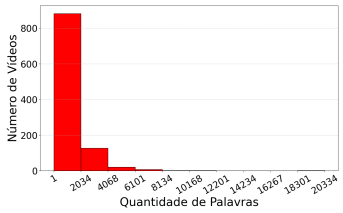

Figura 4: Distribuição de vídeos por quantidade de palavras.

A segunda análise examinou o tamanho do vocabulário de cada vídeo, definido como o número de palavras únicas presentes na transcrição (Figura 5). A média foi de 380 termos distintos por vídeo, com desvio padrão de 314. O valor mínimo foi de apenas uma palavra, enquanto o máximo alcançou 4.516 termos únicos. A mediana foi de 331 palavras, com 75% das transcrições não ultrapassando 519 termos únicos. Além disso, analisou-se a proporção de palavras únicas em relação ao total de palavras do texto. A maioria das transcrições concentrou-se na faixa entre 0,3 e 0,5, indicando que de 30% a 50% das palavras são únicas. Também foram observadas caudas à direita, com valores superiores a 0,6 em textos mais curtos e vocabulário diverso. Por outro lado, valores abaixo de 0,2 caracterizam transcrições longas com alta repetição lexical. Esses dados evidenciam um grau considerável de diversidade lexical no corpus analisado.

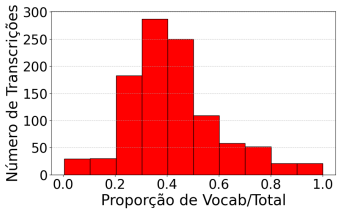

Figura 5: Distribuição do vocabulário por total de palavras.

4.2 Caracterização das Transcrições por Tópico

Na Figura 6, apresentamos os 10 principais termos associados a cada um dos tópicos, em que se observam duas narrativas opostas sobre os jogos de azar. Os tópicos indicam uma divisão temática bem definida entre conteúdos críticos e conteúdos de incentivo ao jogo. O Tópico 0 reúne termos como "jogo", "polícia", "crime", "divulgar", "dinheiro" e "influenciador", o que indica vídeos com viés predominantemente crítico. O vocabulário utilizado sugere uma preocupação com a legalidade dessas práticas, além de destacar o papel de influenciadores digitais na divulgação desses jogos. Já o Tópico 1 é composto por termos como "pagar", "rodada", "jogar", "banca", "estratégia" e "turbo", evidenciando uma linguagem direcionada ao público consumidor de cassinos *online*. Os vídeos desse grupo tendem a focar na dinâmica dos jogos, oferecendo dicas, instruções e estratégias com o objetivo de incentivar a prática do jogo.

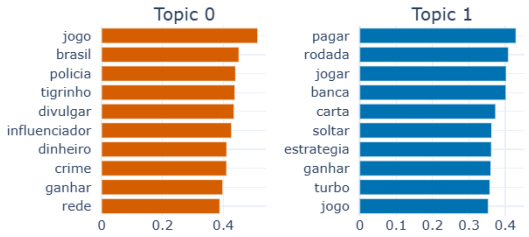

Figura 6: Representação dos tópicos obtidos.

4.2.1 ***Sumarização dos Tópicos.*** As Figuras 7 e 8 apresentam os resumos obtidos por meio do processo de sumarização, conforme detalhado na Seção 3.3.3. O Tópico 0 (Figura 7) enfatiza aspectos criminais e consequências sociais negativas, destacando operações policiais realizadas em diversos estados brasileiros para combater esquemas ilegais promovidos por influenciadores. As ações investigativas revelam crimes como estelionato, lavagem de dinheiro e propaganda enganosa, além de apontarem os danos sociais causados pela exploração da vulnerabilidade emocional e financeira das vítimas.

O Tópico 1 (Figura 8) apresenta o jogo sob a ótica de jogadores e influenciadores que compartilham estratégias, experiências pessoais e conselhos sobre apostas, sugerindo um engajamento mais cotidiano com o jogo. Embora haja tentativas de promover práticas de controle e responsabilidade individual, os vídeos também naturalizam a lógica de ganhos fáceis e incentivam a continuidade das apostas. Essa abordagem transfere a responsabilidade pelos prejuízos exclusivamente ao jogador, invisibilizando o papel das plataformas e dos influenciadores na promoção e manutenção desse ambiente de risco.

Em relação à **RQ1**, a comparação entre os dois tópicos revela uma narrativa dual: **por um lado, um tópico composto por um conjunto de vídeos que expõem uma rede ilícita que se beneficia da manipulação do jogo e da propaganda enganosa,**

abordando denúncias, investigações policiais e consequências sociais negativas do jogo; por outro, um conjunto de vídeos com viés promocional, centrado em estratégias de jogo, experiências e incentivos à prática e que frequentemente desloca a culpa de excessos para quem joga.

Tópico 0

Título do Tópico: Operações Policiais e Crimes Relacionados ao Jogo do Tigrinho
Resumo: Diversas operações policiais em estados como Pará, Alagoas, São Paulo e Paraná revelaram esquemas criminosos envolvendo o jogo do Tigrinho, uma espécie de caça-níquel virtual. Influenciadores digitais são apontados como responsáveis por divulgar o jogo nas redes sociais por meio de contas de demonstração com ganhos fictícios, levando seguidores a apostarem em plataformas ilegais e não regulamentadas no Brasil. Investigações apontam crimes como estelionato, lavagem de dinheiro, ocultação de bens e propaganda enganosa. Muitos influenciadores ostentavam vidas de luxo financiadas por esses ganhos ilícitos, enquanto vítimas relatam perdas financeiras significativas, endividamento e até casos extremos de suicídio. As plataformas geralmente operam fora do país, dificultando ações legais. O impacto social do jogo é agravado pela influência dos criadores de conteúdo, que exploram a vulnerabilidade emocional e financeira de seus seguidores, promovendo falsas promessas de enriquecimento rápido.

Figura 7: Resultado da sumarização para o tópico 0.

Tópico 1

Título do Tópico: Estratégias e Testemunhos sobre o Jogo do Tigrinho (Fortune Tiger)
Resumo: Os vídeos analisados apresentam influenciadores e jogadores compartilhando experiências, estratégias e dicas sobre o jogo Fortune Tiger, popularmente conhecido como "jogo do Tigrinho". As falas destacam métodos como controle de banca, horários e minutos estratégicos para apostar, uso de apostas mínimas ou progressivas ("estratégia da escadinha") e a importância de parar após obter lucro. Alguns usuários promovem plataformas específicas e aplicativos que prometem aumentar a assertividade nas jogadas. Apesar da tentativa de apresentar o conteúdo como entretenimento, os vídeos frequentemente incentivam a prática de jogos de azar e trazem discursos que normalizam os riscos financeiros envolvidos, reforçando a ideia de lucros fáceis e recorrentes.

Figura 8: Resultado da sumarização para o tópico 1.

4.2.2 ***Número de Vídeos por Tópico.*** O número de vídeos classificados em cada um dos tópicos extraídos pela modelagem apresenta uma discrepância significativa entre os grupos. O Tópico 1, associado a conteúdos de incentivo ao jogo, concentra 895 vídeos, enquanto o Tópico 0, com abordagem crítica, reúne apenas 103 vídeos. Essa predominância do Tópico 1 sugere uma maior produção de conteúdos voltados à dinâmica dos cassinos *online*, incluindo dicas, estratégias e estímulos à prática do jogo. Em contrapartida, a menor frequência de vídeos críticos pode refletir um alcance reduzido na plataforma, apesar de sua relevância para o debate social. Cabe destacar que os dois tópicos analisados abrangem 998 dos 1.068 vídeos com transcrição adequada, correspondendo a aproximadamente 93% do corpus total. Tal cobertura garante representatividade suficiente para os objetivos analíticos propostos na **RQ1**.

4.2.3 ***Duração dos Vídeos.*** A análise da duração dos vídeos (Figura 9) acrescenta uma dimensão relevante à distinção entre os grupos temáticos. No Tópico 0, de caráter crítico, a maioria dos vídeos concentra-se em durações inferiores a 500 segundos (cerca de 8 minutos), compatíveis com o formato de notícias rápidas. Ainda assim, há *outliers* com durações superiores a 7.000 segundos (aproximadamente 2 horas), indicando a presença de conteúdos extensos, como documentários. Por outro lado, o Tópico 1, voltado ao incentivo ao jogo, apresenta uma distribuição mais homogênea, com concentração predominante em vídeos de curta a média duração, geralmente abaixo de 600 segundos (aproximadamente 10 minutos). Essas diferenças reforçam que os tópicos se distinguem não apenas pelo conteúdo, mas também pelos formatos de veiculação. O Tópico 0 alterna entre formatos breves e longos, coerentes com abordagens informativas ou investigativas. Já o Tópico 1 tende a adotar uma estrutura estável, com vídeos mais curtos e repetitivos, alinhados a estratégias de engajamento rápido. Assim, a análise de duração contribui para a caracterização estrutural dos conteúdos e se articula diretamente com a **RQ2**, ao evidenciar como distintas formas de produção se associam às narrativas críticas ou promocionais presentes nos dados.

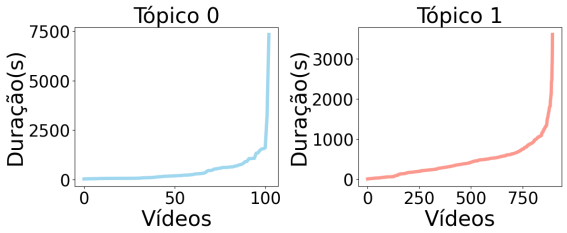

Figura 9: Duração dos vídeos por tópico.

4.2.4 ***Diversidade e Densidade do Vocabulário por Tópico.*** Para estimar a complexidade linguística dos discursos e identificar padrões de repetição lexical, analisamos a diversidade e a densidade do vocabulário em cada grupo temático. A diversidade foi calculada pela razão entre o número de palavras únicas e o total de palavras em cada transcrição. A Figura 10(a) evidencia uma clara diferença entre os tópicos. O Tópico 0 apresenta índices de diversidade mais elevados, concentrando-se entre 0,5 e 0,9, o que indica um vocabulário variado e compatível com a complexidade dos temas abordados, como crimes, investigações e aspectos legais, exigindo maior riqueza lexical para descrever os diferentes contextos. Já o Tópico 1 concentra-se entre 0,4 e 0,5, refletindo uma construção textual mais repetitiva e padronizada. Transcrições muito curtas, com poucas palavras e quase nenhuma repetição, aparecem como ***outliers*** com diversidade próxima de 1,0, mas esse valor elevado não necessariamente representa maior sofisticação linguística.

De forma complementar, a densidade lexical, representada na Figura 10(b), foi calculada como o inverso da diversidade, isto é, a razão entre o total de palavras e o número de palavras únicas. Valores próximos de 1,0 sugerem um vocabulário variado, enquanto valores mais altos indicam maior repetição lexical. O Tópico 0 demonstra baixa densidade (entre 1,0 e 2,0), reforçando seu caráter discursivamente mais complexo. O Tópico 1, por sua vez, revela um padrão bimodal: de um lado, vídeos curtos com baixa densidade; de outro, conteúdos com densidade extremamente alta (valores acima de 40), característicos de narrativas centradas em termos repetitivos, como no caso de transmissões de jogabilidade e tutoriais baseados em jargões e frases de efeito ("vamos girar", "olha o bônus", "grande ganho").

Esses resultados indicam que os conteúdos promocionais operam com uma estrutura linguística menos variada e, frequentemente, centrada em fórmulas repetitivas, enquanto os conteúdos críticos apresentam maior diversidade e sofisticação textual, alinhando-se ao padrão observado nas demais análises estruturais.

4.2.5 ***Visualizações dos Vídeos.*** A análise quantitativa das visualizações (Figura 11(a)) revela que, embora o Tópico 1 concentre um volume significativamente maior de vídeos, o Tópico 0 apresenta maior média de visualizações por vídeo (21.736 contra 10.461). No entanto, como a média é sensível a valores extremos, essa discrepância pode estar inflacionada por vídeos com grande audiência.

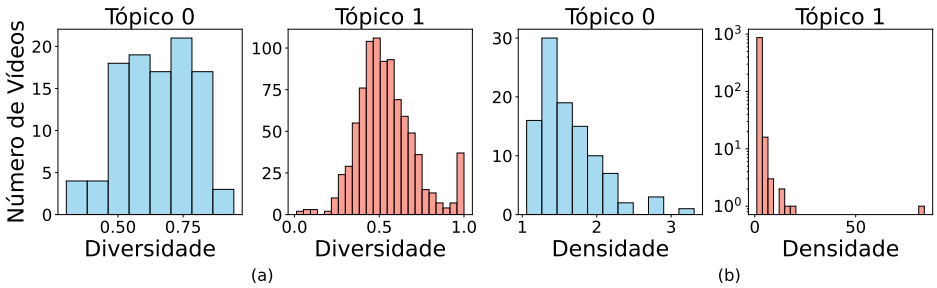

Figura 10: Diversidade do vocabulário por tópico (à esquerda) e densidade (à direita).

Para avaliar o desempenho típico dos vídeos, a mediana fornece uma medida mais robusta. Sob essa perspectiva, o Tópico 0 registra uma mediana de 6.469 visualizações, enquanto o Tópico 1 apresenta uma mediana bastante inferior, de apenas 228 visualizações. Esses valores indicam realidades distintas: a maioria dos vídeos críticos alcança níveis relevantes de audiência, enquanto a maior parte dos vídeos promocionais atrai pouco público, com metade deles sendo assistidos por menos de 228 usuários. Esse cenário de baixo desempenho está associado a estratégias de canais que publicam vídeos curtos com alta frequência, resultando em um volume massivo de conteúdo com baixo engajamento individual. Ainda assim, o Tópico 1 contém vídeos que viralizam com grande intensidade, atingindo até 1.500.000 visualizações, número muito superior ao vídeo mais assistido do Tópico 0, que alcança 309.000 visualizações. Essa dispersão é evidenciada pelo elevado desvio padrão das visualizações no Tópico 1 (75.365), valor aproximadamente sete vezes maior que sua média, sugerindo forte assimetria e grande variabilidade no alcance dos vídeos promocionais.

4.2.6 ***Curtidas dos Vídeos.*** A análise do número de curtidas por vídeo (Figura 11(b)) reflete o padrão de engajamento já observado nas visualizações. O Tópico 0 apresenta uma média de 1.560 curtidas por vídeo, com mediana de 241, enquanto o Tópico 1 possui média de 511 curtidas e mediana de apenas 17. Esses dados sugerem que os vídeos críticos tendem a gerar um engajamento mais consistente por parte dos usuários. Apesar do desempenho mediano reduzido no Tópico 1, os vídeos mais populares desse grupo atingem picos de curtidas mais elevados, alcançando até 84 mil, valor superior ao máximo de 46 mil curtidas observado no Tópico 0. O elevado desvio padrão do Tópico 1 (3.896), substancialmente maior que sua média, indica uma distribuição altamente desigual. Nota-se que, enquanto a maioria dos vídeos recebe pouca ou nenhuma aprovação, poucos conteúdos virais concentram grande volume de curtidas, superando inclusive os níveis máximos do conteúdo crítico.

4.2.7 ***Comentários dos Vídeos.*** Em relação ao número de comentários (Figura 11(c)), o Tópico 0 apresenta maior engajamento, com média de 108 comentários por vídeo, superando os 60 comentários do Tópico 1. A mediana reforça essa diferença: 26 comentários no Tópico 0, contra apenas 5 no Tópico 1. Esses dados indicam que o conteúdo crítico tende a ser mais eficaz na mobilização do público. Apesar disso, o vídeo com maior número de comentários está no Tópico 1, com mais de 4.400 interações, valor significativamente superior ao pico de 1.454 comentários registrado no Tópico 0. Essa assimetria sugere que, embora a maioria dos vídeos de incentivo gere pouco debate, uma parcela muito pequena de conteúdos virais alcança altos níveis de engajamento, distorcendo a média. Assim como nas demais métricas de engajamento, essa análise contribui para uma compreensão mais ampla das manifestações discursivas em cada grupo, compondo a resposta à **RQ2** ao lado dos indicadores linguísticos explorados na próxima subseção.

4.2.8 ***Características Linguísticas.*** A Tabela 1 apresenta as 10 características linguísticas com maior valor de *Gini Gain*, além de uma breve descrição de cada atributo. Observa-se que algumas características fornecem informações complementares entre si, enquanto outras apresentam redundância. Por esse motivo, aplicou-se um processo de seleção com o objetivo de mitigar sobreposição semântica entre atributos. Como resultado, quatro características com maior capacidade discriminativa entre os Tópicos 0 e 1 foram selecionadas (destacadas em cinza na tabela).

As quatro características finais selecionadas foram: (i) **Diversidade de Vocabulário (não linear)** — métrica que aplica uma transformação não linear para captar variações na variedade lexical; (ii) **Escolaridade** — nível educacional estimado necessário para compreensão do texto; (iii) **Proporção de Substantivos Próprios** — frequência relativa de nomes próprios, como pessoas, lugares e organizações; e (iv) **Diversidade de Vocabulário (linear)** — proporção direta entre vocabulário único e total de palavras. A análise conjunta dessas *features* oferece múltiplas perspectivas linguísticas e estruturais sobre os conteúdos. Na Tabela 2, são apresentados os valores médios dessas características para cada tópico. Valores acompanhados por um $*$ indicam diferença estatisticamente significativa, conforme o teste de *Mann-Whitney U* aplicado na Seção 3.3.4.

Conforme os resultados, todos os valores médios associados ao Tópico 1 (conteúdos de incentivo ao jogo) são significativamente inferiores, indicando construções linguísticas mais simples e padronizadas. No caso da **diversidade de vocabulário** (linear e não linear), observa-se uma repetição frequente de termos, reforçando a hipótese de homogeneidade textual nas mensagens promocionais. A versão não linear da métrica, em especial, evidencia com maior sensibilidade essas diferenças, justificando sua inclusão. A análise da **escolaridade** sugere que os vídeos do Tópico 1 exigem menor nível de instrução para compreensão, tornando-os mais acessíveis a públicos com menor escolaridade. Essa acessibilidade pode potencializar o impacto social desses conteúdos, sobretudo em audiências vulneráveis. A **proporção de substantivos próprios** também se destaca como um fator discriminativo: os vídeos críticos (Tópico 0) fazem uso mais intenso de nomes de instituições, figuras públicas

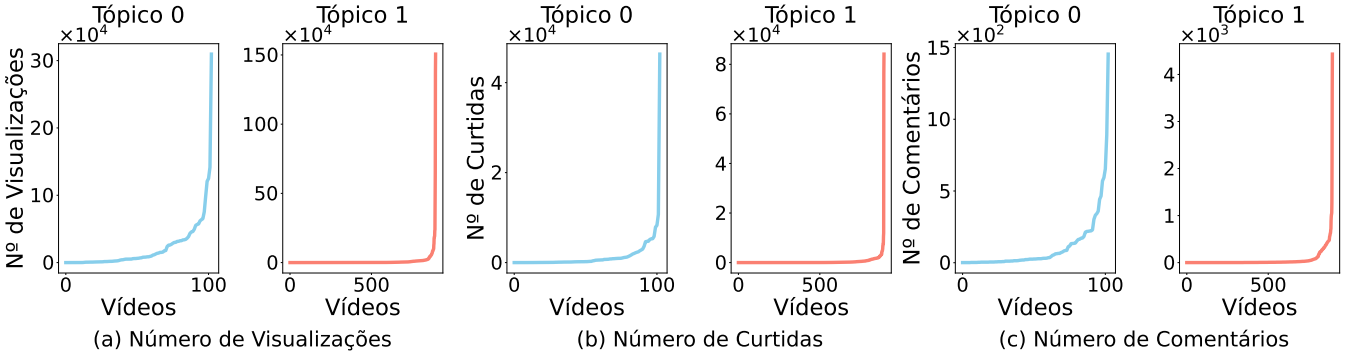

Figura 11: Número de visualizações por tópico (à esquerda), curtidas (no centro) e comentários (à direita).

e entidades, como "polícia", "influenciadores" e "Brasil", reforçando o teor investigativo e político desses conteúdos.

Característica	Descrição
uber_ttr_no_lem **(Diversidade do Vocabulário (mais complexo))**	Analisa a diversidade do vocabulário de um texto, mas utilizando logs, permitindo a análise de forma mais complexa, dado que a diversidade do vocabulário não cresce, necessariamente, de forma linear (não usa termos lematizados)
uber_ttr	Analisa a diversidade do vocabulário de um texto, mas utilizando logs, permitindo a análise de forma mais complexa, dado que a diversidade do vocabulário não cresce, necessariamente, de forma linear (usa termos lematizados)
cole **(Escolaridade)**	Índice que indica a escolaridade necessária para se compreender um texto
a_char_pw	Número médio de caracteres por palavra
a_syll_pw	Número médio de sílabas por palavra
bilog_ttr_no_lem	Analisa a diversidade do vocabulário de um texto, mas utilizando logs, permitindo a análise de forma mais complexa, dado que a diversidade do vocabulário não cresce, necessariamente, de forma linear. (Não utiliza a diferença dos valores de log entre o tamanho do texto e do vocabulário, e não utiliza termos lematizados)
bilog_ttr	Analisa a diversidade do vocabulário de um texto, mas utilizando logs, permitindo a análise de forma mais complexa, dado que a diversidade do vocabulário não cresce, necessariamente, de forma linear. (Não utiliza a diferença dos valores de log entre o tamanho do texto e do vocabulário, e utiliza termos lematizados)
corr_propn_var **(Proporção de substantivos)**	Proporção entre a quantidade de substantivos próprios únicos e o tamanho total de um texto
root_propn_var	Proporção entre a quantidade de substantivos próprios únicos e o tamanho total de um texto
corr_ttr_no_lem **(Diversidade do vocabulário (menos complexo))**	Analisa a diversidade do vocabulário de um texto, mas analiza de forma linear (denominador é a raíz quadrada de 2 vezes o tamanho do total do texto)

Tabela 1: Características mais discriminativas.

Tópico	Diversidade de vocabulário (não linear)	Escolaridade	Proporção de substantivos	Diversidade de vocabulário (linear)
0	57.048	25.257	5.752	6.327
1	32.991*	21.649*	4.747*	5.207*

Tabela 2: Valores das características linguísticas para transcrições favoráveis e contrárias.

Esses achados complementam as análises anteriores e reforçam a robustez da segmentação temática realizada. Todos os resultados, analisados em conjunto, respondem à **RQ2, demonstrando que as narrativas promocionais e críticas sobre o *Fortune Tiger* no YouTube divergem significativamente em conteúdo, estrutura linguística e engajamento. As narrativas favoráveis ao jogo são menos complexas e mais genéricas do ponto de vista linguístico, enquanto os conteúdos críticos apresentam maior riqueza lexical e sofisticação discursiva.**

5 CONCLUSÕES E TRABALHOS FUTUROS

Este estudo debruçou-se sobre a produção audiovisual no YouTube acerca do jogo *Fortune Tiger*, amplamente reconhecido como "Jogo do Tigrinho", inserido em um cenário de crescente inquietação social diante da disseminação de conteúdos que promovem práticas associadas a jogos de azar em plataformas digitais. Embora o fenômeno já tenha sido objeto de investigações anteriores, identificamos uma lacuna metodológica substancial no que tange à análise estruturada e linguística de tais conteúdos, especialmente sob a perspectiva da atuação de influenciadores brasileiros e em escala ampla, com o suporte de métodos quantitativos robustos.

Com o intuito de suprir essa lacuna, delineamos uma abordagem linguístico-temática que integra transcrição automática, modelagem de tópicos e extração sistemática de atributos textuais. Essa estratégia analítica permitiu mapear a distribuição narrativa dos vídeos entre discursos promocionais e críticos, bem como elucidar suas distinções estruturais e discursivas de maneira sistemática e replicável. Os achados evidenciam uma predominância significativa de vídeos com viés promocional, marcados por reduzida diversidade lexical, estruturas reiterativas e baixa complexidade linguística — características frequentemente associadas a estratégias de engajamento superficial e de rápida absorção. Em contraposição, os vídeos de cunho crítico demonstraram maior sofisticação textual, com vocabulário mais elaborado, riqueza expressiva e níveis superiores de engajamento, indicando um discurso mais reflexivo e investigativo.

Como trabalhos futuros, propomos a ampliação da análise para uma perspectiva multimodal, que contemple simultaneamente os elementos visuais, sonoros e textuais do conteúdo. Ademais, sugerimos investigações sobre a efetividade de mecanismos automatizados de moderação e rotulagem de conteúdos sensíveis, bem como reflexões aprofundadas sobre os dilemas éticos e os desafios regulatórios decorrentes da promoção de jogos de azar em ambientes digitais, sobretudo em contextos de vulnerabilidade social.

AGRADECIMENTOS

Trabalho financiado por CNPq, INCT-TILD-IAR, Fapemig e AWS.

REFERÊNCIAS

[1] Brett Abarbanel and Mark R Johnson. 2020. Gambling engagement mechanisms in Twitch live streaming. *International Gambling Studies* (2020).

[2] Khawla Ali Abd Al-Hameed. 2022. Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications* 13, 1 (2022), 3249–3255.

[3] Mariana Arantes, Flavio Figueiredo, and Jussara M Almeida. 2016. Understanding video-ad consumption on YouTube: a measurement study on user behavior, popularity, and content properties. In *ACM Conference on Web Science*.

[4] Pippa Boering, Matthew Jones, Kishan Patel, Daniel Leightley, and Simon Dymond. 2025. A scoping review of routinely collected linked data in research on gambling harm. 8, 1 (2025).

[5] Alex Bradley and Richard JE James. 2019. How are major gambling brands using Twitter? *International Gambling Studies* (2019).

[6] Deniz Cemiloglu, Emily Arden-Close, Sarah Hodge, Theodoros Kostoulas, Raian Ali, and Maris Catania. 2020. Towards Ethical Requirements for Addictive Technology: The Case of Online Gambling. In *2020 1st Workshop on Ethics in Requirements Engineering Research and Practice (REthics)* (2020-08). 1–10. doi:10.1109/REthics51204.2020.00007

[7] Andika Yulianto Chamil, Safiera Amanda Djuanda, and Nurina Septaviana. 2024. A Comprehensive Communication Approach to Navigate the Crisis Caused by Online Gambling: Insights from Kemencast# 44 on Youtube. *Ilomata International Journal of Social Science* (2024).

[8] Jian Chen, Shenao Zheng, Yanan Cheng, and Zhaoxin Zhang. 2024. Data mining based analysis of online gambling sites and illicit financial flows. In *Proceedings of the 2024 International Conference on Cloud Computing and Big Data* (New York, NY, USA, 2024) *(ICCBD '24)*. Association for Computing Machinery, 205–211. doi:10.1145/3695080.3695116 event-place: Dali, China.

[9] Steven Coats. 2023. A pipeline for the large-scale acoustic analysis of streamed content. In *International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2023)*.

[10] Jessica Costa, Geovana Oliveira, Guilherme Fonseca, Davi Reis, Giancarlo Oliveira Teixeira, Washington Cunha, Leonardo Rocha, and Carlos H. G. Ferreira. 2025. Characterizing YouTube's Role in Online Gambling Promotion: A Case Study of Fortune Tiger in Brazil. In *Proceedings of the 17th ACM Web Science Conference 2025*. 42–51. doi:10.1145/3717867.3717905

[11] Nathan Critchlow, Martine Stead, Crawford Moodie, Phillip WS Purves, Gerda Reith, Amber Morgan, and Fiona Dobbie. 2019. The Effect of Gambling Marketing and Advertising on Children, Young People and Vulnerable People. (2019).

[12] Saul Sousa da Rocha, Carlos Henrique do Vale, Carlos HG Ferreira, Glauber Dias Gonçalves, Jussara Marques de Almeida, et al. 2024. Monitorando a Opinião Pública sobre Operações Policiais no Brasil via Comentários de Vídeos no YouTube. In *Brazilian Workshop on Social Network Analysis and Mining*.

[13] Mohammadreza Davoudi, Sheida Shirvani, Aliakbar Foroughi, and Fereshteh Rajaeiramsheh. 2022. Online gambling in Iranian social media users: prevalence, related variables and psychiatric correlations. *Journal of gambling studies* (2022).

[14] Aline Dias, Richardy R Tanure, Jussara M Almeida, Helen CSC Lima, and Carlos HG Ferreira. 2024. Análise da Percepção do Uso de Cigarros Eletrônicos no Brasil por meio de Comentários no YouTube. In *30th Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC, 45–53.

[15] Sally M Gainsbury. 2015. Online gambling addiction: the relationship between internet gambling and disordered gambling. *Current addiction reports* (2015).

[16] Lucas Rafael Stefanel Gris, Ricardo Marcacini, Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, and Sandra Maria Aluísio. 2023. Evaluating OpenAI's Whisper ASR for Punctuation Prediction and Topic Modeling of life histories of the Museum of the Person. *arXiv preprint arXiv:2305.14580* (2023).

[17] Maarten Grootendorst. 2024. BERTopic - Parameter Tuning. https://maartengr.github.io/BERTopic/getting_started/parameter%20tuning/parametertuning.html Accessed: July 2025.

[18] Hanifah Harahap and Farid Ridho. 2024. Detection of Online Gambling Web Defacement in University Domains Using Attack Signatures. In *2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD)* (2024-08). 73–78. doi:10.1109/ICoABCD63526.2024.10704413

[19] Sovy Muti Ardianty Hasibuan, Andri Soemitra, and Muhammad Syukri Albani. 2024. Analysis of Socioeconomic and Situational Factors, Winning Probability, and Perception of Convenience on Online Gambling Addiction Among Gen Z. *Jurnal Ad'ministrare* (2024), 357–366.

[20] Niklas Hopfgartner, Thorsten Ruprechter, and Denis Helic. 2022. Retention and Relapse in Gambling Self-help Communities on Reddit. In *International Conference on Social Informatics*.

[21] Scott Houghton, Andrew McNeil, Mitchell Hogg, and Mark Moss. 2019. Comparing the Twitter posting of British gambling operators and gambling affiliates: A summative content analysis. *International Gambling Studies* (2019).

[22] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction* (2020).

[23] IBOPE. 2023. Video Audience Share Percentage in Brazil. https://kantaribopemedia.com/conteudo/relatorios/april-2023/.

[24] Christian Jacques, Daniel Fortin-Guichard, Pierre Bergeron, Catherine Boudreault, David Lévesque, and Isabelle Giroux. 2016. Gambling content in Facebook games: A common phenomenon? *Computers in Human Behavior* (2016).

[25] Richard JE James and Alex Bradley. 2021. The use of social media in research on gambling: A systematic review. *Current Addiction Reports* (2021).

[26] Sang-Jo Ko, Jeong-Eun Seo, and Hun-Yeong Kwon. 2024. A Study on the Jurisdiction and Regulation of offshore Online Gambling between trading countries. In *Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance* (New York, NY, USA, 2024) *(ICEGOV '24)*. Association for Computing Machinery, 166–175. doi:10.1145/3680127.3680138

[27] Åsa Kroon. 2020. Converting gambling to philanthropy and acts of patriotism: The case of "The world's most Swedish gambling company". *Discourse, Context & Media* (2020).

[28] Erika Langham, Hannah Thorne, Matthew Browne, Phillip Donaldson, Judy Rose, and Matthew Rockloff. 2016. Understanding gambling related harm: a proposed definition, conceptual framework, and taxonomy of harms. *BMC Public Health* 16 (2016), 80. doi:10.1186/s12889-016-2747-0

[29] Tiina Latvala, Tomi Lintonen, and Anne Konu. 2019. Public health effects of gambling – debate on a conceptual model. *BMC Public Health* 19, 1 (2019), 1077. doi:10.1186/s12889-019-7391-z

[30] Larissa Malagoli, Giovana Piorino, Carlos Ferreira, and Ana Silva. 2024. Twitter and the 2022 Brazilian Elections Portrait: A Network and Content-Driven Analysis. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 283–291. doi:10.5753/webmedia.2024.241926

[31] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50 – 60. doi:10.1214/aoms/1177730491

[32] Geovana Oliveira, Otávio Venâncio, Vinícius Vieira, Jussara Almeida, Ana Silva, Ronan Ferreira, and Carlos Ferreira. 2024. Um Framework para Análise Bidimensional de Disseminação de Informações em Plataformas de Mídias Sociais. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 301–309. doi:10.5753/webmedia.2024.241957

[33] Antônio Pereira, Felipe Viegas, Diego Roberto Colombo Dias, Elisa Tuler, Ana Cláudia Machado, Guilherme Fonseca, Marcos André Gonçalves, and Leonardo Rocha. 2025. "Are the current topic modeling evaluation metrics enough?" Mitigating the limitations of topic modeling evaluation metrics using a multi-perspective game theoretic approach. *Knowledge-Based Systems* 320 (2025), 113634. doi:10.1016/j.knosys.2025.113634

[34] Antônio Pereira, Felipe Viegas, Marcos André Gonçalves, and Leonardo Rocha. 2023. Evaluating the Limits of the Current Evaluation Metrics for Topic Modeling. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web (WebMedia '23)*. 119–127. doi:10.1145/3617023.3617040

[35] L.E. Raileanu and K. Stoffel. 2004. Theoretical Comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence* 41 (2004), 77–93. doi:10.1023/B:AMAI.0000018580.96245.c6

[36] Melanie Revilla and Carlos Ochoa. 2017. Ideal and maximum length for a web survey. *International Journal of Market Research* 59, 5 (2017), 557–565.

[37] Johannes Singer, Vadim Kufenko, Andrea Wöhr, Marius Wuketich, and Steffen Otterbach. 2022. How do Gambling Providers Use the Social Network Twitter in Germany? An Explorative Mixed-Methods Topic Modeling Approach. 39, 3 (2022), 1371–1398. doi:10.1007/s10899-022-10158-y Publisher: Springer Science and Business Media LLC.

[38] Sonia Singh and Priyanka Gupta. 2014. Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey. *International Journal of Advanced Information Science and Technology (IJAIST)* 27, 27 (2014), 97–103.

[39] Elke Smith, Simon Michalski, Kilian H. K. Knauth, Nils Kaspar, and Jan Peters. 2023. Large-Scale Web Scraping for Problem Gambling Research: A Case Study of COVID-19 Lockdown Effects in Germany. 39 (2023).

[40] Richardy R. Tanure, Aline M. Dias, Lucas A. Camelo, Jussara Almeida, Helen C. S. C. Lima, and Carlos H. G. Ferreira. 2025. Caracterizacao do debate online sobre cigarro eletronico no Brasil: Uma analise de topicos de discussao no YouTube. In *Anais do Brazilian Symposium on Multimedia and the Web (Brasnam)*. Sociedade Brasileira de Computação (SBC), Recife, Brasil.

[41] Prasert Teppap, Panudech Tipauksorn, Somnuek Surathong, Wirot Ponglangka, and Prasert Luekhong. 2024. Automating Hidden Gambling Detection in Web Sites: A BeautifulSoup Implementation. In *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)* (2024-06). 132–139. doi:10.1109/JCSSE61278.2024.10613687 ISSN: 2642-6579.

[42] Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information processing & management* 50, 1 (2014), 104–112.

[43] Mark Van Der Maas and Jim Samuel. 2025. Online gambling forums as a potential target for harm reduction: an exploratory natural language processing analysis of a reddit.com forum. 22, 1 (2025). doi:10.1186/s12954-025-01220-0 Publisher:

Springer Science and Business Media LLC.
[44] Otávio Venâncio, Gabriel Gonçalves, Carlos Ferreira, and Ana Silva. 2024. Evidências de disseminação de conteúdo no Telegram durante o ataque aos órgãos públicos brasileiros em 2023. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 385–389. doi:10.5753/webmedia.2024.241972
[45] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of the 12a ACM WSDM*. 753–761.
[46] Heather Wardle, Louisa Degenhardt, Virve Marionneau, Gerda Reith, Charles Livingstone, Malcolm Sparrow, Lucy T Tran, Blair Biggar, Christopher Bunn, Michael Farrell, et al. 2024. The Lancet Public Health Commission on gambling. *The Lancet Public Health* (2024).
[47] Ye Yuan, Liji Wu, and Xiangmin Zhang. 2021. Gini-Impurity Index Analysis. *IEEE Transactions on Information Forensics and Security* 16 (2021), 3154–3169.