Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.8462

Artigo de Pesquisa/Research Paper

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

Gambling Narratives on Video Platforms: A Linguistic-Thematic Study of Fortune Tiger ("Jogo do Tigrinho") on YouTube

Gabriel Prenassi [Universidade Federal de São João del-Rei | *prenassigabriel@aluno.ufsj.edu.br*]
Ana Machado [Universidade Federal de São João del-Rei | *anaclaudiamachado211@aluno.ufsj.edu.br*]
Carlos Ferreira [Universidade Federal de Ouro Preto | *chgferreira@ufop.edu.br*]
Leonardo Rocha [Universidade Federal de São João del-Rei | *lcrocha@ufsj.edu.br*]

Departamento de Ciência da Computação, Universidade Federal de São João del-Rei, BR-494, s/n, Campus CTan, São João del-Rei, MG, 36301-360, Brasil.

Resumo. A promoção de jogos de azar online por influenciadores digitais tem se tornado uma preocupação social crescente. No Brasil, o jogo *Fortune Tiger*, popularmente conhecido como Jogo do Tigrinho, ganhou visibilidade no YouTube por meio de narrativas que prometem ganhos rápidos. Este estudo analisa como esses conteúdos são construídos e difundidos na plataforma por meio de uma análise linguístico-temática em larga escala. Com base em transcrição automática, modelagem de tópicos e extração de características linguísticas, identificamos duas narrativas predominantes: uma promocional, majoritária e linguisticamente mais simples, e outra crítica, com maior diversidade lexical e engajamento mais consistente. Os resultados evidenciam um desequilíbrio estrutural no discurso sobre jogos de azar online.

Abstract. The promotion of online gambling by digital influencers has become a growing social concern. In Brazil, Fortune Tiger, popularly known as Jogo do Tigrinho, gained visibility on YouTube through narratives that promise quick profits. This study analyzes how such content is constructed and disseminated on the platform through a large-scale linguistic-thematic analysis. Based on automatic transcription, topic modeling, and linguistic feature extraction, we identify two predominant narratives: a promotional one, which is dominant and linguistically simpler, and a critical one, marked by greater lexical diversity and more consistent engagement. The results reveal a structural imbalance in the discourse on online gambling.

Palavras-chave: Jogos de Azar Online, Influenciadores Digitais, YouTube, Jogo do Tigrinho, Análise Linguístico-Temática

Keywords: Online Gambling, Digital Influencers, YouTube, Fortune Tiger, Linguistic-Thematic Analysis

Recebido/Received: 14 June 2026 • **Aceito/Accepted:** 16 June 2026 • **Publicado/Published:** 10 July 2026

1 Introdução

O mercado de apostas online tem registrado crescimento acelerado, impulsionado pela digitalização, pela acessibilidade via dispositivos móveis e pela expansão regulatória em diferentes países, com projeções de perdas líquidas globais próximas a US$ 700 bilhões até 2028 [Wardle *et al*., 2024]. Esse cenário amplia a exposição a riscos e tem sido associado ao agravamento de vulnerabilidades socioeconômicas e a prejuízos financeiros significativos. No Brasil, essa dinâmica se expressa de maneira emblemática no *Fortune Tiger* (Jogo do Tigrinho), amplamente promovido por influenciadores, em um contexto no qual dados do Banco Central indicam que o mercado de apostas online mobiliza volumes expressivos de recursos, inclusive entre beneficiários de programas assistenciais[1].

Embora estudos anteriores tenham analisado a promoção de jogos de azar em plataformas digitais [Kroon, 2020; Chamil *et al*., 2024], o YouTube, apesar de sua centralidade no consumo de vídeo no Brasil [Hussein *et al*., 2020], permanece pouco explorado de forma sistemática, especialmente no que se refere ao conteúdo audiovisual. Nesse contexto, este trabalho investiga como influenciadores brasileiros constroem narrativas sobre o *Jogo do Tigrinho* na plataforma, buscando responder às seguintes questões:

- **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 à **RQ1**, realizamos transcrição automática (*Whisper*) de 1.068 vídeos com qualidade adequada após filtragem acústica e aplicamos o BERTopic para identificar agrupamentos semânticos dominantes, complementados por sumarização dos tópicos. Identificamos duas narrativas centrais: uma promocional, que representa cerca de 90% dos vídeos e se concentra em estratégias e incentivos ao jogo, e outra crítica, voltada a denúncias e impactos sociais. Para a **RQ2**, conduzimos uma análise quantitativa das transcrições com base em métricas de diversidade lexical, densidade de vocabulário, duração e engajamento (visualizações, curtidas e comentários), além de atributos linguísticos extraídos por meio do LFTK (*Linguistic Feature Toolkit*). Os

[1] Estudos Especiais do Banco Central

resultados indicam que vídeos promocionais são mais curtos e linguisticamente mais simples e repetitivos, enquanto vídeos críticos apresentam maior diversidade lexical, maior complexidade discursiva e engajamento mais consistente. Este trabalho de iniciação científica resultou na publicação de um artigo no WebMedia 2025 (A3) [Prenassi *et al*., 2025].

2 Trabalhos Relacionados

A literatura sobre jogos de azar *online* tem se expandido nos últimos anos, impulsionada pela digitalização do setor e por eventos como a pandemia de COVID-19 [Ko *et al*., 2024; Smith *et al*., 2023]. Parte dos estudos dedica-se à identificação de práticas ilícitas e à análise estrutural do ecossistema de apostas, propondo métodos para detectar anúncios ocultos [Teppap *et al*., 2024], identificar domínios comprometidos [Harahap and Ridho, 2024] e mapear redes de sites de apostas [Chen *et al*., 2024]. Tais esforços articulam-se a discussões mais amplas sobre governança, ética e responsabilidade no setor, incluindo o *design* de tecnologias potencialmente viciantes [Cemiloglu *et al*., 2020] e os desafios da operação de serviços *offshore* [Ko *et al*., 2024]. Em paralelo, outra vertente concentra-se na análise do comportamento de usuários e das dinâmicas comunicacionais em plataformas digitais, examinando comentários em fóruns e redes sociais para identificar padrões discursivos e estratégias de promoção [van der Maas and Samuel, 2025; Smith *et al*., 2023; Singer *et al*., 2022]. No que se refere ao YouTube, entretanto, os estudos ainda são limitados e tendem a concentrar-se em análises qualitativas [Chamil *et al*., 2024], anúncios publicitários [Kroon, 2020] ou comentários textuais [Costa *et al*., 2025].

Apesar dessas contribuições, análises quantitativas do conteúdo audiovisual em larga escala ainda são raras, especialmente no contexto brasileiro. **Este trabalho preenche essa lacuna por meio de uma análise linguístico-temática de vídeos sobre o *Fortune Tiger* no YouTube, combinando transcrição automática, modelagem e sumarização de tópicos e métricas linguísticas e de engajamento para caracterizar diferenças estruturais e discursivas entre narrativas promocionais e críticas**.

3 Metodologia

3.1 Coleta e Pré-processamento de Dados

Este estudo utiliza a base de dados de [Costa *et al*., 2025], coletada via API do YouTube entre janeiro de 2023 e julho de 2024, com foco em vídeos sobre o *Fortune Tiger*, incluindo formatos tradicionais e *Shorts*, no contexto brasileiro e em língua portuguesa. Em cada mês, foram selecionados os 50 vídeos mais bem ranqueados, adotando-se o ranqueamento como uma aproximação da exposição dos usuários. Ao final, a base compreende 3.983 canais, 7.587 vídeos distintos e 60.086 usuários únicos que interagiram com esses conteúdos. A partir desse conjunto, os áudios foram obtidos via *yt-dlp*, desconsiderando vídeos removidos, privados ou com restrição de idade, para os quais o *download* não é permitido pela plataforma, totalizando 1.956 áudios coletados, os quais foram associados às respectivas métricas de engajamento (visualizações, curtidas e comentários), empregadas na análise do alcance e do impacto dos discursos no contexto da **RQ2**.

3.2 Filtragem Acústica e Seleção de Áudios

Como as análises dependem da qualidade das transcrições automáticas, adotamos uma estratégia de filtragem acústica para selecionar apenas áudios adequados. Inicialmente, foram extraídas características com a biblioteca *Librosa*, incluindo a duração do áudio, a frequência de amostragem (associada à preservação dos detalhes sonoros), métricas de amplitude (média e máxima, indicativas da captação e de picos de volume), a razão entre essas amplitudes (como medida de consistência do sinal), a relação sinal-ruído (associada à proporção entre conteúdo falado e ruído), a taxa de cruzamento por zero (associada à distinção entre fala, ruído e silêncio) e a detecção de regiões de silêncio (que permite identificar pausas ou falhas na gravação). Amostras com valores nulos foram removidas, resultando em 1.930 áudios válidos.

Os áudios foram então agrupados por *K-Means*, com normalização via *StandardScaler*, ambos empregados com seus parâmetros padrão. O número ótimo de *clusters* foi definido por uma variação do método do cotovelo baseada no índice *BetaCV* [Pereira *et al*., 2025], que avalia a razão entre coesão intra-*cluster* e separação inter-*cluster*. Foram testados valores entre 5 e 100 grupos, sendo 35 o valor identificado como ponto de melhor equilíbrio. A partir desses grupos, selecionamos uma amostra proporcional de 100 áudios, transcritos com o *Whisper* (*Small*), escolhido por sua alta acurácia em português e baixo custo computacional [Gris *et al*., 2023], e avaliados manualmente por três anotadores, sendo 58% classificados, via moda, como "bons".

Com base nessas anotações, foi treinado um classificador de árvore de decisão (ID3), utilizando entropia como critério de divisão. O modelo foi ajustado por meio de *Grid Search*, com validação cruzada estratificada (5 *folds*), e avaliado com a métrica *Macro-F1*. Em seguida, o classificador foi aplicado aos áudios não utilizados na etapa de treinamento, correspondentes ao restante do conjunto de 1.930 áudios, dos quais 55% foram considerados adequados para transcrição. As etapas analíticas subsequentes utilizaram apenas os áudios classificados como adequados, somados às amostras previamente rotuladas como "boas", totalizando 1.068 transcrições.

3.3 Processamento e Análise Linguístico-Temática

Para examinar as transcrições, empregamos um conjunto de procedimentos de Processamento de Linguagem Natural. Inicialmente, realizamos o **pré-processamento**, visando remover elementos linguísticos irrelevantes e aprimorar a representação textual. Adotamos procedimentos de lematização; normalização textual (conversão para minúsculas, remoção de acentuação e redução de repetições de caracteres); remoção de URLs, pontuação e números; exclusão de palavras com até três letras; e eliminação de *stopwords*.

Em seguida, procedemos à **modelagem de tópicos** por meio do BERTopic. Durante a aplicação do modelo, foram identificados agrupamentos de baixa representatividade (menos de cinco transcrições), bem como documentos classificados como *outliers* pelo HDBSCAN. Conforme as diretrizes do modelo [Grootendorst, 2024], tais grupos tendem a apresentar fragilidade semântica. Ademais, a análise qualitativa e hierárquica indicou convergência para

dois núcleos temáticos dominantes. Assim, mantivemos apenas os dois tópicos principais, que concentravam 998 vídeos (93% do total com transcrição adequada). Com base nesses agrupamentos, realizamos a **sumarização dos tópicos**, selecionando as 10 palavras mais representativas segundo o BERTopic e 10 transcrições por tópico, sendo três indicadas pelo modelo como mais representativas e sete selecionadas aleatoriamente. Essas informações foram utilizadas para gerar títulos e resumos com auxílio do ChatGPT.

Por fim, conduzimos a **extração de características linguísticas**, etapa diretamente relacionada à **RQ2**, com o objetivo de caracterizar os grupos temáticos identificados na **RQ1**. Para isso, extraímos mais de 200 atributos textuais por meio do LFTK, incluindo métricas estatísticas (e.g., contagem de palavras e tamanho médio de sentenças), índices de complexidade linguística (como densidade lexical) e propriedades psicométricas, como idade média de aquisição do vocabulário e nível educacional estimado. Como a análise manual desse conjunto seria excessivamente custosa, realizamos uma seleção criteriosa das características com maior potencial discriminativo, ranqueando-as com base na métrica de *Gini Gain*, que avalia diferentes limiares de separação e mede a redução da *Gini Impurity*, indicador da pureza dos grupos formados. Quanto maior o *Gini Gain*, maior a capacidade de separação entre os tópicos.

Com base no ranqueamento pelo *Gini Gain*, selecionamos as 10 características com maior capacidade discriminativa. Como atributos distintos podem apresentar significados semelhantes, aplicamos uma etapa adicional de filtragem para reduzir redundâncias, calculando a correlação de *Spearman* entre os pares de variáveis. Adotamos um procedimento sequencial no qual, partindo da característica com maior *Gini Gain*, removemos aquelas com correlação superior a 0.7 (limiar considerado alto na literatura [Ali Abd Al-Hameed, 2022]), até obter um conjunto final não redundante. Com base nesse conjunto, segmentamos as transcrições pelos tópicos identificados e calculamos as médias por grupo. Para avaliar a significância das diferenças observadas, aplicamos o teste não paramétrico de *Mann-Whitney U*.

4 Resultados e Discussões

4.1 Caracterização Global das Transcrições

Nesta etapa, as análises consideraram o conjunto completo de transcrições de qualidade adequada. Foram realizadas análises quantitativas das propriedades estruturais, iniciando pela distribuição do número de palavras por vídeo (Figura 1). Com base nas transcrições sem pré-processamento, a extensão variou de 1 a 20.334 palavras, com média de 1.182, desvio padrão de aproximadamente 1.307 e mediana de 864. Além disso, 75% das transcrições possuem até 1.611 palavras, indicando assimetria na distribuição e a presença de vídeos substancialmente mais longos, o que explica a diferença entre a média e o valor máximo.

A segunda análise examinou o tamanho do vocabulário de cada vídeo, calculado a partir das transcrições sem pré-processamento e definido como o número de palavras únicas por documento. A média foi de 380 termos distintos por vídeo, com desvio padrão de 314, mínimo de 1 e máximo de 4.516 termos únicos. A mediana foi de 331 palavras, e

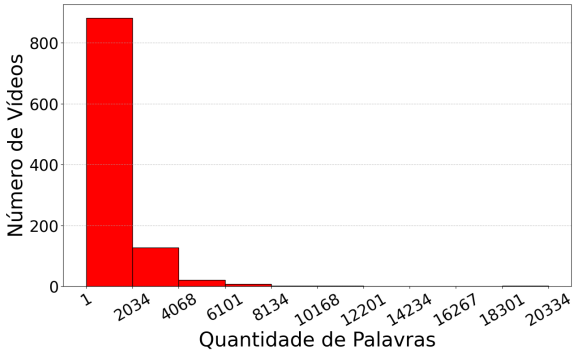

Figura 1. Distribuição dos vídeos por quantidade de palavras.

75% das transcrições não ultrapassam 519 termos distintos. Também foi analisada a proporção de palavras únicas em relação ao total de palavras (Figura 2), cuja distribuição se concentrou entre 0,3 e 0,5. Valores superiores a 0,6 ocorreram em textos mais curtos e lexicalmente diversos, enquanto valores inferiores a 0,2 caracterizam transcrições mais longas com maior repetição lexical. **Esses resultados indicam um grau considerável de diversidade lexical no corpus analisado**.

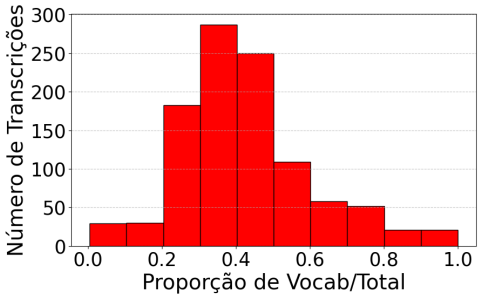

Figura 2. Distribuição do vocabulário por total de palavras.

4.2 Caracterização das Transcrições por Tópico

Após a identificação dos dois grupos temáticos dominantes, analisamos suas características lexicais distintivas. A Figura 3 apresenta os 10 principais termos de cada tópico, revelando duas narrativas distintas sobre os jogos de azar. Observa-se uma clara divisão entre conteúdos críticos e promocionais. O Tópico 0 reúne termos como "jogo", "polícia", "crime", "divulgar", "dinheiro" e "influenciador", indicando um viés predominantemente crítico, com ênfase na legalidade das práticas e no papel de influenciadores na sua divulgação. Já o Tópico 1 inclui termos como "pagar", "rodada", "jogar", "banca", "estratégia" e "turbo", associados a uma linguagem voltada ao público de cassinos *online*, com foco na dinâmica dos jogos e em estratégias de incentivo à prática.

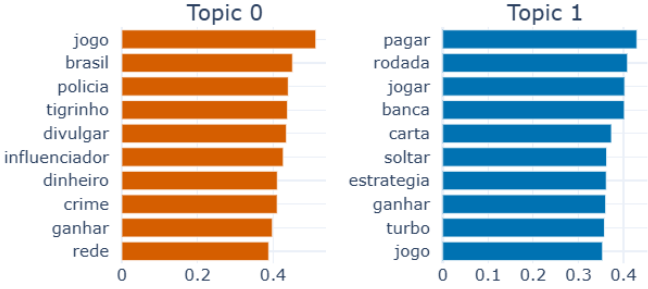

Figura 3. Representação dos tópicos obtidos.

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.

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 4. Resultados da sumarização para os tópicos identificados.

4.2.1 Sumarização dos Tópicos

Complementarmente à análise dos termos mais representativos, a Figura 4 apresenta os resumos gerados para os dois tópicos identificados. O Tópico 0 enfatiza aspectos criminais e impactos sociais negativos, mencionando operações policiais, denúncias e práticas como estelionato, lavagem de dinheiro e propaganda enganosa, além da exploração de indivíduos em situação de vulnerabilidade. Por outro lado, o Tópico 1 descreve o jogo sob a perspectiva de jogadores e influenciadores que compartilham estratégias e experiências, promovendo a prática das apostas sob a lógica de ganhos recorrentes. Essa narrativa tende a atribuir eventuais prejuízos à conduta individual do jogador, com menor ênfase no papel das plataformas e dos influenciadores. Assim, em relação à **RQ1**, os resultados indicam **uma divisão temática consistente entre conteúdos críticos, voltados à denúncia e às consequências sociais do jogo, e conteúdos promocionais, centrados em estratégias e incentivos à prática**.

4.2.2 Número de Vídeos por Tópico

A distribuição dos vídeos entre os tópicos evidencia a predominância do Tópico 1 (promocional), com 895 vídeos, enquanto o Tópico 0 (crítico) reúne 103. Esse predomínio sugere maior produção de conteúdos voltados a estratégias e dinâmicas de cassinos *online*. Em contrapartida, a menor presença de vídeos críticos pode indicar um alcance mais restrito desse tipo de conteúdo na plataforma, apesar de sua relevância para o debate social. Ao todo, os dois tópicos abrangem 998 dos 1.068 vídeos com transcrição adequada (93% do corpus), assegurando representatividade para as análises da **RQ1**.

4.2.3 Duração dos Vídeos por Tópico

A análise da duração dos vídeos (Figura 5) acrescenta uma dimensão estrutural à distinção entre os tópicos, dialogando diretamente com a **RQ2**. No Tópico 0 (crítico), a maioria dos vídeos possui menos de 500 segundos (cerca de 8 minutos), embora haja *outliers* superiores a 7.000 segundos (aproximadamente 2 horas), evidenciando a coexistência de conteúdos mais curtos, como notícias rápidas, e produções significativamente mais longas. Já o Tópico 1 (promocional) apresenta distribuição mais concentrada em vídeos de curta a média duração, predominantemente com duração inferior a 600 segundos (cerca de 10 minutos). Esses resultados sugerem diferenças não apenas no conteúdo, mas também no formato: enquanto o Tópico 0 alterna entre produções breves e longas, o Tópico 1 adota um padrão mais uniforme e conciso, alinhado a estratégias de engajamento rápido.

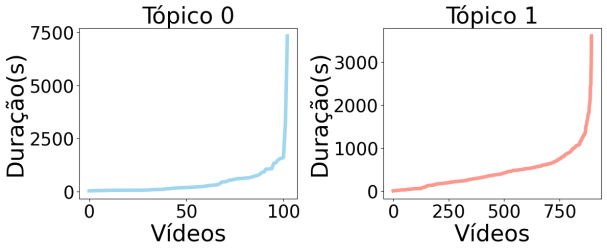

Figura 5. Duração dos vídeos por tópico.

4.2.4 Diversidade Lexical por Tópico

Para estimar a complexidade linguística dos discursos, analisamos a diversidade lexical, calculada após o pré-processamento como a razão entre o número de palavras únicas e o total de palavras de cada transcrição. A Figura 6 evidencia uma distinção consistente entre os tópicos. O Tópico 0 (crítico) concentra índices mais elevados, predominantemente entre 0,5 e 0,9, indicando maior variedade vocabular, compatível com a complexidade temática dos conteúdos críticos. Já o Tópico 1 (promocional) apresenta valores mais baixos, majoritariamente entre 0,4 e 0,5, refletindo construções textuais mais repetitivas e padronizadas. Observa-se, contudo, a presença de transcrições muito curtas que aparecem como *outliers* com diversidade próxima de 1,0, embora esse valor elevado decorra do baixo número total de palavras e não represente necessariamente maior sofisticação linguística. Em conjunto, esses resultados indicam que conteúdos promocionais tendem a operar com estrutura lexical mais restrita e recorrente, enquanto conteúdos críticos mobilizam maior variedade vocabular, contribuindo diretamente para a caracterização proposta na **RQ2**.

4.2.5 Engajamento dos Vídeos por Tópico

A análise das métricas de engajamento aprofunda as diferenças entre as narrativas identificadas. A Figura 7 apresenta as distribuições de visualizações, curtidas e comentários por tópico. Em termos de visualizações (Figura 7(a)), embora

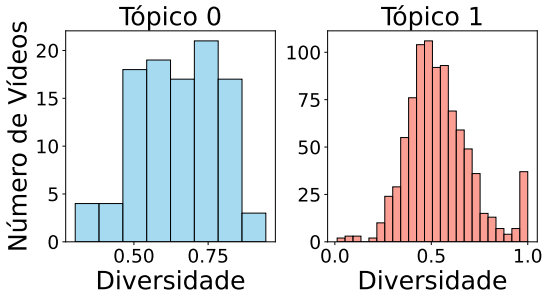

Figura 6. Diversidade do vocabulário por tópico.

o Tópico 1 (promocional) concentre um número maior de vídeos, o Tópico 0 (crítico) apresenta maior média por conteúdo (21.736 contra 10.461) e mediana substancialmente superior (6.469 contra 228), indicando engajamento mais consistente nos vídeos críticos. Em contraste, o Tópico 1 exibe forte assimetria, com picos de até 1.500.000 visualizações (frente a 309.000 no Tópico 0), evidenciando concentração de audiência em poucos conteúdos virais.

Padrão semelhante é observado nas curtidas e comentários (Figuras 7(b) e (c)). O Tópico 0 (crítico) apresenta médias e medianas superiores tanto em curtidas, com média de 1.560 e mediana de 241, quanto em comentários, com média de 108 e mediana de 26. Já o Tópico 1 (promocional) registra valores mais baixos, com média de 511 e mediana de 17 curtidas, além de média de 60 e mediana de 5 comentários, embora concentre os maiores picos absolutos, alcançando 84 mil curtidas e 4.400 comentários. Em conjunto, os resultados indicam que conteúdos críticos tendem a gerar engajamento mais estável e distribuído, enquanto conteúdos promocionais apresentam alta variabilidade, marcada por um grande volume de vídeos com baixo desempenho e poucos casos de viralização intensa, o que contribui diretamente para a caracterização proposta na **RQ2**.

4.2.6 Características Linguísticas por Tópico

Além das diferenças estruturais e de engajamento, investigamos quais características linguísticas distinguem melhor os dois tópicos identificados. A Tabela 1 apresenta as 10 métricas obtidas a partir do processo descrito na Seção 3.3. Como parte desses atributos é complementar ou potencialmente redundante, aplicamos um procedimento adicional de seleção para reduzir sobreposição semântica, resultando em quatro características com maior poder discriminativo entre os Tópicos 0 e 1, destacadas em cinza: **diversidade de vocabulário** (linear e não linear), **escolaridade** e **proporção de substantivos próprios**. A Tabela 2 apresenta os valores médios por tópico, sendo $*$ indicativo de diferença estatisticamente significativa segundo o teste de *Mann-Whitney U*.

Os resultados indicam que o Tópico 1 (promocional) apresenta médias significativamente inferiores em todas as métricas, evidenciando construções linguísticas mais simples e padronizadas. As medidas de diversidade lexical indicam maior repetição nos conteúdos promocionais, enquanto o Tópico 0 (crítico) mobiliza um vocabulário mais variado. De forma consistente, menores valores de **escolaridade** no Tópico 1 sugerem menor complexidade textual, ao passo que a maior **proporção de substantivos próprios** no Tópico 0 reflete maior presença de referências institucionais e atores sociais, coerente com o caráter investigativo dessas narrativas.

Em síntese, as diferenças estruturais, linguísticas e de engajamento confirmam distinções sistemáticas entre narrativas promocionais e críticas sobre o *Fortune Tiger*. Assim, respondemos à **RQ2** ao demonstrar que **conteúdos promocionais tendem a ser mais curtos, linguisticamente mais simples e marcados por engajamento concentrado em poucos vídeos virais, enquanto conteúdos críticos apresentam maior complexidade discursiva e engajamento mais consistente.**

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.

5 Limitações

Uma possível limitação deste estudo está relacionada à redução do conjunto inicial de vídeos até a formação do corpus efetivamente analisado. Embora esse processo tenha sido necessário para aumentar a confiabilidade das transcrições automáticas e das análises linguísticas subsequentes, a diferença entre os vídeos inicialmente identificados, os áudios efetivamente baixados e as transcrições consideradas adequadas pode ter afetado a composição da amostra. Essa redução decorre, em parte, de restrições da própria plataforma, como vídeos removidos, privados ou com restrição de idade, além da exclusão de áudios de baixa qualidade acústica ou inadequados à transcrição.

Esse possível efeito deve ser considerado em relação aos padrões observados nos resultados. As análises indicaram que conteúdos promocionais tendem a ser mais curtos, repetitivos e linguisticamente mais simples, enquanto conteúdos críticos apresentam maior diversidade lexical, maior presença de referências institucionais e engajamento mais consistente. Nesse sentido, diferentes formatos podem ter sido afetados de maneiras distintas pelo processo de coleta e filtragem. Vídeos promocionais, por exemplo, podem incluir gravações caseiras, demonstrações de jogo, capturas de tela ou narrações em condições menos controladas, o que poderia aumentar a chance de exclusão devido à baixa qualidade acústica ou à indisponibilidade. Por outro lado, vídeos críticos podem estar mais associados a formatos informativos, jornalísticos ou analíticos, possivelmente com condições mais favoráveis de áudio e transcrição.

Ainda assim, como os conteúdos promocionais permaneceram predominantes no corpus final, esse tipo de narrativa não foi excluído pelas etapas de filtragem. Portanto, a redução da amostra não elimina os padrões observados, mas indica que sua generalização para todo o universo de vídeos inicialmente identificado deve considerar o recorte imposto pela disponibilidade e pela qualidade dos áudios. Assim, os achados devem ser interpretados como representativos do subconjunto de vídeos com áudio acessível e qualidade suficiente para análise textual.

6 Conclusões e Trabalhos Futuros

Este estudo analisou a produção audiovisual no YouTube acerca do "Jogo do Tigrinho", em um contexto de preocupação social diante da disseminação de conteúdos que

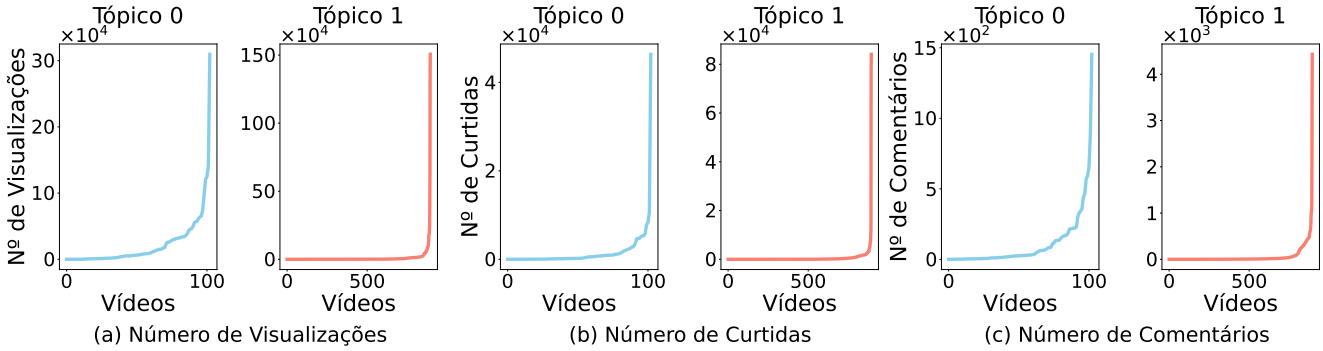

Figura 7. Número de visualizações por tópico (à esquerda), curtidas (no centro) e comentários (à direita).

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 analisa de forma linear (denominador é a raiz quadrada de 2 vezes o tamanho do total do texto)

Tabela 1. Características mais discriminativas.

promovem práticas associadas a jogos de azar em plataformas digitais. Embora o fenômeno tenha sido investigado sob diferentes perspectivas, identificamos uma lacuna na análise estruturada e linguística desses conteúdos, especialmente considerando a atuação de influenciadores brasileiros e a aplicação de métodos quantitativos em larga escala. Para suprir essa lacuna, propusemos uma abordagem linguístico-temática capaz de mapear a distribuição entre discursos promocionais e críticos e evidenciar, de forma sistemática e replicável, diferenças estruturais.

Os resultados obtidos evidenciam um desequilíbrio na forma como o "Jogo do Tigrinho" é representado na plataforma. Identificamos a predominância de conteúdos promocionais, associados a estratégias, dinâmicas de jogo e incentivos à prática de apostas, em contraste com um conjunto menor de vídeos críticos, voltados a denúncias, impactos sociais e aspectos criminais relacionados ao jogo. Além disso, as análises estruturais, linguísticas e de engajamento mostraram que vídeos promocionais tendem a ser mais curtos, linguisticamente mais simples e marcados por maior repetição lexical, enquanto vídeos críticos apresentam maior diversidade lexical, maior complexidade discursiva e engajamento mais consistente. Esses achados indicam que as narrativas promocionais recorrem a uma linguagem mais padronizada, repetitiva e de fácil assimilação, ao passo que as narrativas críticas apresentam vocabulário mais variado e maior presença de referências institucionais e de atores sociais. Esse contraste evidencia um cenário preocupante, no qual conteúdos de incentivo ao jogo aparecem em maior volume do que discursos voltados à problematização de seus riscos sociais e financeiros.

Além dos achados empíricos, o estudo demonstra a viabilidade de analisar transcrições de vídeos como fonte para compreender fenômenos sociais em plataformas digitais. Essa perspectiva é relevante porque permite observar não apenas a presença de conteúdos sobre jogos de azar online, mas também as formas como esses conteúdos são formulados, diferenciados e potencialmente direcionados ao público. Com isso, o trabalho aproxima métodos de Inteligência Artificial e de Processamento de Linguagem Natural a uma discussão social atual, relacionada à influência digital, à promoção de apostas e à circulação de conteúdos sensíveis no ambiente online.

Como desdobramentos futuros, propomos incorporar dimensões visuais e sonoras, desenvolver modelos de detecção de narrativas promocionais e críticas e aprofundar a análise das implicações éticas e regulatórias associadas à promoção de jogos de azar em ambientes digitais. Essas extensões podem oferecer uma compreensão mais ampla das estratégias de persuasão presentes nos vídeos, indo além do conteúdo textual das transcrições, e apoiar iniciativas futuras de monitoramento em larga escala. Além disso,

podem contribuir para discussões sobre transparência, responsabilidade das plataformas e mitigação da circulação de conteúdos sensíveis relacionados a jogos de azar online.

Declarações complementares

Financiamento

Esta pesquisa foi financiada por: CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde e Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR; 408490/2024-1).

Contribuições dos autores

G. Prenassi, A. Machado, C. Ferreira e L. Rocha contribuíram para a concepção do estudo (*Conceptualization*) e análise formal dos resultados (*Formal analysis*). G. Prenassi foi responsável pelo desenvolvimento dos códigos (*Software*), investigação (*Investigation*) e escrita do rascunho original (*Writing – original draft*). L. Rocha atuou na supervisão da pesquisa (*Supervision*). A. Machado e C. Ferreira colaboraram na validação da metodologia (*Validation*). Todos os autores participaram da revisão (*Writing – review & editing*) e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

Os conjuntos de dados (e/ou softwares) gerados e/ou analisados durante o estudo atual serão feitos mediante solicitação.

Outras informações relevantes

Ferramentas de Inteligência Artificial Generativa foram utilizadas como suporte à revisão gramatical do texto. Os autores realizaram uma revisão completa do texto e assumem total responsabilidade pela integridade das informações apresentadas.

Referências

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

Cemiloglu, D., Arden-Close, E., Hodge, S., Kostoulas, T., Ali, R., and Catania, M. (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)*, pages 1–10. IEEE.

Chamil, A. Y., Djuanda, S. A., and Septaviana, N. (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*.

Chen, J., Zheng, S., Cheng, Y., and Zhang, Z. (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*, pages 205–211.

Costa, J., Oliveira, G., Fonseca, G., Reis, D., Oliveira Teixeira, G., Cunha, W., Rocha, L., and Ferreira, C. H. (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*, pages 42–51.

Gris, L. R. S., Marcacini, R., Junior, A. C., Casanova, E., Soares, A., and Aluísio, S. M. (2023). Evaluating openai's whisper asr for punctuation prediction and topic modeling of life histories of the museum of the person.

Grootendorst, M. (2024). Bertopic - parameter tuning. Accessed: July 2025.

Harahap, H. and Ridho, F. (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)*, pages 73–78. IEEE.

Hussein, E., Juneja, P., and Mitra, T. (2020). Measuring misinformation in video search platforms: An audit study on youtube. *Proceedings of the ACM on Human-Computer Interaction*.

Ko, S.-J., Seo, J.-E., and Kwon, H.-Y. (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*, pages 166–175.

Kroon, Å. (2020). Converting gambling to philanthropy and acts of patriotism: The case of "the world's most swedish gambling company". *Discourse, Context & Media*.

Pereira, A., Viegas, F., Dias, D. R. C., Tuler, E., Machado, A. C., Fonseca, G., Gonçalves, M. A., and Rocha, L. (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*, 320:113634.

Prenassi, G., Machado, A., Souza, E., Brito, M., Reis, D., Costa, J., Oliveira, G., Ferreira, C., and Rocha, L. (2025). Narrativas de jogos de azar em plataformas de vídeo: Um estudo linguístico-temático sobre o jogo do tigrinho no youtube. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 366–375. SBC.

Singer, J., Kufenko, V., Wöhr, A., Wuketich, M., and Otterbach, S. (2022). How do gambling providers use the social network twitter in germany? an explorative mixed-methods topic modeling approach. *Journal of Gambling Studies*, 39(3):1371–1398.

Smith, E., Michalski, S., Knauth, K. H., Kaspar, K., Reiter, N., and Peters, J. (2023). Large-scale web scraping for problem gambling research: a case study of covid-19 lockdown effects in germany. *Journal of Gambling Studies*, 39(3):1487–1504.

Teppap, P., Tipauksorn, P., Surathong, S., Ponglangka, W., and Luekhong, P. (2024). Automating hidden gambling detection in web sites: A beautifulsoup implementation. In *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 132–139. IEEE.

van der Maas, M. and Samuel, J. (2025). Online gambling forums as a potential target for harm reduction: an exploratory natural language processing analysis of a reddit. com forum. *Harm reduction journal*, 22(1):77.

Wardle, H., Degenhardt, L., Marionneau, V., Reith, G., Livingstone, C., Sparrow, M., Tran, L. T., Biggar, B., Bunn, C., Farrell, M., *et al.* (2024). The lancet public health commission on gambling. *The Lancet Public Health*.