

Adultização Infantil e Regulação Digital: Uma Análise do Debate sobre a Lei nº 15.211/2025 no YouTube

Ester Souza¹, Nathan Lima¹, Vítor Amorim¹, Mateus Brito¹, Gabriel Prenassi¹, Arthur Assis¹, Henrique Teixeira¹, Ana Cláudia Machado¹, Diego Roberto Colombo Dias², Leonardo Rocha¹

¹Universidade Federal de São João del-Rei, ²Universidade Federal do Espírito Santo,
(estermariasouza2005,nathan116lima,voamorim345,mateusdeoliveirabritto,
prenassigabriel,axassis11,rickmteixeira10,anaclaudiamachado211)@aluno.ufsj.edu.br
lrocha@ufsj.edu.br;diego.dias@ufes.br

ABSTRACT

The public controversy surrounding child adultification and digital regulation gained major visibility in Brazil after the viral repercussion of Law No. 15.211/2025, popularly known as the “Lei Felca”. Although prior studies have examined child exposure, platform risks, and engagement dynamics in digital media, we still know little about how a recent and socially sensitive controversy is structured, circulated, and contested within large-scale audiovisual platforms. In this paper, we present a large-scale computational analysis of YouTube videos and comments discussing child adultification and the so-called Lei Felca. Our approach combines validated automatic transcriptions, metadata collection, topic modeling, sentiment classification, and linguistic analysis to examine how creators frame the debate, how audiences react to these frames, and how different publication formats shape visibility and interaction. The results show that the controversy is organized around distinct thematic frames, such as regulation, child exploitation, privacy, games, and political polarization, which are associated with different audience responses. We also find that traditional videos and YouTube Shorts play different roles in the platform ecology: long-form videos concentrate more technical, argumentative, and interaction-intensive discussion, whereas Shorts favor rapid diffusion and visibility, but with limited conversion into effective debate. Finally, linguistic analyses reveal systematic adaptations in both creators’ discourse and audience comments according to topic, format, and polarization. Taken together, these findings show that YouTube is not merely a neutral host of this controversy, but an active environment that shapes how it is framed, amplified, and publicly negotiated.

KEYWORDS

Child Adultification, YouTube, Public Controversy, Short-form Video, Natural Language Processing

1 INTRODUÇÃO

A crescente presença de crianças e adolescentes em plataformas digitais ampliou a visibilidade de práticas de exposição

precoce, monetização da infância e circulação de conteúdos que tensionam os limites entre entretenimento, consumo e proteção de direitos [19, 39]. No Brasil, esse cenário se tornou particularmente sensível diante do debate sobre a adultização infantil, entendida como a inserção de menores em contextos, estéticas e performances associados ao universo adulto, frequentemente atravessados por lógicas de engajamento, visibilidade e monetização. Embora o problema não seja novo, sua intensificação em plataformas digitais recolocou a discussão em termos de responsabilidade parental, atuação das plataformas e proteção integral de crianças e adolescentes [3, 4, 16, 23].

Esse debate ganhou escala nacional em agosto de 2025, quando o influenciador Felca publicou um vídeo de denúncia que repercutiu amplamente ao expor o processo de adultização infantil em ambientes digitais e ao questionar o papel das plataformas na circulação e na amplificação desse tipo de conteúdo. A repercussão impulsionou discussões midiáticas, mobilização social e iniciativas legislativas associadas à chamada “Lei Felca”, em referência à Lei nº 15.211/2025 (ECA Digital) e ao debate sobre proteção digital e responsabilização de plataformas. Mais do que um episódio de grande visibilidade, esse momento tornou explícita uma controvérsia contemporânea sobre a infância, a regulação digital, a liberdade de circulação de conteúdo e os limites éticos da exploração da imagem de menores no ambiente online.

Embora a literatura já tenha examinado a exposição infantil em plataformas digitais e, em paralelo, dinâmicas de engajamento e circulação de narrativas em ecossistemas audiovisuais [6, 8, 20, 25, 29, 32, 36, 40, 43, 45], ainda conhecemos pouco sobre como uma controvérsia pública recente e socialmente sensível se estrutura discursivamente no YouTube. Esse problema é particularmente relevante porque a plataforma não apenas hospeda vídeos sobre temas controversos, mas também condiciona sua circulação e suas formas de interação por meio de mecanismos de recomendação, métricas de engajamento e formatos distintos de consumo [2, 14, 15].

Este artigo apresenta uma análise em larga escala de vídeos e comentários do YouTube relacionados à adultização infantil e à Lei nº 15.211/2025. Para isso, coletamos vídeos e seus comentários, geramos e validamos transcrições automáticas, distinguimos conteúdos nos formatos tradicional e *YouTube Shorts* e aplicamos um *pipeline* de Processamento de Linguagem Natural baseado em modelagem de tópicos, classificação de sentimentos e análise linguística. Com esse desenho, buscamos compreender como o debate é enquadrado pelos criadores,

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia’2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

como a audiência reage a esses enquadramentos e de que modo o formato do vídeo interfere na circulação, na profundidade e na polarização da controvérsia. A partir desse objetivo, o estudo é guiado pelas seguintes perguntas de pesquisa:

- RQ1:** *Quais enquadramentos temáticos do debate sobre adultização infantil e a Lei nº 15.211/2025 no YouTube, e como eles se relacionam às reações da audiência?*
- RQ2:** *Como vídeos tradicionais e YouTube Shorts diferem na circulação do debate em termos de temas, engajamento e propensão à interação?*
- RQ3:** *Que diferenças linguísticas caracterizam o discurso dos criadores e a resposta da audiência, considerando tema, formato e polarização dos comentários?*

De forma sintética, os resultados mostram que o debate se organiza em torno de enquadramentos temáticos distintos, como regulação, exploração de menores, adultização e impactos psicossociais, privacidade e verificação etária, *games* e polarização política, e que esses enquadramentos influenciam diretamente a postura da audiência nos comentários. Mostram também que vídeos tradicionais e *Shorts* cumprem funções diferentes na plataforma: enquanto os primeiros concentram temas mais técnico-jurídicos e apresentam maior capacidade de converter visualizações em discussão efetiva, os segundos favorecem a rápida disseminação de controvérsias, subnichos e repercussões ligadas a figuras públicas, mas com baixa propensão ao debate. Por fim, a análise linguística evidencia que tanto os criadores quanto a audiência adaptam seu discurso ao tema, ao formato e ao grau de polarização: tópicos mais regulatórios mobilizam uma linguagem mais complexa e impessoal, enquanto comentários negativos tendem a ser mais densos e argumentativos do que os positivos. Em conjunto, esses achados indicam que a arquitetura do YouTube afeta não apenas a circulação do conteúdo, mas também a forma como ele é enunciado, interpretado e disputado.

Como contribuições, o artigo oferece: (i) uma caracterização computacional de um debate recente e socialmente sensível sobre infância, regulação digital e responsabilização de plataformas; (ii) uma análise integrada entre o discurso dos criadores e a repercussão da audiência nos comentários; (iii) uma comparação sistemática entre vídeos tradicionais e *Shorts* como formatos de circulação discursiva; e (iv) evidências de que a arquitetura da plataforma não apenas distribui conteúdos, mas condiciona a forma, a profundidade e a intensidade com que controvérsias públicas são negociadas no YouTube.

Este artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve os procedimentos metodológicos de coleta, transcrição, validação e análise dos dados. A Seção 4 discute os resultados empíricos, organizados em torno das perguntas de pesquisa. Por fim, a Seção 5 apresenta as conclusões do estudo, suas limitações e direções para trabalhos futuros.

2 TRABALHOS RELACIONADOS

A literatura relacionada a este trabalho pode ser organizada em três frentes: (i) estudos sobre exposição infantil, adultização e monetização da infância em ambientes digitais; (ii)

pesquisas sobre proteção de menores, regulação e riscos em plataformas; e (iii) investigações sobre circulação de conteúdos, engajamento e formatos em ecossistemas audiovisuais.

Na primeira frente, diferentes trabalhos têm mostrado que a exposição precoce de crianças a plataformas digitais pode produzir efeitos relevantes sobre privacidade, identidade e desenvolvimento. Estudos sobre *sharenting* indicam que o compartilhamento recorrente de imagens e informações de menores por responsáveis pode ser percebido como invasivo e comprometer a autonomia infantil [1, 16, 23, 42]. Em paralelo, pesquisas recentes destacam a crescente mercantilização da infância nas redes sociais, em especial quando crianças e adolescentes passam a atuar como ativos de visibilidade e de monetização [10]. No caso da adultização infantil, [4] ressaltam os efeitos psicossociais da erotização midiática e do consumo estético precoce, enquanto [3] e [34] mostram que conteúdos com maior exposição corporal de menores tendem a ser favorecidos em termos de circulação e engajamento.

Na segunda frente, a literatura tem enfatizado a insuficiência dos mecanismos atuais de proteção infantil diante da lógica de funcionamento das plataformas. [18] argumenta que algoritmos orientados por performance e visibilidade intensificam os riscos de exploração, mercantilização e violência simbólica contra crianças, ao passo que [11] mostra que adolescentes frequentemente internalizam a culpa pelos danos sofridos nas redes. No contexto brasileiro, [5] destaca o descompasso entre os instrumentos jurídicos existentes e a realidade digital, enquanto [37] defende que a proteção de menores exige não apenas atualização normativa, mas também responsabilização social e institucional pelo uso abusivo das mídias. Pesquisas também têm avançado na identificação automática de conteúdos nocivos, como demonstram os trabalhos de [24] sobre o YouTube Kids e de [35] sobre objetificação infantil no TikTok.

Na terceira frente, estudos sobre plataformas de vídeo têm mostrado que a circulação de conteúdos e as respostas da audiência variam conforme as métricas de engajamento, os formatos e os componentes multimodais. [25] demonstra que visualizações, curtidas e comentários estão associados ao tempo de permanência em vídeos no YouTube, enquanto [43] identifica diferenças relevantes entre *Shorts* e vídeos tradicionais. De modo complementar, [45] mostra que títulos, transcrições, áudio e elementos visuais influenciam as respostas emocionais da audiência. Em diálogo com essa literatura, estudos recentes [6, 8, 20, 29, 32, 36, 40] também têm explorado debates públicos mediados por plataformas audiovisuais sobre temas politicamente sensíveis.

Nosso trabalho se aproxima dessas três frentes ao tratar, simultaneamente, da vulnerabilidade infantil em ambientes digitais, das disputas em torno da regulação de plataformas e das dinâmicas de circulação discursiva no YouTube. Ao mesmo tempo, diferencia-se delas em um ponto central: em vez de focalizar apenas o risco, a proteção normativa ou o engajamento em abstrato, investigamos como uma controvérsia pública recente, em torno da adultização infantil e da Lei nº 15.211/2025, se organiza discursivamente na plataforma. Para isso, articulamos o discurso dos criadores, a repercussão da audiência nos comentários e a comparação entre vídeos

tradicionais e *Shorts*, oferecendo uma caracterização integrada de como temas, formatos e estilos linguísticos moldam a circulação desse debate no YouTube.

3 METODOLOGIA

Para realizar nossa análise, desenvolvemos um *pipeline* estruturado em seis etapas: (i) busca de vídeos e extração de metadados, (ii) coleta de comentários e filtragem de idioma, (iii) classificação do formato de consumo, (iv) transcrição automatizada dos áudios, (v) análise acústica e validação da qualidade das transcrições e (vi) processamento de linguagem natural, abrangendo modelagem de tópicos, análise de sentimentos e extração de características linguísticas.

3.1 Busca de Vídeos e Extração de Metadados

A coleta foi realizada por meio da YouTube Data API v3¹, utilizando os termos “Lei Felca”, “ECA Digital”, “Felca”, “Polêmica Felca”, “Adultização” e “15.211/2025”. O período de coleta compreendeu o intervalo entre 6 de agosto de 2025, data da denúncia inicial que impulsionou a controvérsia, e 26 de março de 2026. Como a API retorna cerca de 500 resultados por busca, implementamos uma estratégia de particionamento temporal em janelas de 15 dias, ampliando a cobertura da coleta. Após a unificação dos resultados e a remoção de duplicatas, obtivemos um corpus inicial de 10.267 vídeos únicos. Para cada vídeo, extraímos metadados descritivos e de engajamento, detalhados na Tabela 1.

Tabela 1: Metadados extraídos via YouTube Data API.

Atributo	Descrição
videoId	Identificador único do vídeo na plataforma.
title	Título oficial do vídeo.
channelTitle	Nome do canal do criador de conteúdo.
channelId	Identificador único do canal.
publishedAt	Data e hora exata da publicação.
description	Texto de descrição fornecido pelo criador.
tags	Palavras-chave associadas ao vídeo.
duration	Duração total do vídeo.
viewCount	Número total de visualizações.
likeCount	Número total de curtidas.
commentCount	Número total de comentários.
algorithmTopics	Categorias temáticas atribuídas pelo algoritmo do YouTube.
madeForKids	Flag indicando se o conteúdo é direcionado para crianças.
defaultAudioLanguage	Idioma principal detectado no áudio.

3.2 Coleta de Comentários e Filtro de Idioma

A etapa seguinte consistiu na coleta das interações do público. Dos 10.267 vídeos, 6.462 continham comentários habilitados e pelo menos uma interação. Limitamos a coleta a um máximo de 1.000 comentários por vídeo, privilegiando os comentários ranqueados pela própria plataforma. Essa decisão buscou equilibrar custo computacional e relevância analítica, afetando apenas uma fração reduzida da base, já que 75% dos vídeos possuíam 32 comentários ou menos. O processo resultou em 754.159 comentários. Para restringir a análise a vídeos em português, utilizamos inicialmente a métrica `defaultAudioLanguage` fornecida pela API, unificando as categorias “pt”, “pt-BR” e “pt-PT”. Ao final dessa etapa, seguimos com 9.270 vídeos validados em português.

¹<https://developers.google.com/youtube/v3/getting-started>

3.3 Shorts vs. Vídeos Tradicionais

Como a API do YouTube não informa diretamente se um conteúdo pertence ao formato *Shorts*, desenvolvemos uma rotina baseada no comportamento de roteamento da plataforma. Para cada `videoId`, realizamos requisições ao padrão `/shorts/videoId`; quando o conteúdo correspondia a um vídeo tradicional, o servidor redirecionava automaticamente a URL para o padrão `/watch?v=`. A classificação final foi, portanto, definida pela URL de destino. Com esse procedimento, identificamos 5.916 *YouTube Shorts* e 3.354 vídeos tradicionais no corpus final em português.

3.4 Download e Transcrição de Áudios

Para capturar o áudio dos vídeos, utilizamos a ferramenta *yt-dlp*, em linha com procedimentos adotados em estudos recentes sobre vídeo online [7, 29], para extrair as faixas de áudio dos conteúdos. Dos 9.270 vídeos em português, obtivemos com sucesso 8.973 arquivos de áudio. As falhas ocorreram, principalmente, em vídeos removidos, privatizados ou protegidos por restrições de acesso. A conversão da fala em texto foi realizada com o modelo *Whisper* [12]. Nessa etapa, 1.359 transcrições retornaram vazias e foram descartadas, em geral associadas a vídeos sem fala perceptível, com trilha sonora predominante ou com ruído intenso.

3.5 Limpeza dos Áudios pela Qualidade Acústica

Como a qualidade das transcrições é central para as análises subsequentes, realizamos uma etapa de validação acústica e textual, seguindo o trabalho [29]. Utilizando a biblioteca *Librosa*, extraímos oito características físicas do som: (i) duração total, (ii) frequência de amostragem, (iii) amplitude média, (iv) amplitude máxima, (v) razão entre amplitudes, (vi) relação sinal-ruído (SNR), (vii) *Zero Crossing Rate* e (viii) detecção de regiões de silêncio. Foram removidas 21 amostras com sinal nulo, resultando em 8.952 áudios válidos. Esses áudios foram agrupados com K-Means após normalização por *StandardScaler*, com o objetivo de representar adequadamente a diversidade acústica do corpus. O número de agrupamentos foi definido com base no índice BetaCV [27], testando valores de k entre 5 e 100, com incrementos de 5. O ponto ótimo foi identificado pelo método do cotovelo, resultando em $k = 80$.

A partir desses grupos, extraímos uma amostra estratificada de 600 áudios para validação manual. Seguindo recomendações da literatura [31], utilizamos apenas trechos com até 2 minutos, reduzindo a carga cognitiva dos avaliadores. Cada item foi anotado por três avaliadores independentes (os próprios autores do artigo) como “boa” ou “ruim”, com decisão final por maioria. A concordância entre anotadores foi medida pelo AC1 de Gwet [44], resultando em 0,78, o que indica boa concordância. Com base nessa amostra, treinamos um classificador *Random Forest* [26], com hiperparâmetros ajustados por meio de *Grid Search* e validação cruzada em 5 *folds*. O modelo obteve *Macro-F1* de 0,73, com os seguintes hiperparâmetros ótimos: `n_estimators=350`, `max_depth=9`, `min_samples_split=17`, `min_samples_leaf=4`, `class_weight='balanced'`, `max_features='log2'` e `max_samples=0.82`.

Aplicado aos áudios não anotados, o classificador rotulou 85,6% das transcrições como adequadas e 14,4% como inadequadas. Nas etapas seguintes, utilizamos apenas as transcrições consideradas adequadas, totalizando 6.506 instâncias.

3.6 Análise Estrutural e Linguística

Com a base textual validada, conduzimos três frentes de análise: modelagem de tópicos das transcrições, classificação de sentimentos dos comentários e extração de características linguísticas. Em todos os textos, o pré-processamento seguiu a rotina de limpeza validada por [28, 41], com seis operações sequenciais: (i) lematização, (ii) normalização para caixa baixa e remoção de acentuação, (iii) redução de caracteres repetidos, (iv) eliminação de URLs, numerais e pontuação, (v) descarte de *tokens* com até três caracteres e (vi) remoção de *stopwords*.

3.6.1 Modelagem de Tópicos e Sumarização do Discurso.

Para agrupar semanticamente as transcrições, utilizamos BERTopic [13], explorando diferentes combinações de parâmetros nos módulos de representação vetorial, redução de dimensionalidade e clusterização. Avaliamos dois modelos de *embeddings*: *paraphrase-multilingual-mpnet-base-v2* e *neuralmind/bert-base-portuguese-cased*. No UMAP, variamos *n_neighbors* em 10, 20, *n_components* em 5, 10 e *min_dist* em 0.0, 0.1, 0.5. Na etapa de clusterização, testamos K-Means com $k \in \{2, 3, 4, 5, 10, 20, 30, 50\}$. O desempenho de cada combinação foi comparado pelas métricas: coerência [22], diversidade [9] e *Silhouette Score* [33], complementadas por inspeção qualitativa dos modelos mais bem avaliados. A configuração final utilizou o modelo *neuralmind/bert-base-portuguese-cased*, UMAP com 20 vizinhos, distância mínima de 0,1 e 5 componentes, e K-Means com $k = 50$ tópicos. Para melhorar a interpretabilidade, aplicamos ainda uma filtragem por c-TF-IDF com lista ampliada de *stopwords* do domínio. Após a extração dos 50 tópicos, utilizamos o modelo *GPT-5.4 Thinking* para gerar rótulos, sumarizações e agrupamentos semânticos em macrotópicos. Para cada tópico, o modelo recebeu suas 10 palavras-chave mais representativas e uma amostra de 10 documentos, composta pelos três itens mais representativos segundo o BERTopic e sete amostras aleatórias do mesmo agrupamento. O *prompt* está ilustrado na Figura 1. Para assegurar a validade semântica do processo, os títulos e as descrições geradas pelo modelo foram validados manualmente pelos autores. Adicionalmente, conduziu-se uma inspeção qualitativa por meio da visualização de uma amostra de vídeos de cada macrotópico, confirmando a consistência entre os agrupamentos sugeridos e o conteúdo efetivamente abordado.

3.6.2 Classificação de Sentimentos nos Comentários.

Para identificar a postura da audiência, realizamos classificação de sentimentos dos comentários. Inicialmente, utilizamos um modelo RoBERTa para descartar automaticamente os comentários não escritos em português. Em seguida, construímos uma base anotada de 1.000 comentários, classificados em positivo, negativo ou neutro. Cada comentário foi inicialmente anotado por cinco avaliadores (autores do artigo); os casos sem consenso foram submetidos a um sexto anotador. A concordância entre anotadores, medida pelo AC1 de Gwet

Prompt

```
# CONTEXT You are a data analyst specializing in social media dynamics. I am analyzing a corpus of 6432 transcripts extracted from YouTube videos that discuss the 'adultification of childhood,' the exploitation of minors on the internet, and the mobilization for digital regulation known as the 'Felca Law.' The data collection period spans from August 2025 to March 2026.
I applied topic modeling (BERTopic with embeddings in Portuguese + UMAP + K-Means) and obtained 50 topics. Each topic, presented below, is represented by its 10 most characteristic keywords and 10 document examples (3 being the most mathematically representative and 7 randomly selected). The keywords and transcripts are written in Brazilian Portuguese.
# TASK Based on the keywords and ad examples for each topic, do TWO things:
1) SUMMARY — for EACH topic, provide a short label (3 to 6 words) and a one-line description explaining what the topic is about.
2) MACRO-TOPICS — propose a grouping of the topics into coherent macro-topics (macro-themes). Suggest an appropriate number of macro-topics (generally between 5 and 12) and justify it. For each macro-topic, list the IDs of the topics that compose it and explain in 1 to 2 lines the semantic criterion that unites them.
# IMPORTANT NOTES - Some topics may NOT be relevant to the central theme: greetings, generic engagement requests ('subscribe to the channel', 'like'), poorly transcribed audio noise, or videos unrelated to the topic. Identify them explicitly and group them into a separate category (e.g., 'Digital Noise / Generic Engagement'), instead of forcing them to fit into a macro-theme.
- Prioritize grouping by the SUBSTANTIVE AGENDA OF THE DEBATE (e.g., regulation and laws, platform responsibility, parental/guardian negligence, mental health and impacts on childhood, specific influencer cases, etc.).
- IMPORTANT: write your ENTIRE response in Brazilian Portuguese (pt-BR), even if these instructions are in English.
# RESPONSE FORMAT (content in Portuguese) ## Parte 1 — Rótulo dos Tópicos Tópico {id} | {rótulo} | {descrição de uma linha}
## Parte 2 — Macro-tópicos propostos MACRO {nome do macrotópico} Tópicos: {lista de IDs} Critério: {1-2 linhas}
## Parte 3 — Tópicos não-substantivos (ruídos / engajamento genérico) {lista de IDs+ breve justificativa}
```

Figura 1: *Prompt* utilizado para sumarização dos tópicos.

[44], foi de 0,57, indicando concordância moderada. Para a classificação automática, utilizamos o modelo *BERTimbau* [38], escolhido por sua adequação ao português brasileiro. Os hiperparâmetros foram ajustados por otimização bayesiana via *Optuna*, resultando em taxa de aprendizado de $1,83 \times 10^{-5}$, *batch size* 8, *weight decay* de 0,024 e 4 épocas. O modelo atingiu *Macro-F1* médio de 0,7019 em validação cruzada, com desvio padrão de $\pm 0,0193$.

3.6.3 Extração de Características Linguísticas e Seleção de Variáveis.

Por fim, extraímos métricas linguísticas das transcrições e dos comentários com a biblioteca *Linguistic Feature Toolkit* (LFTK) [17]. Nas transcrições, essas métricas foram utilizadas para comparar diferenças entre macrotópicos e entre formatos de vídeo. Nos comentários, foram empregadas para analisar como a estrutura linguística varia entre classes de sentimento. Para reduzir dimensionalidade e garantir robustez estatística, aplicamos um *pipeline* em duas etapas. Primeiro, ranqueamos as variáveis por *Gini Gain* [30, 46]. Em seguida, removemos iterativamente a colinearidade com base na correlação de Spearman, adotando limiar de $\rho \geq 0.7$. Na etapa inferencial, utilizamos o teste de Kruskal-Wallis [21]; quando aplicável, diferenças entre grupos foram detalhadas com o teste *post-hoc* de Dunn com correção de Bonferroni.

4 RESULTADOS

Esta seção apresenta os resultados em três níveis complementares. Primeiro, caracterizamos a estrutura geral do debate por meio dos macrotópicos extraídos das transcrições e da distribuição dos sentimentos nos comentários. Em seguida,

analisamos como os enquadramentos temáticos se relacionam às reações da audiência (RQ1), como os formatos de publicação moldam a circulação do debate (RQ2) e, por fim, como o ecossistema se diferencia linguisticamente em função do tema, do sentimento e do formato (RQ3).

4.1 Caracterização geral do debate

A aplicação do modelo *BERTopic* às transcrições dos vídeos resultou na identificação de 50 tópicos distintos. Após a etapa de sumarização e agrupamento semântico via LLM, parte desses *clusters* foi classificada como ruído digital (e.g., pedidos genéricos de inscrição, trilhas sem fala ou conteúdos não relacionados) e removida da análise. Com isso, seguimos com 5.011 documentos, consolidados em 9 macrotópicos que sintetizam os principais eixos do debate sobre a regulação no YouTube:

- **Regulação, ECA Digital e tramitação legislativa (1064 vídeos):** debates sobre marco legal, tramitação, fiscalização e atualização do ECA para o ambiente digital.
- **Felca como ator midiático e símbolo público (632 vídeos):** conteúdos sobre a centralidade de Felca no debate, sua repercussão pública e disputas de legitimidade.
- **Exploração, abuso e casos de exposição de menores (615 vídeos):** denúncias, casos concretos e discussões sobre exploração infantil, pedofilia e responsabilização.
- **Roblox, chat e mobilização de usuários (601 vídeos):** reações de usuários a possíveis impactos da lei nos jogos, nos chats e nas práticas de sociabilidade digital.
- **Adultização, desenvolvimento infantil e impactos psicossociais (567 vídeos):** efeitos da adultização na saúde mental, no comportamento e desenvolvimento infantil.
- **Privacidade, verificação etária e infraestrutura técnica (493 vídeos):** debates sobre biometria, reconhecimento facial, mecanismos técnicos de controle etário.
- **Games, loot boxes e mercado gamer (483 vídeos):** discussões sobre jogos, monetização, recompensas aleatórias e adaptações do mercado à regulação.
- **Polarização política, censura e guerra cultural (412 vídeos):** interpretações da lei como disputa ideológica, censura e liberdade de expressão.
- **Educação, prevenção e mediação social (144 vídeos):** conteúdos voltados à prevenção, internet segura, mediação educativa e campanhas de conscientização.

No que tange à audiência, a etapa de análise de sentimentos atuou nos comentários únicos vinculados aos vídeos desses 9 macrotópicos, resultando em um corpus final de 305.517 interações. A aplicação do classificador *BERTimbau* revelou que o debate é marcado por um tom majoritariamente neutro e crítico: 103.357 comentários (33.83%) foram classificados como negativos, 162.071 (53.05%) como neutros e 40.089 (13.12%) como positivos. Com essa base estrutural, as subseções seguintes articulam as dimensões temáticas, emocionais e linguísticas para responder às três perguntas de pesquisa.

4.2 RQ1: Enquadramentos temáticos e Audiência

Para responder à RQ1, cruzamos os macrotópicos extraídos das transcrições com os sentimentos predominantes nos comentários. Os resultados mostram que o enquadramento

temático adotado pelos criadores se associa diretamente à forma como a audiência reage ao conteúdo, tanto em termos de tom emocional quanto de intensidade da controvérsia.

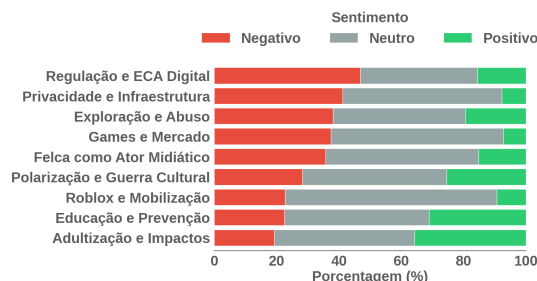


Figura 2: Proporção de Sentimentos por Macrotópico.

Conforme apresentado na Figura 2, os macrotópicos “Regulação e ECA Digital” e “Privacidade e Infraestrutura” lideram a proporção de comentários negativos. Nesse contexto, a negatividade não necessariamente representa a rejeição ao vídeo, mas sim a indignação coletiva diante da impunidade, de falhas nas plataformas ou de preocupações com os possíveis impactos das medidas propostas, como o receio de fornecer dados pessoais para a verificação etária. Em contraste, os macrotópicos “Educação e Prevenção” e “Adultização e Impactos” apresentam maior proporção de comentários positivos, indicando uma audiência mais engajada em propor soluções e demonstrar solidariedade. Os resultados também mostram que alguns temas atuam como focos privilegiados de dissenso. Macrotópicos como “Polarização política, censura e guerra cultural”, “Adultização, desenvolvimento infantil e impactos psicossociais” e “Educação, prevenção e mediação social” apresentam proporções expressivas de comentários positivos e negativos, indicando forte disputa interpretativa. Nesses casos, o debate parece menos orientado por consenso moral e mais por controvérsia quanto às causas, aos limites e aos efeitos da regulação.

Na Figura 3 apresentamos a distribuição da negatividade por vídeo para cada macrotópico. Podemos observar que, em temas regulatórios, a crítica da audiência não se concentra em poucos conteúdos virais, mas se distribui de forma relativamente consistente ao longo do conjunto. Em outras palavras, o tom negativo associado a certos macrotópicos é uma característica estrutural do enquadramento temático, e não um artefato decorrente de casos isolados. Isso sugere que, no YouTube, o tema escolhido pelo criador não apenas organiza o conteúdo do vídeo, mas também condiciona o horizonte emocional da discussão que se forma nos comentários.

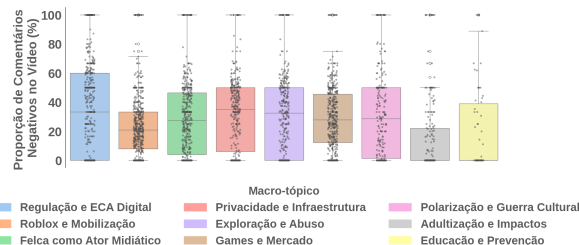


Figura 3: Consistência do Sentimento Negativo Intra-Tópico.

4.3 RQ2: Formato de publicação e Debate

Para responder à **RQ2**, analisamos de que modo os formatos de publicação (i.e. vídeos tradicionais e *YouTube Shorts*) influenciam a circulação do debate em termos de temas abordados, volume de engajamento e propensão à interação. Os resultados mostram que os dois formatos cumprem papéis distintos na plataforma: enquanto os vídeos tradicionais concentram conteúdos mais densos e maior conversão em debate, os *Shorts* favorecem visibilidade rápida e disseminação de controvérsias, mas com menor aprofundamento interacional.

A primeira evidência desse padrão aparece na própria composição do corpus recuperado pelo termo de busca. Como mostra a Figura 4, buscas associadas a termos técnico-jurídicos, como “15.211/2025” e “ECA Digital”, retornam, majoritariamente, vídeos tradicionais. Em contrapartida, buscas vinculadas à repercussão pública e à figura do influenciador, como “Felca” e “Polêmica Felca”, são dominadas por *Shorts*. Esse contraste sugere uma adequação estrutural entre pauta e formato: discussões sobre regulação e seus mecanismos tendem a demandar maior duração e encadeamento argumentativo, ao passo que polêmicas e repercussões imediatas se ajustam melhor à lógica breve e reativa dos vídeos curtos.

Esse padrão se mantém quando observamos o conteúdo efetivamente abordado nos vídeos. A Figura 5 mostra que os vídeos tradicionais concentram, com maior frequência, macro-tópicos como “Regulação e ECA Digital” e “Privacidade

e Infraestrutura”, associados a explicações técnicas, à tramitação legislativa e a implicações regulatórias. Já os *Shorts* apresentam maior incidência de “Felca como Ator Midiático” e “Roblox e Mobilização”, indicando maior aderência a conteúdos centrados na repercussão, na reação e na circulação de tendências. Em outras palavras, os formatos não apenas distribuem o mesmo debate em velocidades diferentes, mas também selecionam tipos distintos de enquadramento temático.

A diferença entre os formatos também se reflete no padrão de engajamento. Como apresentado na Figura 6, os *Shorts* alcançam, em média, um volume maior de visualizações, o que confirma sua capacidade de ampliar rapidamente a exposição dos conteúdos. No entanto, essa vantagem em termos de alcance não se traduz em maior participação discursiva. Os vídeos tradicionais concentram, em média, um número maior de comentários, o que indica maior capacidade de transformar a visualização em interação efetiva. Esse resultado sugere que alcance e debate não são fenômenos equivalentes na plataforma: um formato pode maximizar difusão sem, necessariamente, fomentar discussão.

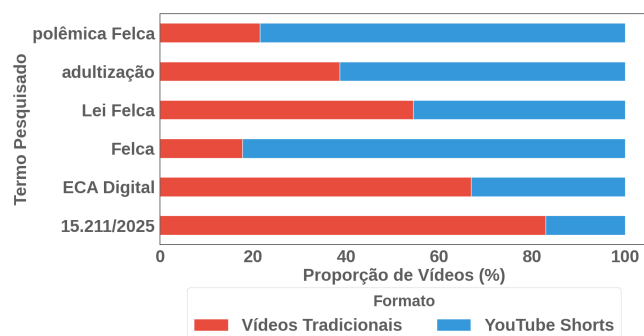


Figura 4: Proporção de formatos de acordo com a busca.

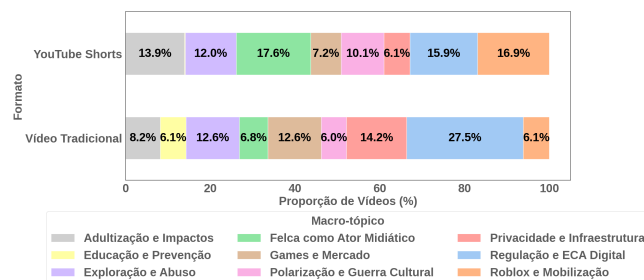


Figura 5: Distribuição dos macro-tópicos por formato de vídeo.

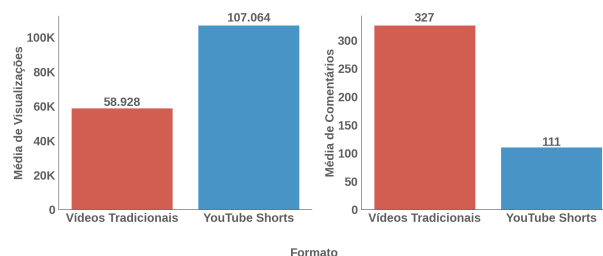


Figura 6: Média de visualizações e de comentários por formato.

Esse comportamento torna-se ainda mais claro quando observamos a taxa de conversão entre visualizações e comentários. A Figura 7 apresenta a distribuição acumulada dessa taxa e evidencia que a ampla maioria dos *Shorts* possui conversão próxima de zero. Já os vídeos tradicionais exibem uma distribuição mais deslocada para valores superiores, indicando maior propensão a transformar audiência em conversa. O teste de Mann-Whitney U confirmou que essa diferença é estatisticamente significativa ($p < 0.001$), reforçando que os formatos se distinguem não apenas em volume bruto de audiência, mas na capacidade de sustentar engajamento discursivo.

Em conjunto, esses resultados evidenciam um *paradoxo do engajamento*: os *Shorts* operam como mecanismo eficiente de visibilidade e rápida circulação, mas pouco convertem esse alcance em debate efetivo; os vídeos tradicionais, por sua vez, alcançam menos visualizações médias, mas concentram as condições necessárias para discussão mais densa e participativa. Assim, o formato de publicação não atua apenas como embalagem do conteúdo, mas como uma dimensão estruturante da própria controvérsia, influenciando quais temas circulam com mais força e em que medida a audiência se engaja com eles de forma argumentativa.

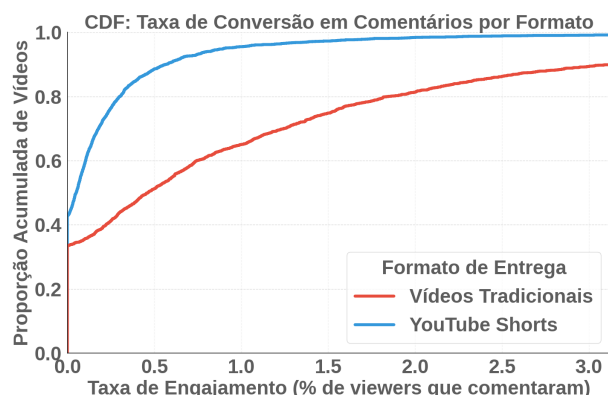


Figura 7: Taxa de conversão de visualizações em comentários.

4.4 RQ3: Arquitetura linguística do debate

Para responder à RQ3, analisamos como o discurso dos criadores e a resposta da audiência variam linguisticamente em função do tema, do formato do vídeo e da polarização dos comentários. Os testes de Kruskal-Wallis confirmaram diferenças estatisticamente significativas ($p < 0.001$) em todas as comparações, indicando que a linguagem mobilizada se adapta sistematicamente às condições de circulação do debate.

A Tabela 2 mostra que os criadores modulam sua linguagem conforme o enquadramento temático do vídeo. O macrotópico “Regulação e ECA Digital” apresenta os maiores níveis de complexidade textual, com destaque para o índice de legibilidade (*índice de escolaridade* = 8.803), a maior proporção de adposições por palavra (0.132) e a maior frequência relativa de adjetivos (*proporção de adjetivos* = 0.057). Em conjunto, essas métricas sugerem um discurso mais denso, impessoal e estruturado, compatível com conteúdos voltados à explicação normativa e à interpretação técnico-jurídica da lei. Em contraste, macrotópicos como “Felca como Ator Midiático” e “Roblox e Mobilização” apresentam os menores valores de complexidade e maior incidência de marcas interacionais, como a maior proporção de pronomes (0.102), o que reflete uma linguagem mais direta, coloquial e orientada à proximidade com a audiência. Além disso, alguns temas apresentam assinaturas linguísticas próprias. “Educação e Prevenção”, por exemplo, registra o maior volume médio de palavras com três ou mais sílabas (405.889), sugerindo um vocabulário mais elaborado e voltado à mediação pedagógica. Já “Roblox e Mobilização” combina baixa complexidade geral com maior frequência de nomes próprios (0.064), o que indica forte ancoragem em entidades, personagens e referências específicas do universo gamer. Assim, a linguagem dos criadores varia não apenas em complexidade, mas também em função do tipo de enquadramento que cada tema exige.

No caso da audiência, a polarização também se manifesta linguisticamente. A Tabela 3 mostra que comentários negativos tendem a ser mais longos, lexicalmente mais diversos e estruturalmente mais densos do que os comentários positivos e neutros. Eles apresentam, por exemplo, médias mais altas em palavras com 2 ou mais sílabas (5.089), diversidade lexical (44.023), número total de determinantes (2.898) e total de

pontuações (2.972), o que sugere textos mais elaborados e com maior esforço argumentativo. Em vez de simples reações impulsivas, a negatividade aparece aqui frequentemente como crítica elaborada, sustentada por vocabulário mais variado e por cadeias discursivas mais longas. Em contraste, os comentários positivos tendem a ser mais curtos, menos diversos e mais formulaicos, embora apresentem maior proporção relativa de adjetivos (0.074), sugerindo avaliações mais sintéticas e valorativas. Já os comentários neutros ocupam, em geral, uma posição intermediária entre esses dois polos. Esse padrão é importante porque mostra que, nesse debate, a polarização não se expressa apenas na orientação do sentimento, mas também na forma da linguagem: a crítica mobiliza maior elaboração textual, enquanto o apoio tende a se condensar em formas discursivas mais breves.

Por fim, a Tabela 4 evidencia que a própria gramática do discurso é afetada pela arquitetura temporal da plataforma. Os vídeos tradicionais apresentam maiores volumes de palavras longas (248.742) e maior presença de interjeições (13.608), o que é compatível com uma fala mais extensa, pausada e próxima da conversação natural. Já os *Shorts* exibem maior variação gramatical em classes como verbos (0.777), adjetivos (0.837), auxiliares (0.936) e conjunções coordenativas (0.854), indicando uma fala mais condensada e sintaticamente compacta, adaptada à retenção imediata da atenção. Em outras palavras, o formato curto não apenas reduz a duração do discurso: ele impõe uma reorganização estrutural da própria linguagem.

Em conjunto, os resultados das Tabelas 2, 3 e 4 mostram que a arquitetura linguística do debate é moldada por três fatores centrais: o tema, que orienta o grau de tecnicidade e formalidade do discurso dos criadores; o sentimento, que afeta a densidade argumentativa da audiência; e o formato, que impõe ritmos e estruturas distintas à fala. Assim, a controvérsia sobre a Lei nº 15.211/2025 no YouTube é tematizada de diferentes formas e também linguisticamente configurada pelas condições sociotécnicas de circulação da plataforma.

5 CONCLUSÃO E TRABALHOS FUTUROS

Este artigo apresentou uma análise computacional do debate sobre adultização infantil e a Lei nº 15.211/2025 no YouTube, articulando o discurso dos criadores de conteúdo, a reação da audiência e o papel dos diferentes formatos de publicação na circulação da controvérsia. Ao combinar transcrições de vídeos, modelagem de tópicos, análises de sentimentos e linguística, o estudo permitiu caracterizar não apenas os principais enquadramentos temáticos do debate, mas também as formas pelas quais esses enquadramentos mobilizam diferentes padrões de engajamento e de resposta pública.

Os resultados mostraram que a controvérsia se organiza em torno de enquadramentos temáticos distintos, como regulação, privacidade, exploração de menores, impactos psicossociais, universo gamer e polarização política, e que esses enquadramentos influenciam diretamente a forma como a audiência reage nos comentários. Também evidenciamos que vídeos tradicionais e *YouTube Shorts* desempenham papéis distintos na plataforma: enquanto os primeiros concentram discussões

Característica	Adultização e Impactos	Educação e Prevenção	Exploração e Abuso	Felca como Ator Midiático	Games e Mercado	Polarização e Guerra Cultural	Privacidade e Infraestrutura	Regulação e ECA Digital	Roblox e Mobilização
Índice de Escolaridade	6.806	7.294	6.641	3.554	5.673	5.166	6.232	8.803	3.820
Proporção de Adposições	0.095	0.107	0.107	0.081	0.100	0.096	0.101	0.132	0.092
Proporção de Pronomes	0.089	0.071	0.077	0.102	0.072	0.083	0.077	0.052	0.079
Proporção de Adjetivos	0.045	0.046	0.041	0.035	0.038	0.038	0.043	0.057	0.033
Palavras com 3 ou mais sílabas	90.679	405.889	136.797	51.440	80.449	133.325	119.154	198.078	25.952
Proporção de Nomes Próprios	0.012	0.024	0.037	0.030	0.052	0.041	0.029	0.039	0.064
Frequência Lexical por Sentença	30.890	47.800	31.989	27.602	51.750	28.451	43.960	38.121	24.939
Frequência Lexical por Palavra	1.354	1.435	1.426	1.317	1.468	1.439	1.430	1.564	1.449
Proporção de Verbos	0.129	0.119	0.129	0.137	0.132	0.130	0.130	0.112	0.131
Proporção de Pontuação	0.131	0.133	0.129	0.153	0.122	0.135	0.132	0.113	0.141

Tabela 2: Valores médios das características linguísticas (LFTK) por macrotópico nas transcrições. Destaques em verde indicam o maior valor da métrica, enquanto destaques em vermelho indicam o menor valor.

Característica	Negativo	Neutro	Positivo
Palavras com 2 ou mais sílabas	5.089	3.523	3.087
Total de Determinantes	2.898	2.097	1.794
Diversidade Lexical	44.023	33.211	29.269
Dificuldade de Leitura	3.648	2.689	2.597
Total de Pontuações	2.972	2.060	2.080
Variação de Pontuações	0.459	0.387	0.418
Proporção de Substantivos	2.878	2.393	1.903
Proporção de Adjetivos	0.053	0.045	0.074
Variação de Auxiliares	0.542	0.436	0.386
Variação de Verbos	1.667	1.392	1.184

Tabela 3: Valores médios das características linguísticas por sentimento. Destaques em verde indicam o maior valor, e em vermelho, o menor.

Característica	Tradicionais	Shorts	p-valor
Palavras com 3 ou mais sílabas	248.742	12.798	< 0.001
Variação de Verbos	0.488	0.777	< 0.001
Variação de Adjetivos	0.631	0.837	< 0.001
Variação de Auxiliares	0.598	0.936	< 0.001
Variação de Conjunções	0.544	0.854	< 0.001
Variação de Advérbios	3.156	2.347	< 0.001
Total de Interjeições	13.608	0.647	< 0.001
Variação de Símbolos	0.247	0.039	< 0.001
Variação de Adposições	1.442	1.815	< 0.001
Variação de Numerais	0.601	0.551	0.165

Tabela 4: Valores médios das características linguísticas (LFTK) comparando Vídeos Normais e *Shorts*. Destaques em verde indicam o maior valor da métrica, enquanto destaques em vermelho indicam o menor valor.

mais densas e argumentativas, os segundos favorecem visibilidade rápida e disseminação de controvérsias, mas com menor conversão em debate efetivo. Mostramos que a arquitetura linguística do debate também varia de forma sistemática, tanto no discurso dos criadores quanto na resposta da audiência, em função do tema, do formato e do grau de polarização.

Esses achados reforçam que o YouTube não atua como um hospedeiro neutro de controvérsias públicas, mas como um ambiente sociotécnico que condiciona a forma, a intensidade

e a profundidade com que debates socialmente sensíveis são enquadrados, amplificados e disputados. Nesse sentido, o artigo contribui ao oferecer uma caracterização computacional de um debate recente sobre infância e regulação digital, ao integrar vídeos e comentários em uma mesma análise e ao demonstrar que os formatos de circulação da plataforma influenciam não apenas o alcance dos conteúdos, mas também a qualidade do engajamento e os estilos discursivos mobilizados.

Como desdobramento, futuras investigações podem ampliar essa agenda em pelo menos três direções. A primeira consiste em comparar a dinâmica observada no YouTube com a de plataformas nativamente orientadas ao vídeo curto, como TikTok e Instagram *Reels*, permitindo avaliar em que medida o paradoxo do engajamento se mantém em ecossistemas distintos. A segunda envolve incorporar abordagens multimodais e técnicas de visão computacional, a fim de examinar também como a adultização infantil é construída visualmente nos vídeos. A terceira diz respeito à ampliação longitudinal da coleta, possibilitando acompanhar como enquadramentos, formatos e reações da audiência evoluem ao longo do tempo à medida que a controvérsia se reconfigura publicamente. Mais do que um caso pontual, o debate em torno da Lei nº 15.211/2025 mostra como plataformas audiovisuais se tornaram espaços centrais para a disputa pública sobre proteção infantil, regulação digital e os limites éticos da visibilidade no ambiente online.

6 LIMITAÇÕES

O escopo da pesquisa restringe-se ao YouTube, o que limita a generalização dos achados a redes com lógicas algorítmicas distintas, como TikTok ou Instagram. A coleta também limitou-se a palavras-chaves pré-definidas e aos 1.000 comentários mais bem ranqueados por vídeo. Embora exclua opiniões periféricas, essa escolha garante uma amostra representativa focada nas discussões de maior visibilidade e impacto efetivo na audiência **segundo os critérios do YouTube**. Por fim, modelos como o BERTimbau carregam vieses de treinamento que podem limitar a detecção de ironias e nuances características da internet.

7 CUIDADOS ÉTICOS

Este estudo utilizou exclusivamente dados publicamente disponíveis no YouTube, incluindo metadados, transcrições automáticas de vídeos e comentários públicos relacionados ao debate sobre a adultização infantil e à Lei nº 15.211/2025. A análise foi conduzida em nível agregado, com foco em padrões temáticos, emocionais e linguísticos, sem tentativa de reidentificação, exposição individual ou inferência sobre as identidades reais dos usuários para além das informações publicamente acessíveis na plataforma. Dessa forma, o tratamento dos dados observou os princípios de finalidade, minimização e necessidade previstos na Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018). Por se basear em informações de acesso público e não envolver intervenção ou interação direta com participantes humanos, o estudo também se enquadra nas diretrizes da Resolução CNS nº 510/2016.

ACKNOWLEDGMENTS

Esta pesquisa foi financiada por: CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde e INCT-TILD-IAR (grant 408490/2024-1). Ferramentas de Inteligência Artificial Generativa foram utilizadas como suporte à revisão gramatical do texto e, conforme detalhado na Seção 3.6.1, na etapa metodológica de sumarização e agrupamento semântico dos tópicos. Os autores realizaram uma auditoria humana e uma revisão completa de todo o conteúdo gerado pelo modelo, assumindo total responsabilidade pela integridade das informações apresentadas.

REFERÊNCIAS

- [1] Mary Jean Amon, Nika Kartvelishvili, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2022. Sharenting and Children's Privacy in the United States: Parenting Style, Practices, and Perspectives on Sharing Young Children's Photos on Social Media. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 116 (April 2022), 30 pages. <https://doi.org/10.1145/3512963>
- [2] 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*.
- [3] Encarnación Ayala and Verónica Cala. 2022. TikTok and Child Hypersexualization: Analysis of Videos and Narratives of Minors. *American Journal of Sexuality Education* (07 2022). <https://doi.org/10.1080/15546128.2022.2096734>
- [4] Giulia Casagrande Bonilha, Celso Taques Saldanha, Alberto Stessel Sadalla Peres, Rafael Pimentel Saldanha, Julia Keiko Kofuji, Ana Júlia Loubet da Costa Cardoso, Ryllare Tawane Dourado Barros, Ana Karoline Lima do Nascimento, Rebeca Soares Silva, and Maria Eduarda Benevides dos Anjos. 2025. Adultização precoce: revisão narrativa crítica de um fenômeno multifatorial na infância e adolescência. *ARACÊ* 7, 10 (2025), e8774–e8774.
- [5] Ana Luiza Bornia Carbonieri, Maria Eloisa Benevides da Silva, and Thiago Silva Prado. 2025. Adultização infantil no Brasil e o trabalho de menores nas redes sociais digitais. *Aurum Editora* (2025), 130–141.
- [6] Raíssa Carvalho and Humberto Marques-Neto. 2024. Crianças e Propagandas no TikTok: identificando publicidade infantil na rede social TikTok. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 98–105. <https://doi.org/10.5753/webmedia.2024.242912>
- [7] 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)*.
- [8] Aline Dias, Richardy Tanure, Jussara Almeida, Helen Lima, and Carlos Ferreira. 2024. Análise da Percepção do Uso de Cigarros Eletrônicos no Brasil por meio de Comentários no YouTube. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 45–53. <https://doi.org/10.5753/webmedia.2024.241958>
- [9] Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* 8 (2020).
- [10] Tom Divon, Taylor Annabell, and Catalina Goanta. 2025. Children as concealed commodities: Ethnographic nuances and legal implications of kidfluencers' monetisation on TikTok. *New Media Society* 28 (01 2025), 1190–1218. <https://doi.org/10.1177/1461448241304657>
- [11] Ritika Gairola, Colin M Gray, Jingxin Dong, Kyung Jin Jeong, Ege Otenen, and Juan J Sarria. 2026. "Social Media Killed Our Generation": Teenagers' Felt Experiences of Harm on Social Media. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [12] 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).
- [13] Maarten Grootendorst. 2024. BERTopic - Parameter Tuning. https://maartengr.github.io/BERTopic/getting_started/parameter%20tuning/parametertuning.html Accessed: July 2025.
- [14] 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).
- [15] IBOPE. 2023. Video Audience Share Percentage in Brazil. <https://kantaribopemedia.com/conteudo/relatorios/april-2023/>.
- [16] Kamil Kopecky, Rene Szotkowski, Inmaculada Aznar-Díaz, and José-Maria Romero-Rodríguez. 2020. The phenomenon of sharenting and its risks in the online environment. Experiences from Czech Republic and Spain. *Children and Youth Services Review* 110 (2020), 104812. <https://doi.org/10.1016/j.childyouth.2020.104812>
- [17] Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted Features in Computational Linguistics. (July 2023), 1–19. <https://aclanthology.org/2023.bea-1.1>
- [18] REGINALDO SOARES DE SOUSA LIMA. 2025. Vulnerabilidade digital e riscos da adultização de menores em plataformas de mídia social. *Periódicos Brasil. Pesquisa Científica* 4, 2 (2025), 321–331.
- [19] Sonia Livingstone, Kjartan Ólafsson, and Elisabeth Staksrud. 2011. Social networking, age and privacy.
- [20] Victor Martins, Sophia Serafim, Lívia Pereira, Amanda Alves, Carlos Ferreira, and Jussara Almeida. 2025. A Misoginia no YouTube Brasileiro: Um Estudo de Caso sobre o Conteúdo Produzido pela Comunidade Red Pill. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 28–37. <https://doi.org/10.5753/webmedia.2025.16070>
- [21] Patrick E McKight and Julius Najab. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology* (2010), 1–1.
- [22] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proc. EMNLP '11*.
- [23] Gaëlle Ouvrein and Karen Verswijvel. 2019. Sharenting: Parental adoration or public humiliation? A focus group study on adolescents' experiences with sharenting against the background of their own impression management. *Children and Youth Services Review* 99 (2019), 319–327. <https://doi.org/10.1016/j.childyouth.2019.02.011>
- [24] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 522–533.
- [25] M. Park, M. Naaman, and J. Berger. 2021. A Data-Driven Study of View Duration on YouTube. *Proceedings of the International AAAI Conference on Web and Social Media* 10, 1 (2021), 651–654. <https://doi.org/10.1609/icwsm.v10i1.14781>
- [26] Aakash Parmar, Rakesh Kataria, and Vatsal Patel. 2018. A review on random forest: An ensemble classifier. In *International conference on intelligent data communication technologies and internet of things*. Springer, 758–763.

- [27] 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. <https://doi.org/10.1016/j.knosys.2025.113634>
- [28] 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. <https://doi.org/10.1145/3617023.3617040>
- [29] Gabriel Prenassi, Ana Machado, Ester Souza, Mateus Brito, Davi Reis, Jessica Costa, Geovana Oliveira, Carlos Ferreira, and Leonardo Rocha. 2025. Narrativas de Jogos de Azar em Plataformas de Vídeo: Um Estudo Linguístico-Temático sobre o jogo do Tigriho no YouTube. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC, 366–375.
- [30] 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. <https://doi.org/10.1023/B:AMAI.0000018580.96245.c6>
- [31] Melanie Revilla and Carlos Ochoa. 2017. Ideal and maximum length for a web survey. *International Journal of Market Research* 59, 5 (2017), 557–565.
- [32] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira Jr. 2019. Auditing Radicalization Pathways on YouTube. *CoRR* abs/1908.08313 (2019). arXiv:1908.08313 <http://arxiv.org/abs/1908.08313>
- [33] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987).
- [34] Miriam Schirmer, Angelina Voggenreiter, and Jürgen Pfeffer. 2024. More Skin, More Likes! Measuring Child Exposure and User Engagement on TikTok. arXiv:2408.05622 [cs.CY] <https://arxiv.org/abs/2408.05622>
- [35] Miriam Schirmer, Angelina Voggenreiter, Jürgen Pfeffer, and Agnes Horvat. 2025. Detecting child objectification on social media: Challenges in language modeling. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*. 396–412.
- [36] Cristian A. Silva and Daniel M. Barbosa. 2019. Analyzing the Acceptance of the 2018 Brazilian Presidential Election' Main Candidates Based on YouTube Comments. In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro). SBC, Porto Alegre, RS, Brasil, 377–384. <https://sol.sbc.org.br/index.php/webmedia/article/view/8050>
- [37] Alethea Sommer, Angélica Sassi, Bruna Nardon, Márcio Von Die-men, Rafaela Kunrath, Greice Carvalho, and Henrique Alexander Keske. 2026. ACELERANDO A INFÂNCIA NO MEIO DIGITAL: IMPACTOS PSICOLÓGICOS DA ADULTIZAÇÃO EM CRIANÇAS. *Periódicos Brasil. Pesquisa Científica* 5, 2 (2026), 920–942.
- [38] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERT-Timbau: pretrained BERT models for Brazilian Portuguese. In *Brazilian conference on intelligent systems*. Springer, 403–417.
- [39] Di "Chelsea" Sun, Vaishnavi Melkote, and Ahmed Sabbir Arif. 2020. Exploratory study of young children's social media needs and requirements. In *Proceedings of the 2020 ACM Interaction Design and Children Conference: Extended Abstracts* (London, United Kingdom) (*IDC '20*). Association for Computing Machinery, New York, NY, USA, 332–337. <https://doi.org/10.1145/3397617.3397836>
- [40] Richardy Tanure, Aline Dias, Lucas Camelo, Jussara Almeida, Helen Lima, and Carlos Ferreira. 2025. Caracterização do debate online sobre cigarro eletrônico no Brasil: Uma análise de tópicos de discussão no YouTube. In *Anais do XIV Brazilian Workshop on Social Network Analysis and Mining* (Maceió/AL). SBC, Porto Alegre, RS, Brasil, 54–64. <https://doi.org/10.5753/brasnam.2025.8544>
- [41] Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information processing & management* 50, 1 (2014), 104–112.
- [42] Karen Verswijvel, Michel Walrave, Kris Hardies, and Wannes Heirman. 2019. Sharenting, is it a good or a bad thing? Understanding how adolescents think and feel about sharenting on social network sites. *Children and Youth Services Review* 104 (2019), 104401. <https://doi.org/10.1016/j.childyouth.2019.104401>
- [43] Caroline Violot, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. 2024. Shorts vs. Regular Videos on YouTube: A Comparative Analysis of User Engagement and Content Creation Trends. In *Proceedings of the 16th ACM Web Science Conference* (Stuttgart, Germany) (*WEBSCI '24*). Association for Computing Machinery, New York, NY, USA, 213–223. <https://doi.org/10.1145/3614419.3644023>
- [44] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology* 13, 1 (2013), 61.
- [45] Niloofar Yousefi, Mert Can Cakmak, and Nitin Agarwal. 2024. Examining Multimodal Emotion Assessment and Resonance with Audience on YouTube. In *Proceedings of the 2024 9th International Conference on Multimedia and Image Processing* (Osaka, Japan) (*ICMIP '24*). Association for Computing Machinery, New York, NY, USA, 85–93. <https://doi.org/10.1145/3665026.3665039>
- [46] Ye Yuan, Liji Wu, and Xiangmin Zhang. 2021. Gini-Impurity Index Analysis. *IEEE Transactions on Information Forensics and Security* 16 (2021), 3154–3169.