Ecos de Radicalização: Auditando o Telegram como Veículo Difusor de Conteúdo de Radicalização Proveniente do YouTube

Marcelo M. R. Araújo

Abstract

Este trabalho explora o uso do Telegram como ferramenta auxiliar na disseminação de conteúdo radicalizador através de vídeos provenientes de plataformas de vídeos, principalmente do YouTube. Neste segundo trabalho, buscamos realizar um *fine-tuning* a fim de discriminar o conteúdo semântico e textual de vídeos que majoritariamente foram submetidos à processos de moderação em comparação com vídeos ainda disponíveis. Isso permite a detecção do conteúdo que está disponível mas que apresenta proximidade semântica com o conteúdo

Introdução

O crescente aumento de popularidade e uso das redes sociais permitiu uma rápida proliferação de conteúdo de radicalização nesses ambientes (Thompson 2011), impulsionando uma série de estudos sobre os mecanismos por trás da disseminação desse conteúdo em plataformas moderadas, como no YouTube (Ribeiro et al. 2020; Fabbri et al. 2022) e Twitter (Linhares et al. 2022; da Rosa Jr et al. 2022) e nas plataformas não moderadas, como no Telegram (Walther and McCoy 2021; Schulze et al. 2022). A necessidade de compreender a disseminação de conteúdo de radicalização online decorre de suas potenciais consequências para a ordem social, tendo em vista que casos de radicalização política facilitada pelas redes sociais já levaram a consequências tangíveis nas sociedades modernas. Por exemplo, o planejamento do ataque ao Capitólio nos Estados Unidos em 6 de janeiro de 2021 (Times 2021) e a tentativa de golpe de Estado no Brasil em 8 de Janeiro de 2023, que envolveu a invasão do Supremo Tribunal, Palácio Presidencial e Congresso (de S. Paulo 2023), foram ambos alimentados por ambientes online, especialmente o Telegram (Stone 2021; UOL 2023).

Embora estudos anteriores tenham examinado a disseminação de conteúdo de radicalização em diversas redes sociais (Ribeiro et al. 2020; Ai et al. 2021; Goel et al. 2023; Habib, Srinivasan, and Nithyanand 2022), incluindo o Telegram (Walther and McCoy 2021; Cavalini et al. 2023), nenhum se aprofundou no papel do Telegram como um canal para disseminar conteúdo de radicalização de outras redes sociais, especialmente plataformas de vídeo. De fato, pesquisas anteriores já revelaram a ampla compartilhamento de links para plataformas de vídeo externas em grupos do Telegram (Júnior et al. 2021), incluindo grupos de extrema-direita (Bovet and Grindrod 2022). No entanto, nenhum trabalho anterior investigou o conteúdo associado a esses links.

Em particular, a exploração do conteúdo do YouTube é de grande importância devido às recomendações algorítmicas da plataforma, que podem criar câmaras de eco e caminhos de radicalização ao reforçar e ampliar o alcance de conteúdo extremista (Ribeiro et al. 2020). Assim, o YouTube pode acabar atuando como um mecanismo auxiliar na radicalização de usuários que entram em contato com esses vídeos. Esse processo pode ser ainda mais exacerbado quando o conteúdo encontra uma audiência já inflamada e polarizada, como ocorre em grupos de discussão orientados para a política (como os do Telegram) (Cavalini et al. 2023; Urman and Katz 2022).

O principal objetivo aqui é investigar o uso do Telegram como uma plataforma auxiliar para disseminar conteúdo de vídeo de radicalização de plataformas de vídeo moderadas (ou seja, o YouTube) e não moderadas. O foco será na disseminação de conteúdo de radicalização relacionado a um recente evento social e político de grande impacto na sociedade brasileira, ou seja, a tentativa de golpe de Estado de janeiro de 2023. Devido à sua significativa associação com o evento (de S. Paulo 2023), escolhemos grupos políticos brasileiros no Telegram como o foco do estudo.

No primeiro trabalho, encontramos evidências iniciais do uso do Telegram como ferramenta auxiliar para a disseminação de conteúdo de radicalização de plataformas de vídeo, em especial no Youtube. Com aplicação do BERTopic, encontramos tópicos que indicam a presença de vídeos com teor radicalizador no conjunto de vídeos dos indisponíveis. Neste trabalho, iremos explorar a realização de *fine-tuning* em modelos BERT a fim de realizar a separabilidade semântica entre vídeos que potencialmente possuem conteúdo radicalizador em detrimento daqueles que não possuem.

Trabalhos Relacionados

Sobre a Definição de Radicalização

A literatura carece de consenso sobre o significado preciso do termo *radicalização* (Sedgwick 2010). Alguns autores (por exemplo, (Institute 2009)) definem radicalização como um processo pelo qual uma pessoa (ou grupo) adota um sistema de crenças extremista, envolvendo a aceitação ou endosso do uso da violência, fornecendo suporte para atos violentos ou facilitando a violência como um meio para promover mudanças significativas na sociedade. Portanto, esta definição vincula explicitamente a radicalização à violência. Outros (Borum 2011; Veldhuis and Staun 2009) reconhecem a diferença entre o processo de radicalização e o comprometimento real de um indivíduo com um ato violento, referindo-se a "radicalização violenta" ou "extremismo violento" quando a violência é usada para alcançar mudanças sociais pretendidas (Sas et al. 2020).

Adotamos aqui o quadro teórico proposto por Borum (Borum 2011), que define *radicalização* como *o processo pelo qual os indivíduos desenvolvem ideologias e crenças extremistas.* O autor dissociou o termo da noção de violência, argumentando que muitos indivíduos que possuem ideias radicais não necessariamente participam de atividades violentas. Por extensão, definimos *conteúdo de radicalização* como *qualquer forma de conteúdo que defenda, incentive, dissemine ou promova mensagens, visões, ideias ou crenças extremistas, de maneira que não necessariamente encoraje indivíduos a se envolverem em ações violentas ou terroristas.* Por *crenças extremistas*, referimo-nos a crenças que se opõem aos valores fundamentais da sociedade, aos princípios democráticos e aos direitos humanos universais. Essas crenças frequentemente se manifestam por meio de comportamentos que se desviam das normas sociais e demonstram desconsideração pela vida, liberdade e direitos humanos (Trip et al. 2019).

Em essência, assumimos que *conteúdo de radicalização* refere-se a conteúdo que pode contribuir para o processo de radicalização de indivíduos em termos de suas perspectivas políticas, religiosas, sociais ou ideológicas. Diante disso, nosso objetivo aqui é obter uma compreensão abrangente de como o Telegram funciona como uma plataforma complementar para a disseminação de conteúdo de radicalização encontrado em plataformas de vídeo moderadas e não moderadas.

Radicalização em Redes Sociais

Vários estudos examinaram a radicalização em várias plataformas de mídia social. No Telegram, os autores de (Schulze et al. 2022) avaliaram as atividades de grupos como QAnon, Movimento Identitário e Querdenken, observando como suas narrativas frequentemente apontam para dinâmicas de radicalização. Walther *et al.* caracterizaram as atividades de grupos de extrema-direita e extrema-esquerda nos EUA, concentrando-se em seu envolvimento na disseminação de campanhas de desinformação e teorias da conspiração (Walther and McCoy 2021). Da mesma forma, a formação e consolidação de redes de interação entre

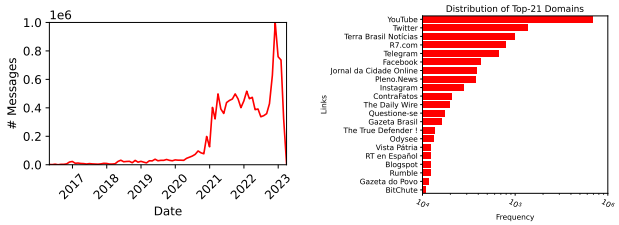

Figure 1: Overview of our Dataset). a) Temporal evolution of the number of messages and b) Top-21 most frequently shared domains.

grupos de extrema-direita nos EUA foi analisada em (Urman and Katz 2022).

A radicalização no YouTube também atraiu a atenção dos pesquisadores. Por exemplo, os autores de (Hosseinmardi et al. 2021) usaram dados de navegação para analisar a tendência dos usuários a se radicalizarem por meio de recomendações e canais tendenciosos. O algoritmo de recomendação do YouTube também foi o foco de outros autores, que demonstraram a existência de um pipeline de radicalização (Ribeiro et al. 2020) e propuseram estratégias para mitigar os caminhos de radicalização (Fabbri et al. 2022). Em outras plataformas, o trabalho de Habib et al. examinou o envolvimento social de usuários do Reddit para entender a disseminação de traços radicais (Habib, Srinivasan, and Nithyanand 2022), enquanto Linhares *et al.* estudaram as diferenças entre contas do Twitter que foram suspensas devido a comportamentos potencialmente radicais em comparação com contas ativas (Linhares et al. 2022).

A radicalização também foi estudada em plataformas periféricas e não moderadas. Por exemplo, Dehghan *et al.* examinaram conversas sobre vacinas no Gab e a radicalização do discurso (Dehghan and Nagappa 2022). Nagappa *et al.* estudaram o processo de radicalização no Gab, focando no fluxo de informações entre o Gab e plataformas similares (Rumble, Bitchute e Gab) (Nagappa and Dehghan 2022), enquanto Goel *et al.* compararam os padrões de radicalização no Gab com plataformas moderadas (Reddit e Twitter) (Goel et al. 2023). Ai et al. coletaram metadados de vídeos de cinco grupos específicos que apoiam ou se opõem a movimentos de esquerda e direita (por exemplo, QAnon, Antifa, Oath Keepers, Black Lives Matter) em quatro plataformas, a saber, YouTube, Bitchute, Vimeo e 4Chan, para analisar características relacionadas à popularidade e persuasão dos vídeos (Ai et al. 2021).

Ao contrário de estudos anteriores, este trabalho planeja examinar de forma única como o Telegram funciona como plataforma de distribuição de conteúdo de radicalização de plataformas de vídeo moderadas e não moderadas. A contribuição consiste na análise em larga escala entre plataformas, comparando especificamente as características de vídeos de uma plataforma amplamente utilizada e moderada, o YouTube, com aqueles originários de plataformas não moderadas, que será realizado num momento.

Caracterização do Dataset

Nossa coleta de dados no Telegram foi iniciada em 25 de novembro de 2022 e encerrada em 3 de março de 2023. No total, identificamos e ingressamos em 264 grupos e 752 canais no Telegram, coletando um total de 13.601.887 mensagens postadas por 230.888 usuários individuais. Os dados abrangem de 4 de junho de 2016 (primeira mensagem) até 3 de março de 2023. No geral, 10.861.263 mensagens contêm apenas conteúdo textual, enquanto 5.404.908 mensagens continham vários tipos de mídia, como fotos, vídeos e documentos PDF. Especificamente, 2.191.181 mensagens contêm referências a domínios externos.

A Figura 1.a) mostra a série temporal do número de mensagens (escala logarítmica no eixo y) compartilhadas nos grupos monitorados. Vale destacar que muitos grupos/canais surgiram durante o período coberto por nossos dados, o que levou a um aumento acentuado no número de mensagens, especialmente a partir de 2021. A Figura 1.b), por sua vez, mostra a distribuição dos 21 domínios mais frequentemente compartilhados (escala logarítmica no eixo x). Revela a presença proeminente do YouTube, que representa mais de $31,65\%$ de todos os links compartilhados no Telegram. Dos 321.665 vídeos únicos do YouTube em nosso conjunto de dados, 210.340 (aproximadamente 65%) ainda estavam disponíveis no momento da coleta de dados. Assim, mais de um terço dos vídeos do YouTube (únicos) com links compartilhados nas mensagens do Telegram foram excluídos pelo usuário que postou o vídeo ou removidos pela plataforma. A fração é semelhante se considerarmos todos os 693.655 links para vídeos (incluindo repetições). Além disso, observamos que um total de 20.610 usuários compartilharam pelo menos um link para um vídeo do YouTube, embora alguns usuários tenham compartilhado até 13.177 links no total. Da mesma forma, descobrimos que 9.903 usuários distintos compartilharam vídeos que não estavam disponíveis no momento de nossa coleta. Um usuário, em particular, compartilhou até 5.453 vídeos distintos que não estavam mais disponíveis. A Figura 1.b) também mostra a presença de um número relevante de links para as três plataformas não moderadas, a saber, Odysee (13.255 links, com 6.600 vídeos únicos), Rumble (12.221 links com 8.372 vídeos únicos) e BitChute (10.927 links e 5.411 vídeos únicos).

Embora essas plataformas possam ter uma frequência menor de compartilhamento, oferecem espaços alternativos para o compartilhamento de conteúdo que pode estar mais alinhado com ideologias extremistas ou narrativas Sua presença na narrativa fornece evidências da diversidade de fontes e do potencial para o conteúdo extremista se espalhar além das plataformas convencionais.

Metodologia

Buscando Evidências de Disseminação de Radicalização de Vídeos do YouTube no Telegram

O grande número e diversidade de vídeos no conjunto de dados, combinados com a noção altamente subjetiva de radicalização, tornam essa tarefa um tipo de desafio de procurar uma ”agulha no palheiro”.

Para enfrentar esse desafio, foi optado por restringir nosso espaço de busca inicial a vídeos com uma chance maior de conter conteúdo de radicalização. Dados os esforços relatados do YouTube para reduzir a radicalização por meio de moderação (Barker and Murphy 2020), especialmente no Brasil (Post 2022), restringimos nossa busca a vídeos que foram considerados indisponíveis no momento em que rastreamos a plataforma. *Não é afirmado que todos esses vídeos contenham conteúdo de radicalização, nem que todos os vídeos com conteúdo de radicalização foram considerados indisponíveis em nossos dados.* Em vez disso, nossa suposição (validada em no experimento) é que, se vídeos com conteúdo de radicalização estiverem presentes nos dados, o conjunto de vídeos indisponíveis pode conter uma alta concentração deles. Portanto, ao examinar esses vídeos, deveríamos ser capazes de encontrar evidências desse conteúdo com mais facilidade.

Observamos que avaliar a presença de radicalização apenas por meio da análise de metadados é difícil, pois avaliações mais precisas exigiriam transcrições de vídeo ou até mesmo avaliação humana. Infelizmente, como este conjunto de vídeos não está disponível, não é possível obter informações adicionais sobre eles que permitiriam uma avaliação mais precisa do assunto (por exemplo, adquirir metadados suplementares para uma análise mais aprofundada ou investigar a causa da indisponibilidade) sem violar os Termos de Uso do YouTube.

No entanto, ainda consideramos alusões explícitas nos metadados a temas associados à radicalização como evidência convincente da disseminação desse conteúdo. Isso é especialmente verdadeiro quando combinado com o sinal de indisponibilidade, que pode estar fortemente relacionado ao processo de moderação do YouTube.

Especificamente, foi utilizado o BERTopic (Grootendorst 2020), um framework que combina um modelo de incorporação e um algoritmo de agrupamento para extrair tópicos do subconjunto de vídeos do YouTube marcados como indisponíveis em nossos dados. Ao gerar representações vetoriais (incorporações), o BERTopic preserva as informações semânticas de um conjunto dado de frases, permitindo que elas sejam agrupadas com base na semelhança. Para obter as incorporações, o BERTopic depende do Sentence BERT (Reimers and Gurevych 2019), que demonstrou bom desempenho em tarefas de semelhança de texto semântico usando um modelo pré-treinado. Em particular, usamos o BERTimbau, um modelo especificamente projetado para texto em português, que mostrou bom desempenho em tarefas semelhantes (Souza, Nogueira, and Lotufo 2020). Em resumo, o BERTopic segue um processo de quatro etapas. Primeiramente, as frases (o conteúdo textual associado a cada vídeo, no nosso caso) são representadas como vetores usando um modelo pré-treinado. Em seguida, o algoritmo UMAP (McInnes et al. 2018) é aplicado para reduzir a dimensionalidade, preservando características globais e locais. Em seguida, o algoritmo HDBSCAN (Campello, Moulavi, and Sander 2013) agrupa as incorporações de baixa dimensionalidade em clusters com representações semânticas semelhantes. Por fim, palavras discriminativas são extraídas de cada cluster usando TF-IDF

baseado em classe (c-TF-IDF), caracterizando os principais tópicos dos vídeos.

A parametrização do BERTopic foi determinada empiricamente para obter clusters mais interpretáveis. Especificamente, para o UMAP, usamos 40 componentes e definimos a distância mínima como 0,2, enquanto o HDBSCAN empregou 30 amostras mínimas e um tamanho mínimo de cluster de 40. Em seguida, reduzimos o número de tópicos para 100, fundindo os clusters mais semelhantes com base em sua representação c-TF-IDF. Esta etapa visava facilitar a análise manual, reduzir outliers e, finalmente, remover stopwords das representações finais.

Foi aplicada inicialmente a abordagem acima à concatenação de títulos e descrições de cada vídeo. No entanto, foi encontrado uma grande presença de termos relacionados a pedidos de doações, inscrições, likes, adesão a grupos e até transferências bancárias nas descrições de vídeos com tópicos muito diferentes. Esses termos afetaram o desempenho do BERTopic, resultando em clusters ruidosos (de baixa qualidade), sem uma separação clara ou tópicos representativos distintos. Assim, foi optado por usar apenas os títulos como representação textual do conteúdo do vídeo, já que sintetizam melhor o assunto do vídeo, o que levou a resultados mais interpretáveis.

Disponibilidade Persistente de Conteúdo de Radicalização no YouTube

Ao aplicar o BERTopic ao conjunto de vídeos do YouTube indisponíveis, descobrimos que a grande maioria desses vídeos está de alguma forma relacionada a tópicos de radicalização, como será exibido na Seção . Portanto, em um próximo passo, usamos o conjunto desses vídeos como *representativos de conteúdo de radicalização*, com o objetivo de revelar mais vídeos semanticamente similares no restante de nosso conjunto de dados.

Especificamente, nos referimos ao conjunto de vídeos do YouTube indisponíveis como *vídeos HPRC*, ou seja, vídeos com *Alto Potencial de ter Conteúdo de Radicalização* e construímos um modelo de classificação baseado em linguagem para distinguir entre vídeos HPRC e não-HPRC. O modelo é primeiro aplicado a todos os vídeos do YouTube disponíveis e indisponíveis, e depois aos vídeos das plataformas de vídeo não moderadas. Fazemos isso com dois objetivos: i) investigar a disponibilidade persistente de conteúdo que mantém forte correlação semântica com HPRC (RQ3) e (ii) comparar a relação semântica entre vídeos do YouTube e de plataformas não moderadas (RQ4). Nossa suposição é que vídeos que não podem ser claramente separados (no espaço de embeddings) de regiões com alta concentração de vídeos HPRC têm uma chance maior de conter conteúdo de radicalização.

Mais uma vez, empregamos um modelo BERT pré-treinado, nomeadamente o BERTimbau (Souza, Nogueira, and Lotufo 2020), para distinguir um conjunto de vídeos de entrada em HPRC e não-HPRC. Para garantir uma melhor representação e separabilidade, propomos ajustar finamente o BERTimbau para essa tarefa específica. A afinação fina é um processo de treinamento iterativo que envolve treinar um modelo BERT pré-treinado em um conjunto de dados rotulado. Durante este processo, o modelo incorpora suas representações de linguagem existentes com conhecimento específico da tarefa ajustando seus parâmetros. Este procedimento melhora o desempenho do modelo na tarefa alvo aproveitando tanto seu conhecimento linguístico geral quanto adaptações específicas da tarefa. Em nosso caso específico, o objetivo é melhorar a separabilidade de classes dos metadados do vídeo e criar uma rede capaz de prever se os metadados de um vídeo são mais semelhantes a conteúdo HPRC ou não-HPRC. Começamos dividindo o conjunto de dados contendo todos os metadados de vídeos HPRC e não-HPRC em duas partes, treinando o modelo em 90% dos dados e testando-o nos 10% restantes usando amostragem estratificada aleatória. Seguimos a metodologia de Sun *et al.* (Sun et al. 2019) para determinar os hiperparâmetros. Ou seja, usamos uma taxa de aprendizado de $2e^{-5}$, uma probabilidade de *dropout* de 0,1 e um tamanho de lote de 16, com *gradient accumulation*. Cada conteúdo de vídeo foi representado pela concatenação de seu título e descrição, que foram então tokenizados em 512 tokens. Para determinar o melhor número de épocas para ajustar finamente a rede BERT, realizamos validação cruzada de 3 *folds* para avaliar a capacidade de generalização do modelo treinado por até 4 épocas. Em seguida, selecionamos os hiperparâmetros do modelo que alcançaram a maior média de pontuação F1 macro durante a etapa de validação cruzada. Este modelo, treinado ao longo de 3 épocas, foi usado para treinar o modelo final usando todo o conjunto de treinamento. Nos dados de teste separados que reservamos para análise, o modelo alcançou uma pontuação F1 macro de 84% e uma precisão de 85%. Treinamos o modelo localmente usando um *RTX 4070* como GPU e um *Ryzen 7 5800x3d* como CPU.

Resultados

Evidências da Disseminação de Vídeos de Radicalização do YouTube para o Telegram

Para entender se e como o Telegram poderia ser utilizado para distribuir conteúdo de radicalização de plataformas de vídeo, especialmente o YouTube, primeiro precisamos determinar se grupos do Telegram compartilham conteúdo de vídeo do YouTube que poderia potencialmente conter material de radicalização. Para fazer isso, realizamos uma análise para examinar como o Telegram é usado para disseminar conteúdo de radicalização de plataformas como o YouTube. Ao aplicar nossa metodologia (descrita na Seção), atribuímos 99.76% dos 111.325 vídeos únicos indisponíveis do YouTube a clusters específicos com base em seus títulos. A Tabela destaca as palavras mais representativas para 9 dos 100 tópicos identificados pelo BERTopic, juntamente com um exemplo traduzido do título do vídeo para fornecer contexto.

A análise revelou descobertas significativas nesses clusters. O Cluster ID 1, o maior, foca predominantemente em conteúdo político relacionado à situação política no Brasil, abordando tópicos como Bolsonaro, CPI (Comissão Parlamentar de Inquérito) e o Supremo Tribunal Federal. Outro cluster significativo (ID 2) inclui vídeos centrados em duas figuras proeminentes na política brasileira: Lula, o atual

Cluster ID	# Vídeos Únicos	Palavras Mais Representativas	Exemplo de Documento Representativo
1	18,375	Bolsonaro, CPI, Ministro, Supremo Tribunal, Agora	"Forças Armadas no governo Bolsonaro - agora vai."
2	13,447	Lula, Olavo de Carvalho, Tudo, Urgente, Mundo	"Urgente! Olavo de Carvalho - Espiões da China, censura, e tudo o que está acontecendo!"
7	2,275	Covid-19, Vacina, Vírus, Vacinação, Hospital	"A ilusão em relação à vacina contra a Covid-19"
9	1,859	Protesto, 12/22, Protestos, Brasília, HQ (Quartel General)	"Protesto em Brasília (11/12/22)"
13	1,490	Militar, Guerra, Comando, Intervenção, Intervenção Militar	"Protesto em frente ao Comando Militar do Sudeste em São Paulo"
24	1,082	C4RTE $UPR4MA, B0LS0NAR0, B0MBA, AG0RA, M0RAES	"B0l$0nar0 toma ação decisiva! alex4ndre de m0rae$ não vai silenciar o Brasil!"
28	976	Voto, Papel, Cédula de Papel, Auditável, Retorno	"O Brasil quer, e o Brasil terá. Votação impressa e auditável agora!"
31	871	Terra, Plana, Terra Plana, NASA, Lua	"Olhe para esta fazenda na Antártida, prova que a Terra é plana!"
58	439	Gay, Homens, Feminismo, Feminista, Mulheres	"A doença do feminismo, o dia em que o feminismo foi exposto."

Table 1: Tabela de Informações sobre Tópicos

presidente, e Olavo de Carvalho, uma figura proeminente que advoga pela disseminação do movimento de extrema direita no Brasil e frequentemente associada a teorias conspiratórias em um contexto político. Em paralelo, o Cluster 24 sugere que seu conteúdo está relacionado a ataques ao Ministro do Supremo Tribunal, Alexandre de Moraes, que tem obstruído vários esforços antidemocráticos de alguns movimentos de extrema direita no Brasil (T. 2023). Também é perceptível que várias tentativas de soletrar palavras são empregadas como estratégia de disseminação. Por exemplo, 'SUPREME' é escrito como 'SUPR4ME', 'NOW' é escrito como 'N0W'. Isso sugere que esses vídeos podem estar circulando em plataformas moderadas, já que são feitas tentativas de deturpar palavras-chave para que possam ser encontradas e removidas. O documento representativo deste cluster apresenta uma narrativa que retrata o Ministro Moraes (escrito como 'M0RAES') como uma figura que busca silenciar e censurar o povo brasileiro, aumentando ainda mais a hostilidade em relação a ele e motivando os apoiadores de Bolsonaro a agir (Reuters 2022b). Isso contribuiu para os eventos crescentes que culminaram em 8 de janeiro de 2023.

Os Clusters 9 e 13 contêm conteúdo que parece estar ainda mais diretamente relacionado à radicalização, especialmente a disseminação de material sobre a ocupação da sede militar por simpatizantes que defendem a intervenção militar para invalidar as eleições de 2022 que Bolsonaro perdeu, levando a várias manifestações (Guardian 2022; Reuters 2022a). Ao mesmo tempo, o Cluster 28 parece estar fortemente associado à proliferação de conteúdo atacando o processo eleitoral brasileiro, visando especificamente a credibilidade e confiabilidade das máquinas de votação eletrônica (EVMs) e advogando pelo uso de cédulas de papel como sistema de substituição. Essa narrativa, frequentemente alimentada pelo Sr. Bolsonaro (Times 2022), alega que o sistema EVM é vulnerável a fraudes e manipulações e seria usado para fraudar as eleições contra ele, mesmo que várias instituições e universidades já tenham certificado a segurança das EVMs (Brasil 2022; BBC 2022).

Além disso, existem vários clusters que apontam para assuntos não políticos, mas ainda mostram uma possível conexão com a radicalização. Por exemplo, o cluster 7 aborda tópicos relacionados à COVID-19 e vacinas, possivelmente envolvendo a disseminação de conteúdo que promove ceticismo ou oposição à vacinação. Além disso, o cluster 31 mostra uma conexão com a disseminação da teoria da Terra Plana, que muitas vezes está interligada com várias teorias conspiratórias. A disseminação dessas teorias pode contribuir para a radicalização ao fomentar um senso geral de ceticismo, inspirar desconfiança em instituições estabelecidas e apoiar crenças em ideologias que se desviam das normas sociais. Finalmente, o cluster 58 aborda predominantemente tópicos relacionados a gênero e sexualidade. O documento representativo deste cluster ilustra uma visão negativa e depreciativa do movimento feminista. Esse conteúdo, que reforça estereótipos negativos e potencialmente justifica ou incita violência contra mulheres ou o movimento feminista, pode estar associado ao processo de radicalização. Por fim, por meio da inspeção manual dos clusters e da aplicação de uma abordagem altamente conservadora que considera apenas clusters com evidências explícitas de conteúdo de radicalização, identificamos que, dos 100 clusters analisados, um total de 60 clusters englobam aproximadamente 87.428 vídeos, cerca de 79%, que estão potencialmente associados à radicalização.

Discussão: A análise fornece insights sobre a distribuição de vídeos do YouTube dentro de grupos políticos brasileiros no Telegram, abordando tanto *RQ1* quanto *RQ2*. A proporção significativa de vídeos do YouTube indisponíveis compartilhados no Telegram sugere que o Telegram serve como uma plataforma para a disseminação de conteúdo que pode

ter passado por moderação no YouTube. Essa descoberta destaca o papel do Telegram em facilitar a propagação de conteúdo de radicalização originado do YouTube. Ao examinar os clusters de vídeo, revelou-se uma ampla variedade de categorizações, abrangendo temas políticos e não políticos. Os clusters políticos estão alinhados com tópicos específicos relevantes para o cenário político brasileiro, enquanto os tópicos não políticos denotam diversos assuntos como teorias da conspiração, movimento anti-vacina, questões de gênero, dentre outros. Um próximo passo seria validar, de maneira manual, em cada um destes clusters, os metadados dos vídeos, de modo que seja possível treinar uma classificador utilizando técnicas de NLP (como BERT fine-tune) para que seja possível treinar um classificador de radicalização.

Disponibilidade Persistente de Vídeos no YouTube com Radicalização

Assim, o próximo passo na investigação foi avaliar a disponibilidade de conteúdo potencialmente radicalizante no YouTube compartilhado por grupos do Telegram. Realizamos um *fine-tuning* num modelo BERT (conforme descrito na Seção) para melhorar a separabilidade de metadados entre as classes de vídeo disponíveis e não disponíveis, nos permitindo prever a disponibilidade de vídeos com base em seus metadados. Ao examinar regiões com alta densidade de vídeos não disponíveis, podemos identificar vídeos disponíveis no YouTube que podem conter conteúdo radicalizante.

As Figuras 2 e 3 exibem os embeddings de baixa dimensionalidade de vídeos gerados antes e depois do ajuste fino da rede BERT. Utilizamos a abordagem de (Andrade et al. 2023) para obter os embeddings, utilizando o token CLS e reduzindo a dimensionalidade com T-SNE (Maaten and Hinton 2008). Ambas as figuras distinguem entre vídeos disponíveis e não disponíveis. A Figura 2 ilustra que, antes do ajuste fino, não há uma separação clara entre as classes de vídeo, destacando a necessidade do ajuste fino para melhorar a separabilidade das classes. Na Figura 3, os embeddings finais após o ajuste fino demonstram uma separabilidade melhorada, permitindo a identificação de regiões com maiores densidades de vídeos nHPRC e HPRC.

No entanto, vale ressaltar que um número significativo de vídeos nHPRC é encontrado em regiões com uma concentração maior de vídeos HPRC, e vice-versa. Isso é esperado, já que alguns vídeos podem estar definidos como privados e não ter conexão com vídeos removidos por moderação, enquanto outros podem ser acessíveis, mas estar em violação das políticas da plataforma. Se o modelo aprendeu com sucesso padrões semânticos significativos para diferenciar entre HPRC e não HPRC, é esperado que os vídeos disponíveis exibam representações dentro da faixa de vídeos não disponíveis, e vice-versa. Isso está alinhado com nossas expectativas e motiva nossa pesquisa a focar em vídeos disponíveis dentro de regiões de conteúdo não disponível, pois esses são particularmente relevantes para nosso estudo.

A Figura 4 exibe os embeddings após a classificação com a rede BERT, representando a classificação final dos vídeos. Notavelmente, a rede tende a categorizar vídeos disponíveis localizados em regiões com alta densidade de vídeos indisponíveis como indisponíveis e vice-versa, permitindo-nos analisar ambas as classes de vídeos.

Para investigar ainda mais, examinamos aleatoriamente uma amostra de vídeos indisponíveis em regiões com alta densidade de vídeos disponíveis. Entre os 20 vídeos analisados, 15 estavam definidos como privados e cobriam uma variedade de tópicos não relacionados à radicalização, como notícias, eventos religiosos e cursos online. Três vídeos estavam relacionados à música e podem ter sido removidos devido a violação de direitos autorais. Dois vídeos estavam relacionados à política e poderiam potencialmente conter conteúdo de radicalização. Hipotetizamos que a maioria dos vídeos indisponíveis nessas regiões provavelmente não estava associada à radicalização, embora alguns poucos pudessem estar relacionados a ela.

Por outro lado, também houve uma frequência notável de regiões com um número significativo de vídeos disponíveis dentro de áreas caracterizadas por uma alta densidade de vídeos indisponíveis. Essa observação sugere que esses vídeos disponíveis podem compartilhar semelhanças na estrutura de metadados com os vídeos indisponíveis e podem potencialmente estar associados à radicalização. Também examinamos um conjunto de 20 vídeos disponíveis aleatoriamente localizados em regiões indisponíveis. Esses vídeos disponíveis exibiam uma estrutura de metadados semelhante aos tópicos descritos na Tabela , predominantemente relacionados à política brasileira, vacinas e teorias da conspiração. Exemplos desses vídeos podem ser encontrados na Tabela 2.

No conjunto de testes de 32.167 vídeos, descobrimos que 1.993 vídeos indisponíveis estavam dentro do espaço de vídeos disponíveis, enquanto 2.681 vídeos disponíveis foram classificados como indisponíveis, indicando a possibilidade de conterem conteúdo de radicalização. Consequentemente, nossos resultados mostram que aproximadamente 13% dos vídeos disponíveis únicos em nosso conjunto de testes podem conter conteúdo de radicalização e permanecer acessíveis para visualização. Esses vídeos cobrem principalmente política, indicado por tags comuns como 'Bolsonaro', 'Brasil', 'Lula' e 'Política', com cerca de 70% postados entre 2021 e 2022. Isso destaca a presença contínua de conteúdo radicalizado no YouTube, apesar dos esforços da plataforma em moderação.

Discussão: A análise destaca uma preocupação contínua apesar da implementação de moderação humana pelo YouTube em 2020 para prevenir a remoção de conteúdo sensível (Barker and Murphy 2020). Foram identificados uma prevalência de vídeos (12,74%) que estão disponíveis na plataforma e fortemente correlacionados com conteúdo moderado indisponível do YouTube, muitos dos quais podem conter material de radicalização. Esses vídeos podem potencialmente contribuir para a disseminação da radicalização, tanto dentro do YouTube quanto em plataformas como o Telegram, onde são compartilhados. Os vídeos identificados predominantemente se concentram em tópicos políticos, evidente pelas suas tags do YouTube, destacando sua potencial influência no processo de radicalização, po-

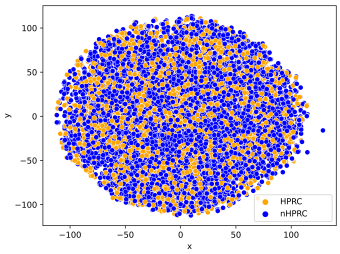

Figure 2: Incorporação de Vídeo Antes do Ajuste Fino

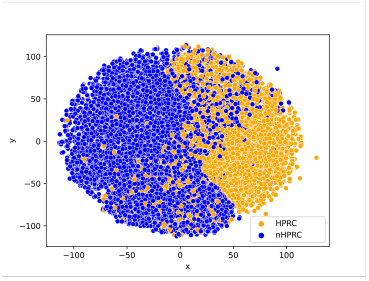

Figure 3: Incorporação de Vídeo Depois do Ajuste Fino

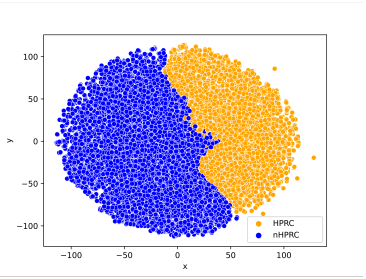

Figure 4: Classificação do Modelo Depois do Ajuste Fino

Figure 5: Representação de Incorporação usando T-SNE de Vídeos do YouTube

Table 2: Exemplos de Vídeos Disponíveis no YouTube que possuem similaridade semântica próxima com vídeos indisponíveis

Title	Description
”INTERVENÇÃO CIRÚRGICA NACIONAL! CORAGEM, FORÇA, SABEDORIA E FÉ!”	”Não há saída fácil para alguém que tenha entregado sua liberdade a chacais em busca de segurança, e não recebeu nada em troca!”
‘A PEDOFILIA E A EROTIZAÇÃO DE CRIANÇAS - Entenda os motivos com Olavo de Carvalho.).	Esquerdistas promovem a pedofilia e a sexualização de crianças. Vejam, por exemplo, a exibição ”queermuseum” promovida pelo banco Santander, que possui diversas esculturas de crianças quando seria natural não ter interesse sexual. Através desse vídeo, com demonstração do Olavo de Carvalho, entenda o porquê dos esquerdistas atacarem crianças.
”Prisão de Ministros do Supremo Tribunal Federal. A Corte Internacional de Justiça Quer Ouvir de Você.”	”Você podeajudar a PRENDER TODOS OS MINISTROS DO SUPREMO TRIBUNAL. Nesta LIVE, eu TE ensino como AJUDAR A ANSP BRASIL a prender todos os ministros do Supremo Tribunal, governadores prefeitos e outros agentes, autoridades e servidores públicos pegos em flagrante cometendo CRIMES GRAVES: Genocídio e Humanicídio.
”Guardiões da Lei e da Ordem. As Forças Armadas, Ministro da Defesa e o Presidente da República.”	”ENTRE EM UM DESTES CANAIS E DIVULGUE AS SEGUINTES HASHTAGS#BrasilCensurado #SOSExércitoBrasileiro #LulaLadrão @Handles Omitidos”

tencialmente introduzindo e aprofundando a influência dos caminhos de radicalização do YouTube nos usuários do Telegram que eventualmente consomem esses vídeos.

Conclusão

Aplicando várias técnicas de Processamento de Linguagem Natural (NLP) de última geração aos metadados de plataformas de vídeo moderadas e não moderadas, encontramos evidências convincentes de que o Telegram está sendo usado como meio para disseminar conteúdo radicalizador do YouTube. Encontramos vídeos que parecem discutir vários tópicos relacionados à radicalização, que vão desde políticos (por exemplo, intervenção militar, cédulas de papel, ocupação de sedes) até não políticos (por exemplo, vacinas, Covid-19, teorias da conspiração, gênero e sexualidade). Também mostramos que há um subconjunto de vídeos disponíveis no YouTube que não foram moderados pela plataforma, apesar de conterem evidências de conteúdo radicalizador. Por fim, mostramos que várias plataformas não moderadas, como Rumble, BitChute e Odysee, hospedam uma quantidade significativa de conteúdo semanticamente relacionado a vídeos do YouTube indisponíveis, muitos dos quais podem conter conteúdo radicalizador. Nossos resultados sugerem que o Telegram pode ser usado para disseminar conteúdo radicalizador de plataformas de vídeo moderadas e não moderadas. Os resultados destacam a importância do monitoramento contínuo e de mecanismos de moderação aprimorados para detectar e remover conteúdo radicalizador em redes sociais, a fim de evitar que plataformas-chave atuem como faróis para a disseminação de conteúdo radicalizador por outros, especialmente plataformas de vídeo, como demonstramos.

Como trabalho futuro, contemplamos investigar mais a fundo o conteúdo de vídeos de plataformas não moderadas, com uma análise e caracterização minuciosas. Também consideramos utilizar os resultados encontrados para aprofundar a análise da ocorrência de radicalização dentro do Telegram. Embora esteja além do escopo deste trabalho, consideramos investigar como outros domínios, não necessariamente relacionados a plataformas de vídeo, podem contribuir para a disseminação de conteúdo de radicalização no Telegram. A investigação visaria lançar luz sobre os padrões de difusão de conteúdo de radicalização interdomínio, para auxiliar governos e instituições a implementar medidas para mitigar os efeitos que a radicalização possa ter para a ordem social.

References

Ai, L.; Kathuria, A.; Panda, S.; Sahai, A.; Yu, Y.; Levitan, S.; and Hirschberg, J. 2021. Identifying the Popularity and Persuasiveness of Right-and Left-Leaning Group Videos on Social Media. In *Proc. of Big Data*.

Andrade, C.; Belém, F.; Cunha, W.; França, C.; Viegas, F.; Rocha, L.; and Gonçalves, M. 2023. On the class separability of contextual embeddings representations – or “The classifier does not matter when the (text) representation is so good!”. *IP&M*, 103336.

Barker, A.; and Murphy, H. 2020. YouTube reverts to human moderators in fight against misinformation. *Financial Times. Retrieved August*.

BBC. 2022. Brazil election: Do voting machines lead to fraud? Https://www.bbc.com/news/63061930.

Borum, R. 2011. Radicalization into Violent Extremism I: A Review of Social Science Theories. *Journal of Strategic Security*, 4.

Bovet, A.; and Grindrod, P. 2022. Organization and evolution of the UK far-right network on Telegram. *ANS*, 7(1): 1–27.

Brasil, A. 2022. Brazil universities attest to safety of new voting machines. Https://agenciabrasil.ebc.com.br/en/politica/noticia/2022-08/brazil-universities-attest-safety-new-voting-machines.

Campello, R.; Moulavi, D.; and Sander, J. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, 160–172.

Cavalini, A.; Malini, F.; Gouveia, F.; and Comarela, G. 2023. Politics and disinformation: Analyzing the use of Telegram's information disorder network in Brazil for political mobilization. *First Monday*.

da Rosa Jr, J. M.; Linhares, R. S.; Ferreira, C. H. G.; Nobre, G. P.; Murai, F.; and Almeida, J. M. 2022. Uncovering discussion groups on claims of election fraud from twitter. In *Socinfo*.

de S. Paulo, F. 2023. Coup-Mongering Bolsonarists Invade Planalto, the Supreme Court and Congress in Brasilia. Https://www1.folha.uol.com.br/internacional/en/brazil/2023/01/coup-mongering-bolsonarists-invade-planalto-the-supreme-court-and-congress-in-brasilia.shtml.

Dehghan, E.; and Nagappa, A. 2022. Politicization and radicalization of discourses in the alt-tech ecosystem: A case study on Gab Social. *Social Media and Society*, 8(3).

Fabbri, F.; Wang, Y.; Bonchi, F.; Castillo, C.; and Mathioudakis, M. 2022. Rewiring What-to-Watch-Next Recommendations to Reduce Radicalization Pathways. In *Proceedings of the WebConf*.

Goel, V.; Sahnan, D.; Dutta, S.; Bandhakavi, A.; and Chakraborty, T. 2023. Hatemongers ride on echo chambers to escalate hate speech diffusion. *PNAS nexus*, 2(3).

Grootendorst, M. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.

Guardian, T. 2022. 'We will not surrender': Bolsonaro militants demand coup as Lula prepares to take power. Https://www.theguardian.com/world/2022/dec/22/brazil-bolsonaro-militants-lefist-lula-president.

Habib, H.; Srinivasan, P.; and Nithyanand, R. 2022. Making a Radical Misogynist: How online social engagement with the Manosphere influences traits of radicalization. In *Proceedings of CSCW*.

Hosseinmardi, H.; Ghasemian, A.; Clauset, A.; Mobius, M.; Rothschild, D.; and Watts, D. 2021. Examining the consumption of radical content on YouTube. *PNAS*, 118(32).

Institute, H. S. 2009. Recruitment and radicalization of school-aged youth by international terrorist groups.

Júnior, M.; Melo, P.; da Silva, A.; Benevenuto, F.; and Almeida, J. 2021. Towards Understanding the Use of Telegram by Political Groups in Brazil. In *Proceedings of WebMedia*.

Linhares, R.; Rosa, J.; Ferreira, C.; Murai, F.; Nobre, G.; and Almeida, J. 2022. Uncovering coordinated communities on twitter during the 2020 us election. In *Proceedings of ASONAM*.

Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Nagappa, A.; and Dehghan, E. 2022. Alt-platformization: Radicalization of discourses and information flows in the alternative ecosystem. In *AoIR-Association of Internet Researchers*.

Post, T. N. Y. T. 2022. To Fight Lies, Brazil Gives One Man Power Over Online Speech. Https://www.nytimes.com/2022/10/21/world/americas/brazil-online-content-misinformation.html.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP*.

Reuters. 2022a. Bolsonaro backers call on Brazil military to intervene after Lula victory. Https://www.reuters.com/world/americas/bolsonaro-backers-call-brazil-military-intervene-after-lula-victory-2022-11-02.

Reuters. 2022b. 'Coup-mongering' Bolsonarista's battle cry reveals a radicalized Brazil. Https://www.reuters.com/world/americas/coup-mongering-bolsonaristas-battle-cry-reveals-radicalized-brazil-2022-12-28/.

Ribeiro, M.; Ottoni, R.; West, R.; Almeida, V.; and Meira, W. 2020. Auditing Radicalization Pathways on YouTube. In *Proceedings of FAT*'20*.

Sas, M.; Ponnet, K.; Reniers, G.; and Hardyns, W. 2020. The Role of Education in the Prevention of Radicalization and Violent Extremism in Developing Countries. *Sustainability*, 12(6): 2320.

Schulze, H.; Hohner, J.; Greipl, M., Simon and; Desta, I.; and Rieger. 2022. Far-right conspiracy groups on fringe platforms: a longitudinal analysis of radicalization dynamics on Telegram. *Convergence*, 28(4).

Sedgwick, M. 2010. The Concept of Radicalization as a Source of Confusion. *Terrorism and Political Violence*, 22(4): 479–494.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceeddings of BRACIS*.

Stone, R. 2021. 'This Is War': Inside the Secret Chat Where Far-Right Extremists Devised Their Post-Capitol Plans. Https://www.rollingstone.com/culture/culture-news/capitol-riot-far-right-extremists-telegram-1120511/.

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2019. How to Fine-Tune BERT for Text Classification? In *Chinese Computational Linguistics*.

T., T. N. Y. 2023. He Is Brazil's Defender of Democracy. Is He Actually Good for Democracy? Https://www.nytimes.com/2023/01/22/world/americas/brazil-alexandre-de-moraes.html.

Thompson, R. 2011. Radicalization and the Use of Social Media. *Journal of Strategic Security*, 4(4): 167–190.

Times, T. N. Y. 2021. The storming of Capitol Hill was organized on social media. Https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html.

Times, T. N. Y. 2022. How Bolsonaro Built the Myth of Stolen Elections in Brazil. Https://www.nytimes.com/interactive/2022/10/25/world/americas/brazil-bolsonaro-misinformation.html.

Trip, S.; Bora, M.; Halmajan, A.; and Drugas, M. 2019. Psychological Mechanisms Involved in Radicalization and Extremism. A Rational Emotive Behavioral Conceptualization. *Frontiers in Psychology*.

UOL. 2023. From Celebration to witch hunt: what we saw on Bolsonarist groups on Telegram in the week of the coup d'état attempt. Https://lupa.uol.com.br/jornalismo/2023/01/13/from-celebration-to-witch-hunt-what-we-saw-on-bolsonarist-groups-on-telegram-in-the-week-of-the-coup-d-etat-attempt.

Urman, A.; and Katz, S. 2022. What they do in the shadows: examining the far-right networks on Telegram. *Information, communication & society*, 25(7): 904–923.

Veldhuis, T.; and Staun, J. 2009. *Islamist radicalisation: A root cause model*. Netherlands Institute of International Relations.

Walther, S.; and McCoy, A. 2021. US extremism on Telegram. *Perspectives on Terrorism*, 15(2): 100–124.