UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Gabriel Martins Juarez

Análise de mensagens com conteúdo potencialmente extremista compartilhadas em grupos do Telegram

Belo Horizonte, Minas Gerais
2024

Análise de mensagens com conteúdo potencialmente extremista compartilhadas em grupos do Telegram

Proposta de pesquisa científica apresentada como requisito para a conclusão da Graduação em Sistemas de Informação da Universidade Federal de Minas Gerais

Orientador(a): Jussara Marques de Almeida (UFMG)

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Belo Horizonte, Minas Gerais
2024

Sumário

1. Resumo

Este artigo apresenta uma investigação sobre a disseminação de conteúdo potencialmente radicalizado e extremista em grupos do Telegram, focando nas mensagens textuais associadas ao compartilhamento de links para vídeos de plataformas moderadas e não-moderadas, em especial o YouTube. A análise abrange observações linguísticas, de sentimentos e das reações dos participantes, utilizando algoritmos de Processamento de Linguagem Natural (PLN). A pesquisa visa contribuir para a compreensão das dinâmicas no compartilhamento de conteúdo potencialmente extremista no Telegram.

2. Introdução

No cenário atual da comunicação online, o Telegram se destaca como um ambiente propício para a disseminação de conteúdo radicalizado e extremista. Essa plataforma de mensagens é conhecida por sua flexibilidade e segurança, fatores que contribuem para sua popularidade nesse contexto. Através da criação de grupos e canais, os indivíduos podem compartilhar informações de forma rápida e, muitas vezes, anônima, aproveitando a criptografia ponta a ponta oferecida pelo Telegram para uma sensação adicional de segurança. Além disso, o compartilhamento de links para vídeos externos em grupos do Telegram destaca a importância da análise desse conteúdo na compreensão das dinâmicas de disseminação de ideias radicais e extremistas.

Esta Monografia em Sistemas de Informação propõe uma análise de dados sobre mensagens compartilhadas em grupos do Telegram, predominantemente orientados à política nacional, com foco nas mensagens com **conteúdo textual** associadas ao compartilhamento de links externos, predominantemente links para o YouTube. Foi observado que uma parcela considerável desses links conduz a vídeos indisponíveis na plataforma, levantando a hipótese de que muitos foram removidos devido a conteúdo extremista. Esta hipótese será investigada em detalhes por outro aluno da orientadora, que tem por objetivo analisar especialmente o conteúdo destes vídeos.

O objetivo do presente estudo é contribuir para a investigação desta hipótese, mas com um foco ortogonal e complementar: objetiva-se analisar propriedades do conteúdo textual compartilhado em associação com estes links, ou seja, as propriedades das mensagens de texto compartilhadas anteriormente e posteriormente ao compartilhamento dos links (durante um período de tempo estipulado). A partir da análise destas propriedades, usando técnicas de processamento de linguagem natural (análise de sentimento, análise de toxicidade, extração de tópicos, etc..), busca-se inferir as reações e percepção dos membros desses grupos em relação aos links compartilhados. Espera-se com esta análise coletar evidências complementares sobre a possível presença de conteúdo extremista associado aos vídeos.

3. Objetivo

3.1. Objetivo Geral

O objetivo geral desta MSI é analisar as propriedades de mensagens textuais em grupos do Telegram, compartilhadas antes e após ao compartilhamento de links para vídeos no YouTube, utilizando algoritmos de Processamento de Linguagem Natural (PLN, ou NLP em inglês), para investigar como os membros desses grupos reagem e percebem conteúdo, especialmente se este for de natureza radicalizada ou extremista.

3.2. Objetivos Específicos

- Compreensão das Reações e Percepções dos Participantes: O estudo envolve a realização de análises linguísticas detalhadas, visando inferir como os participantes dos grupos reagem e percebem o conteúdo externo (vídeos) compartilhado, e a partir destas inferências, buscar evidência da presença de conteúdo radicalizado ou extremista.
- Aplicação de Técnicas de PLN: Ao longo de todo o estudo (incluindo a MSI2), será implementado um conjunto abrangente de técnicas e modelos avançados de processamento de linguagem natural. Essas abordagens incluirão análise de sentimento, fazendo uso de algoritmos como SVM, Naive Bayes, SentiStrength, entre outros. Além disso, serão aplicadas técnicas de análise de toxicidade, utilizando o Google Perspective, e extração de tópicos, por meio do BERTopic e do LIWC. O objetivo é realizar uma análise aprofundada das mensagens, possibilitando a identificação dos sentimentos e opiniões expressas.

4. Referencial Teórico

No cenário atual da comunicação online, as plataformas de mensagens desempenham um papel crucial e cada vez mais influente na disseminação de informações, muitas vezes desafiando as fronteiras tradicionais de comunicação e possibilitando que conteúdos de diversas naturezas alcancem um público amplo e diversificado. Como afirma Castells (2010), "a comunicação é o tecido da nossa vida social", e essa dinâmica se torna ainda mais evidente no ambiente digital. Entre essas plataformas de mensagens, o Telegram merece destaque especial devido à sua popularidade crescente e à natureza altamente flexível e criptografada de suas comunicações.

O Telegram, uma plataforma de mensagens que foi projetada para fornecer privacidade e segurança a seus usuários, tornou-se um espaço onde a disseminação de

conteúdo radicalizado e extremista encontra um ambiente propício para prosperar. Isso se deve em parte à sua capacidade de permitir a criação de grupos e canais, onde os participantes podem compartilhar informações, vídeos e outros tipos de mídia de forma rápida e, em muitos casos, anônima. Como apontado por Berger (2019), "as redes sociais têm o potencial de amplificar mensagens extremistas e permitir que elas se espalhem rapidamente".

Nesses grupos e canais, indivíduos com visões extremistas podem encontrar um espaço para trocar ideias, recrutar seguidores e disseminar sua ideologia sem a mesma exposição que enfrentam em plataformas de mídia social mais tradicionais. A criptografia ponta a ponta oferecida pelo Telegram também aumenta a sensação de segurança e anonimato para os participantes, tornando mais difícil o rastreamento e monitoramento de atividades extremistas. Como destacado por Winter (2020), "a privacidade proporcionada pelo Telegram é uma das razões pelas quais grupos extremistas escolhem essa plataforma para coordenar suas atividades".

O fenômeno da radicalização na sociedade contemporânea é complexo e multifacetado, como destacado por autores como McCauley e Moskalenko (2011). Esse processo envolve o desenvolvimento de crenças extremistas e a disposição de indivíduos a tomar medidas drásticas em busca de objetivos ideológicos. É fundamental compreender que a radicalização não ocorre em um vácuo, mas é influenciada por uma interação complexa de fatores individuais, sociais e políticos (Silke, 2008). O extremismo político, conforme discutido por Neumann (2013), desempenha um papel crucial nesse contexto, fornecendo uma plataforma para a disseminação de ideologias radicais e a mobilização de apoiadores. Além disso, as redes sociais emergiram como um canal fundamental para a propagação de mensagens extremistas (Berger, 2019). Elas proporcionam um alcance sem precedentes e uma capacidade de conectar indivíduos com visões similares, criando assim um ambiente propício para a radicalização e a formação de grupos extremistas online.

O Telegram é particularmente relevante no contexto da disseminação de conteúdo radicalizado em vídeo. Vídeos têm a capacidade de transmitir mensagens poderosas e persuasivas, e a plataforma permite o compartilhamento de links para vídeos hospedados em plataformas externas, ampliando ainda mais o alcance desses conteúdos. Isso torna essencial a análise de mensagens em grupos do Telegram que compartilham esses links de vídeos externos, a fim de compreender como o conteúdo radicalizado é disseminado e recebido por seus membros.

Contudo, a Inteligência Artificial (IA) é uma disciplina da ciência da computação que se concentra na criação de sistemas capazes de realizar tarefas que normalmente exigem inteligência humana (Russell & Norvig, 2016). Dentro da IA, o Aprendizado de Máquina (AM) é um subcampo que desenvolve algoritmos para permitir que as máquinas aprendam com dados e melhorem suas ações ao longo do tempo (Murphy, 2012). O Processamento de Linguagem Natural (PLN) é uma aplicação importante da IA e do AM, envolvendo a capacidade de compreender e gerar linguagem humana (Jurafsky & Martin, 2020).

No contexto das redes sociais, o PLN é essencial. Com o crescimento exponencial de conteúdo textual e a diversidade de idiomas e estilos de comunicação

nas plataformas, o PLN se torna uma ferramenta fundamental para a análise e compreensão desse vasto volume de informações. O BERT, introduzido por Devlin et al. (2018), é um modelo que introduziu uma abordagem bidirecional para o processamento de linguagem, considerando o contexto das palavras em ambas as direções, impulsionando avanços em tarefas como compreensão de texto e tradução automática. O BERTimbau, uma versão adaptada para o português, destaca-se por sua eficácia na interpretação precisa da língua portuguesa, servindo como base para aplicações como o BERTopic, que categoriza tópicos em grandes conjuntos de texto, tornando-se valioso na pesquisa e mineração de dados. Por sua vez, o Perspective, uma iniciativa do Google, foca na detecção de toxicidade e tom na linguagem, sendo crucial para avaliar e classificar o conteúdo em relação a sua potencialidade de ser prejudicial ou extremista. Além do BERT e do Perspective, outros algoritmos de PLN podem desempenhar papéis significativos nessa análise. Como afirmado por Mikolov et al. (2013), o Word2Vec é amplamente utilizado para representação de palavras, enquanto o GPT (Generative Pre-trained Transformer), conforme descrito por Radford et al. (2018), é conhecido por sua capacidade de gerar texto humano-como e tem sido aplicado em tarefas de geração de conteúdo nas redes sociais. Também, o FastText, como destacado por Joulin et al. (2016), é conhecido por sua eficiência na classificação de textos devido à sua capacidade de lidar com palavras fora do vocabulário.

Além desses modelos, algoritmos clássicos de aprendizado de máquina, como o Support Vector Machine (SVM) e o Naive Bayes, têm desempenhado papéis importantes na análise de texto. O SVM, por exemplo, é conhecido por sua capacidade de separar eficientemente dados em espaços multidimensionais, tornando-o útil na classificação de textos em diferentes categorias (Vapnik, 1995). Por outro lado, o Naive Bayes, baseado na teoria probabilística, é amplamente utilizado em tarefas de classificação de texto, como detecção de spam e análise de sentimento, devido à sua simplicidade e eficácia (Rish, 2001). Outra ferramenta relevante é o LIWC (Linguistic Inquiry and Word Count), que se destaca por analisar as características linguísticas e emocionais do texto, oferecendo insights valiosos sobre o conteúdo textual em análises sociais e psicológicas. Esses algoritmos e ferramentas, incluindo o LIWC, contribuem para a riqueza da análise de PLN no contexto das redes sociais, auxiliando na compreensão e mitigação do impacto do conteúdo radicalizado e extremista na sociedade.

Portanto, o estudo das dinâmicas de comunicação online em plataformas como o Telegram e a análise dos algoritmos de processamento de linguagem natural, como o BERT e o Perspective, podem desempenhar um papel fundamental na compreensão e na mitigação do impacto do conteúdo radicalizado e extremista na sociedade brasileira. A pesquisa visa não apenas identificar padrões de disseminação, mas também desenvolver estratégias eficazes para combater a radicalização online e promover uma comunicação mais saudável e segura na era digital. Conforme indicado por Ferrara (2020), "a análise de dados e a pesquisa interdisciplinar desempenham um papel vital na compreensão e na contenção da radicalização online".

5. Contribuição

O presente estudo propõe uma abordagem experimental e multifacetada para a análise da comunicação online em grupos do Telegram, com foco especial na disseminação de conteúdo potencialmente radicalizado e extremista. Ao abordar questões cruciais relacionadas à linguagem, sentimentos e percepções dos participantes desses grupos, busca-se oferecer uma compreensão mais profunda das interações e dos contextos que ocorrem nesses ambientes digitais.

Os dados foram previamente coletados pelo grupo de pesquisa da professora orientadora. Uma análise preliminar revelou um grande número de links externos, sobretudo para o YouTube, muitos dos quais apontavam para vídeos indisponíveis no momento da coleta. Além disso, foram coletados metadados dos vídeos do YouTube como parte do procedimento, destacando uma parcela significativa de links para vídeos já retirados da plataforma. Uma hipótese que está sendo investigada é que esses vídeos foram possivelmente removidos pela moderação da plataforma devido à presença de conteúdo extremista. Uma avaliação inicial qualitativa sugere a plausibilidade dessa hipótese e está sendo aprofundada por outro aluno da professora orientadora. Este estudo visa complementar as pesquisas mencionadas anteriormente.

A coleta de dados foi realizada de maneira sistemática, abrangendo grupos que representam diferentes espectros políticos e interesses opostos. A base de dados, inicialmente fornecida na forma de um "dump", foi posteriormente restaurada em um banco de dados PostgreSQL. Este banco de dados restaurado apresenta uma estrutura organizada em seis tabelas distintas, cada uma projetada para descrever aspectos específicos, como canais de comunicação, usuários envolvidos, mensagens trocadas, mídias compartilhadas e relacionamentos pertinentes. Destaca-se, no entanto, que dentre todas as tabelas, a de mensagens é a mais relevante, compreendendo mais de 13 milhões de registros. Esses registros, representativos das interações entre os usuários, assumem um papel central nesta investigação.

5.1. Metodologia:

Pré-processamento de Dados

O pré-processamento dos dados é uma etapa crucial para garantir a qualidade da análise. A remoção de ruídos de forma ad-hoc, como URLs e caracteres especiais, a lematização e a tokenização foram aplicadas de forma abrangente, entretanto ainda requer uma análise profunda do ruído restante para refinamento. A atenção especial à preservação da integridade das mensagens, aliada ao tratamento cuidadoso dos elementos linguísticos, contribui para a robustez das análises posteriores.

Captura e Definição da Janela de Contexto

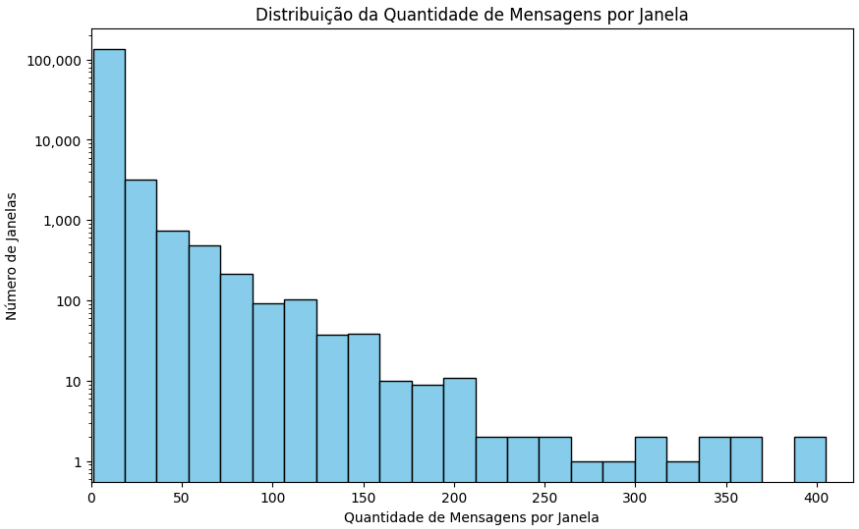
Na fase inicial do trabalho, o foco foi sobre a captura das mensagens pertinentes à análise proposta. A definição da janela de contexto em torno das mensagens contendo links para vídeos externos foi conduzida de forma empírica e avaliada após verificações sucessivas sobre amostras aleatórias. Critérios como a quantidade de mensagens anteriores e posteriores, aliados ao intervalo de tempo entre essas mensagens, consistiam nos atributos iniciais para a definição da janela, entretanto, foi conveniente usar apenas o intervalo de tempo a fim de não limitar a quantidade de mensagens dentro de um período. Optou-se em definir por iniciar a janela um minuto antes da mensagem identificada com o link e encerrá-la três minutos após. Estes métodos asseguram que as mensagens selecionadas tenham maior probabilidade de estarem contextualmente associadas ao compartilhamento de links, proporcionando uma análise mais precisa.

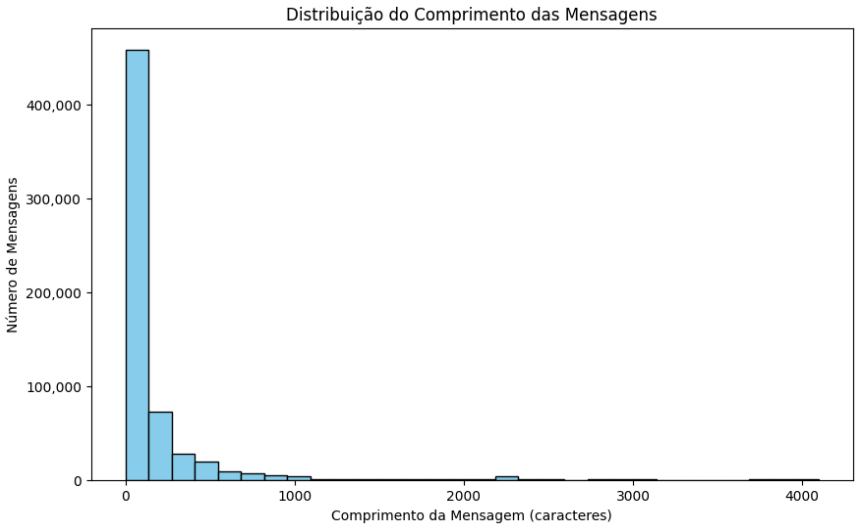
Realizou-se o processamento das janelas em torno das mensagens com links para as plataformas YouTube, Bitchute, Rumble, Odysee. Os dados foram capturados e inseridos em múltiplos arquivos testes, cada um com algum tipo de subconjunto estratégico sobre o conjunto filtrado, como, por exemplo, a exclusão de todas as mensagens repetidas independente da janela e do canal ou a exclusão das mensagens repetidas num canal numa mesma janela. A partir disso, as análises foram conduzidas sobre este segundo subconjunto, considerando se repetida quando enviada pelo mesmo autor.

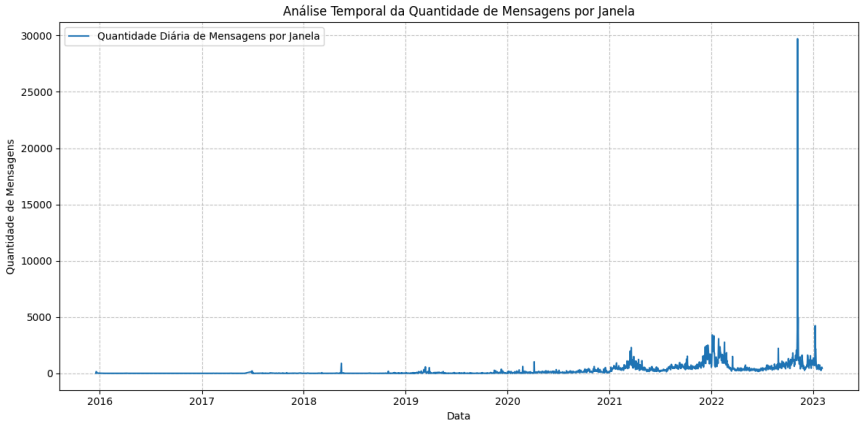
O conjunto de dados capturado a partir das estratégias descritas possui um total de 618.986 mensagens em 140.751 janelas, ao qual 185.778 são mensagens únicas; as janelas possuem uma média de 4,03 mensagens, enquanto que as mensagens possuem, em média, 23,04 palavras.

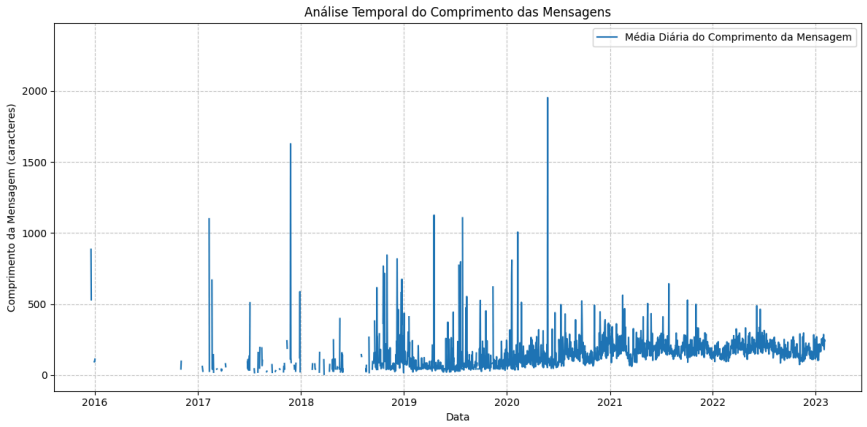
	Valor Médio	Desvio Padrão
Palavras por janela	4,037343	8,858479
Usuários por janela	1,422704	2,354134

A seguir, visões geradas sobre a quantidade de mensagens por janela e sobre o comprimento das mensagens, bem como a avaliação destes atributos ao longo do tempo.


Distribuição da Quantidade de Mensagens por Janela
Número de Janelas
100,000
10,000
1,000
100
10
1
0
50
100
150
200
250
300
350
400
Quantidade de Mensagens por Janela


Distribuição do Comprimento das Mensagens
Número de Mensagens
400,000
300,000
200,000
100,000
0
0
1000
2000
3000
4000
Comprimento da Mensagem (caracteres)


Análise Temporal da Quantidade de Mensagens por Janela
Quantidade Diária de Mensagens por Janela
Quantidade de Mensagens
30000
25000
20000
15000
10000
5000
0
2016
2017
2018
2019
2020
2021
2022
2023
Data



Podemos observar também algumas características do conjunto de dados em relação às janelas, às mensagens e aos usuários, segmentando-as nos domínios selecionados.

	Quantidade de janelas	Quantidade de mensagens	Quantidade de usuários
Bitchute	15.504	59.196	6.230
Odysee	3.593	13.167	1.441
Rumble	13.162	56.952	6.233
Youtube	108.397	438.869	22.759

Quantidade de mensagens, janelas e usuários em cada domínio.

	Valor Médio	Desvio Padrão
Bitchute	4,12539	7,0988
Odysee	4,00362	7,46803
Rumble	4,57582	9,09639
Youtube	4,4299	10,1806

Média e desvio padrão da quantidade de mensagens por janela em cada domínio.

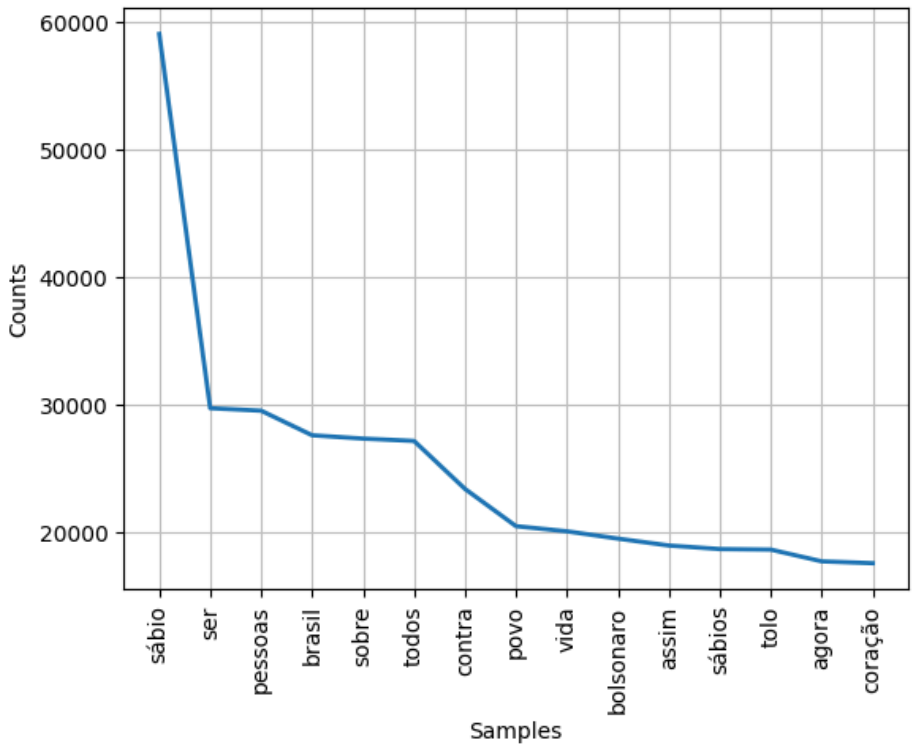
	Valor Médio	Desvio Padrão
Bitchute	1.3223	59.196
Odysee	0.995269	13.167
Rumble	13.162	56.952
Youtube	108.397	438.869

Média e desvio padrão da quantidade de usuários por janela em cada domínio.

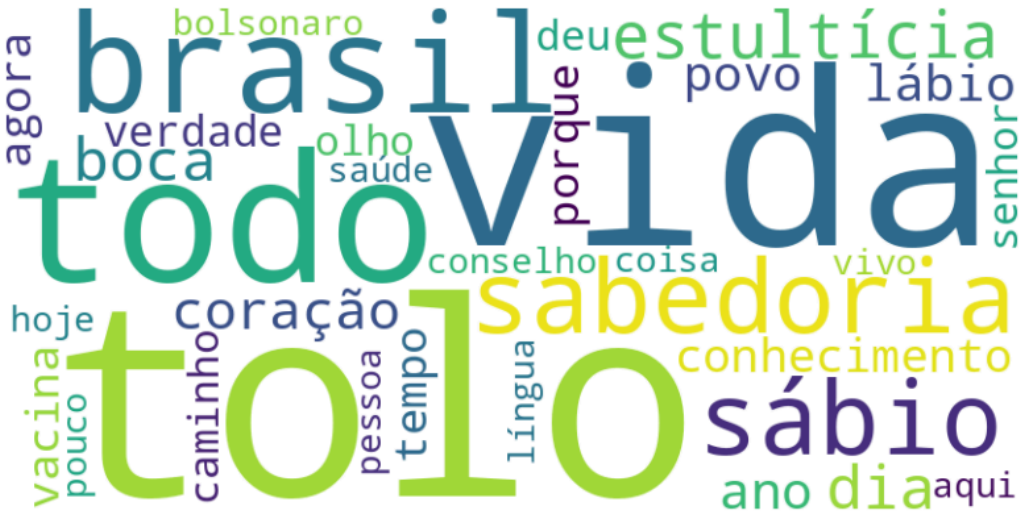
Análise Linguística

Uma abordagem visual através de nuvens de palavras e histogramas destacaram termos de impacto recorrentes nos subconjuntos dos domínios avaliados. A análise da frequência das palavras, vinculada ao tempo, fornece uma visão histórica das palavras compartilhadas que será melhor utilizada na análise qualitativa, na segunda parte deste trabalho. Este duplo enfoque poderá proporcionar insights valiosos sobre a evolução temporal dos temas em discussão nos grupos do Telegram, a fim de uma caracterização do contexto das janelas capturadas. Segue exemplos da frequência de palavras para o domínio geral e para cada um dos domínios segmentados.

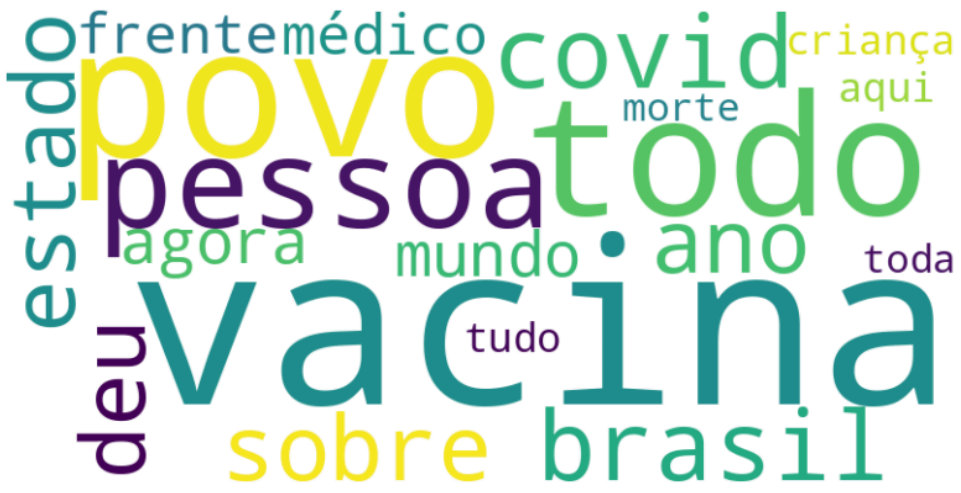
Nuvem de palavras para o conjunto geral.



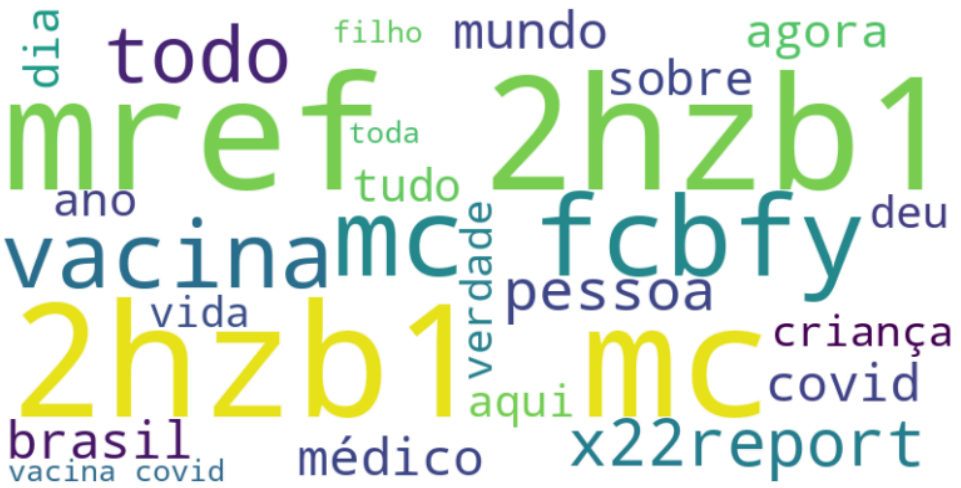
Histograma para o conjunto geral.



Nuvem de palavras para o conjunto do Youtube.



Nuvem de palavras para o conjunto do Bitchute.



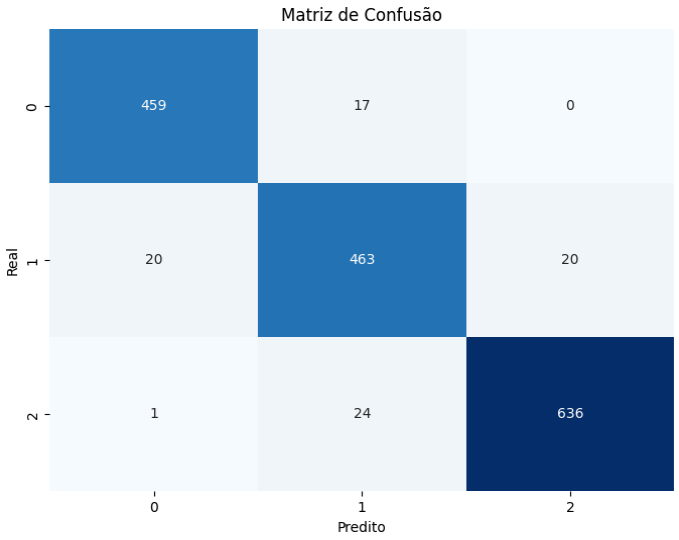
Nuvem de palavras para o conjunto do Rumble.

Nuvem de palavras para o conjunto do Odysee.

A aplicação de técnicas avançadas de PLN é o cerne deste estudo. A priori, algoritmos como SVM, Naive Bayes, BERTopic e LIWC foram propostos para explorar as mensagens com algum nível de profundidade. Essas técnicas não apenas proporcionam uma análise de sentimentos, mas também buscam extrair tópicos, padrões linguísticos e características emocionais, contribuindo assim para uma compreensão mais holística da comunicação em grupos do Telegram. Entretanto, algumas dessas técnicas só serão aplicadas na segunda parte deste trabalho, e outras serão testadas e possivelmente descartadas, embora possam ser aplicadas novamente.

Modelo Naive Bayes

O treinamento e aplicação de um modelo Naive Bayes para a classificação de sentimentos representou uma etapa crucial. Foi utilizado um conjunto de 8.199 mensagens previamente classificadas para treino. A acurácia de 95% atestada após o treinamento valida a eficácia do modelo. O modelo apresentou o seguinte desempenho:



Em que 0 é ‘Negativo’, 1 é ‘Positivo’ e 2 é ‘Negativo’.

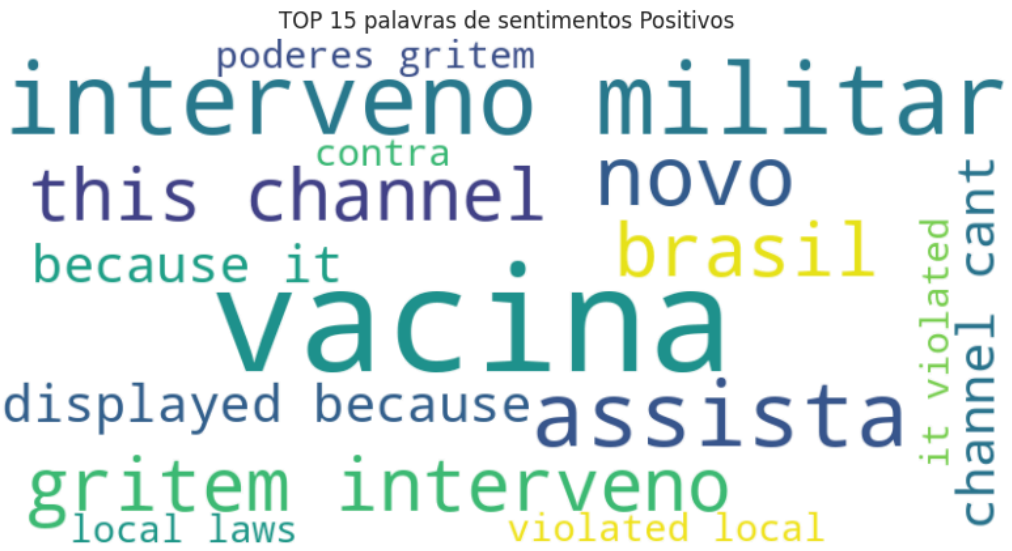
	Precision	Recall	F1-Score	Support
Negativo	0.96	0.96	0.96	476
Neutro	0.92	0.92	0.92	503
Positivo	0.97	0.96	0.97	661
Accuracy	-	-	0.95	1.640
Macro Avg.	0.95	0.95	0.95	1.640
Weighted Avg.	0.95	0.95	0.95	1.640

A partir deste ponto, as análises serão concentradas num subconjunto filtrado em 480.188 mensagens, todas pertencentes à janelas relacionadas à links para o domínio “youtube". Ao se concentrar apenas nas mensagens relacionadas aos links da plataforma YouTube, as análises atuais se posicionam como um ponto de partida para uma análise mais aprofundada e direcionada na próxima fase do trabalho. Após a aplicação do modelo ao subconjunto de dados do YouTube, temos o seguinte resultado:

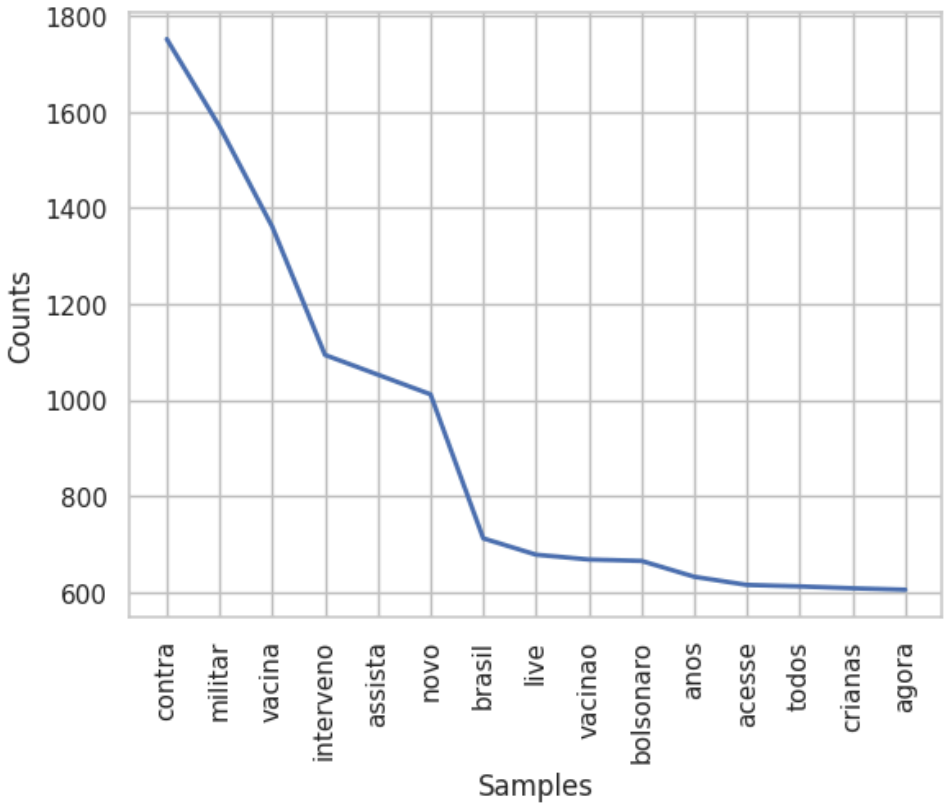
	Quantidade de Ocorrências	Proporção dos Sentimentos
Negativo	213.873	44,64%
Neutro	195.891	48,73%
Positivo	29.105	6,63%

Análise de Sentimentos

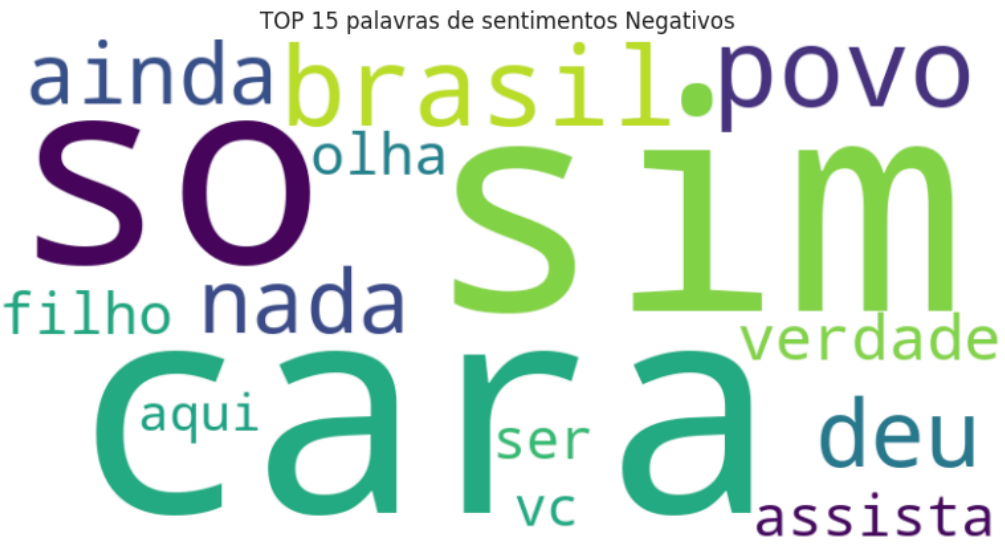
A análise da frequência de palavras em consonância com os sentimentos classificados em tons negativo, neutro e positivo proporciona uma compreensão mais rica das reações dos usuários.



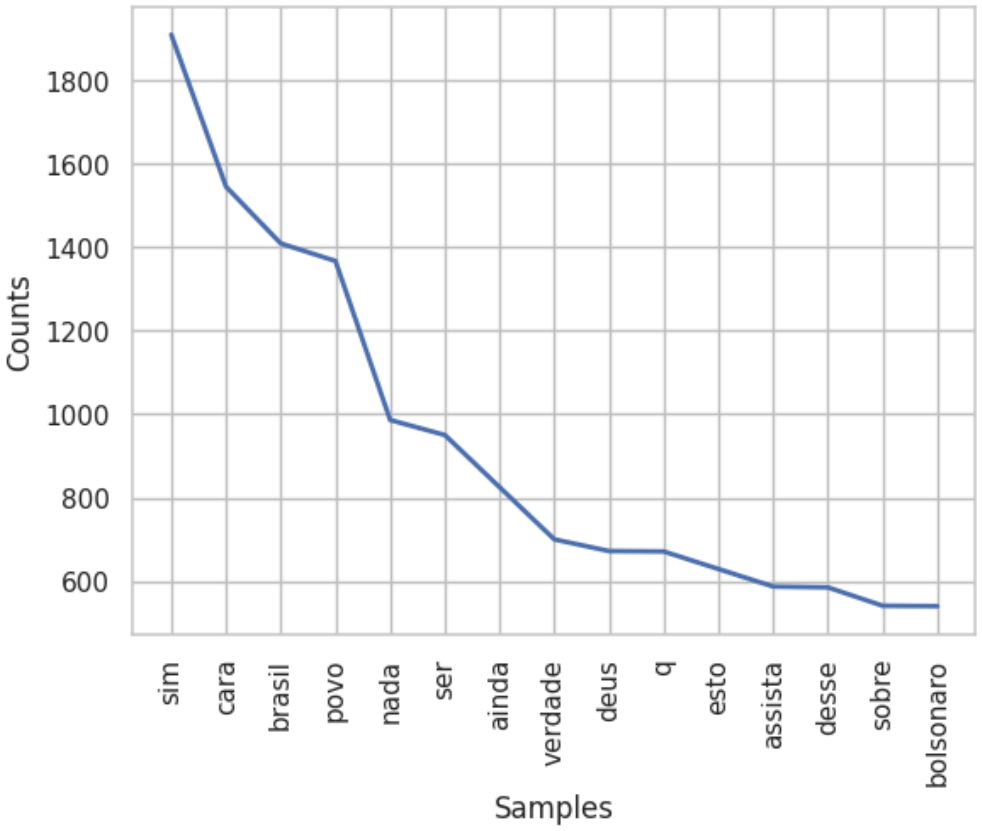
As 15 palavras mais frequentes em ocorrência de sentimentos positivos.



Histograma das palavras em ocorrência de sentimentos positivos.

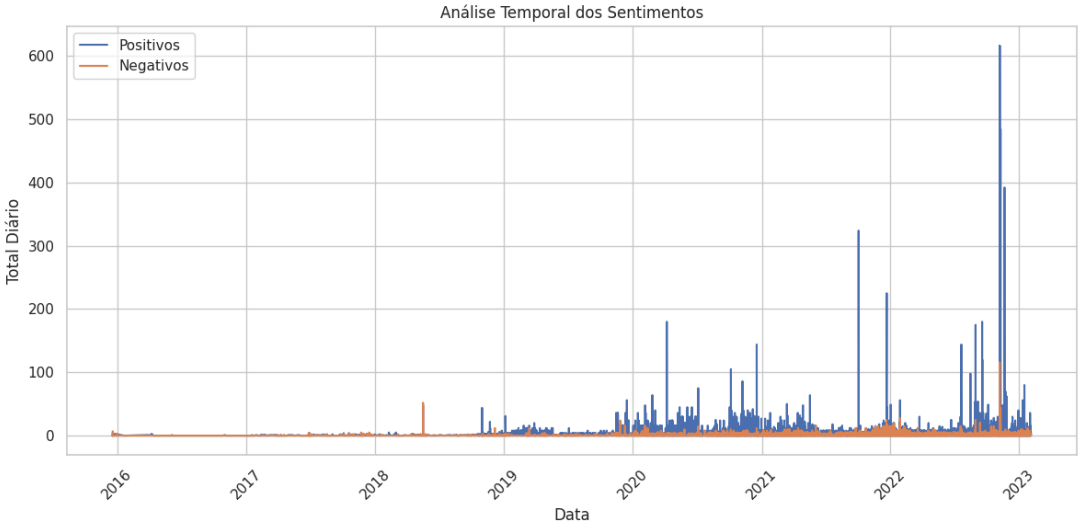


As 15 palavras mais frequentes em ocorrência de sentimentos negativos.



Histograma das palavras em ocorrência de sentimentos negativos.

A dinâmica temporal desses sentimentos é explorada, permitindo a identificação de padrões e flutuações diárias. As visualizações gráficas geradas nesta etapa são essenciais para enriquecer as análises que serão desenvolvidas na MSI2.



Visão temporal das ocorrências de sentimentos positivos e negativos.

Análise Linguística Aprofundada

Este capítulo tem como objetivo a aplicação de técnicas avançadas de análise linguística e qualitativa para compreender melhor as dinâmicas de comunicação em grupos do Telegram, especialmente em relação a links de vídeos do YouTube. Utilizando ferramentas como o LIWC (Linguistic Inquiry and Word Count), o SentiStrength e o Perspective, foi possível categorizar palavras e identificar padrões de discurso que revelam emoções, comportamentos e interações sociais dos participantes.

LIWC

Inicialmente, foi utilizado o LIWC para agrupar palavras por suas funções e contextos. Este processo revelou uma ampla variedade de categorias linguísticas presentes nas mensagens analisadas, que foram agrupadas em tópicos principais como emoções, estilo de linguagem, sociedade e relacionamentos, comportamento e atividades, cognição e pensamento.

Os dados analisados foram categorizados de acordo com as seguintes funções e contextos:

Categoria	Subcategoria	Termo	Ocorrências
Emoções	Positivas	posemo	399.213
		affect	761.660
		leisure	171.339
		health	107.420
		friend	41.056
		assent	32.846
	Negativas	negemo	321.862
		anger	184.453
		sad	66.761
		anx	60.479
		death	20.633
		negate	14.079
Estilo de Linguagem	Pronomes e Determinantes	pronoun	1.479.763
		article	895.966
	Verbos e Advérbios	verb	851.260
		adverb	147.645
	Conjunções e Verbos Auxiliares	conj	761.559
		auxverb	308.815
Social e Relacionamento	Socialização	social	1.686.168
	Família	friend	41.056
		family	20.522
	Pronomes Pessoais	they	201.721
		we	10.579
Comportamento e Atividades	Trabalho	work	258.868
	Lazer	leisure	171.339
	Ingestão	ingest	421.667
	Movimento	motion	420.699
	Corpo	body	193.658
Cognição e Pensamento	Processos Cognitivos	cogmech	2.851.549
		percept	285.188
		insight	411.690
		discrep	389.262

		tentat	573.311
Tempo e Espaço	Tempo	time	486.792
	Espaço	space	1.259.961
Saúde e Bem-Estar	Saúde	health	107.420
	Biologia	bio	458.903
Outros	Palavras Não Fluídas	nonfl	315.196
	Quantidade	quant	248.042
	Humanos	humans	312.374
	Causa	cause	259.158

Para a avaliação do tom emocional, palavras indicativas de emoções positivas e negativas foram classificadas com base em seu significado e estrutura frasal. As palavras foram divididas entre positivas (como afeto, lazer, saúde, amizade, concordância) e negativas (como raiva, tristeza, ansiedade, morte, negação).

Identificou-se termos frequentemente associados a pautas sociais, como gênero, socialização, associação, coletividade, humanidade, família e religião. Esta análise permitiu uma compreensão mais detalhada das interações sociais e das temáticas discutidas nos grupos.

A análise da construção do discurso focou em identificar a intenção do emissor, como busca por validação, inspiração, conexão, imposição ou confiança. Essa etapa é crucial para compreender não apenas o conteúdo das mensagens, mas também o contexto e a motivação por trás das comunicações.

SentiStrength

Neste passo, a ferramenta SentiStrength foi utilizada para avaliar a distribuição e a intensidade dos sentimentos presentes nas mensagens dos grupos do Telegram. A análise focou-se na distribuição dos sentimentos e na força média dos sentimentos positivos e negativos e, para processamento dos dados na ferramenta, foi utilizado uma distribuição open source do léxico de 2007 para o Português..

A ferramenta foi aplicada para analisar a intensidade dos sentimentos nas mensagens, categorizando-as como positivas, negativas ou neutras. O SentiStrength é conhecido por sua capacidade de medir a força dos sentimentos expressos em textos curtos, como mensagens em redes sociais, oferecendo uma escala de -5 (muito negativo) a +5 (muito positivo).

Resultados da análise:

Métrica	**Valor**	**Descrição**

Média da Escala de Sentimentos	0,073	A média próxima de zero sugere que os textos analisados são majoritariamente neutros ou contêm sentimentos ligeiramente positivos.
Textos Positivos	75.523	Quantidade de textos classificados como positivos (terceira posição no resultado trinário: 1).
Textos Negativos	69.668	Quantidade de textos classificados como negativos (terceira posição no resultado trinário: -1).
Textos Neutros	334.998	Quantidade de textos classificados como neutros (terceira posição no resultado trinário: 0).
Média da Força Positiva	1,378	A força média dos textos positivos, indicando sentimentos ligeiramente positivos.
Média da Força Negativa	-1,305	A força média dos textos negativos, indicando sentimentos ligeiramente negativos.

Os resultados mostram que a maioria das mensagens analisadas é neutra, com 334.998 textos classificados nesta categoria. Este dado sugere que as interações nos grupos do Telegram tendem a ser informativas ou descritivas, com pouca expressão de sentimentos intensos. A quantidade de textos positivos (75.523) é ligeiramente maior que a de textos negativos (69.668), corroborando a média da escala de sentimentos de 0,073, que indica uma leve tendência a sentimentos positivos.

A média da força dos sentimentos positivos (1,377) e negativos (-1,304) indica que, quando presentes, os sentimentos expressos nas mensagens são geralmente fracos. Isso sugere que as interações emocionais, tanto positivas quanto negativas, são moderadas e raramente intensas. Portanto, é necessário elevarmos a análise ao nível das exceções para entendermos os casos de possível radicalização.

Amostras de maior força positiva:

Texto	Pontuação
“C U I D A D O patriotas, lembrem-se: este dia será o momento perfeito para a esquerda reproduzir o efeito capitólio. fiquem longe da esplanada, não parem, não precipitem, não retrocedam. a vitória está próxima”	4

“A gente era tremendamente feliz e não sabíamos. Led Zeppelin - Stairway to Heaven. Eu cantava esta bela melodia no conjunto aos 18 anos de idade”	4
"hummm, penso que seja um sinal muito positivo!"	4

Amostras de maior força negativa:

Texto	Pontuação
“Foi Dias Toffoli que começou com o inquérito ilegal, apoiado pelo Bolsolixo inclusive, e o retardado marmita de bandido disse que a foto do encontro do Bolsocorno com o Dias Toffoli era mentira. Bolsonaro foi o presidente que mais fortaleceu o sistema corrupto e mais aprovou leis socialistas”	-4
“O quartel era o único lugar que podíamos ir sem que a mídia dissesse que os protestos eram contra o governo Bolsonaro, distorcendo tudo. Hoje, não tem mais Bolsonaro. podemos parar tudo e a culpa ficará nítida de quem são os corruptos que tomaram o poder.”	-4
“fala pessoal, não gostaria de ter que gravar este vídeo, mas hoje vou falar sobre esse estuprador que cometeu ato escatológico contra a vítima”	-4

Perspective

Em complemento ao SentiStrength, foi utilizado o Perspective para realizar uma análise detalhada dos sentimentos, focando em identificar níveis de toxicidade e insultos nas mensagens. A ferramenta Perspective é desenvolvida para detectar padrões de discurso prejudicial e fornecer insights sobre a qualidade e o impacto das comunicações em plataformas de mídia social. As estatísticas resumidas fornecem uma visão geral das médias de toxicidade e insultos nas mensagens analisadas:

Métrica	Valor
Média de Toxicidade	15,364
Média dos Insultos	8,409

Esses valores indicam que, em média, os textos analisados não apresentam um nível elevado de toxicidade ou insultos, sugerindo que a maioria das comunicações tende a ser relativamente neutra ou moderada. Para uma compreensão mais detalhada, vamos analisar os textos com as pontuações mais extremas de toxicidade e insultos.

Texto mais tóxico:

Texto	Pontuação
"filhos da puta comunista vo queimar no inferno seus lacaios dos vermes do esgoto"	98,176

Este texto apresenta uma pontuação de toxicidade extremamente alta, refletindo o uso de linguagem extremamente ofensiva e violenta. A alta pontuação é apropriada dado o conteúdo altamente prejudicial e agressivo.

Texto menos tóxico:

Texto	Pontuação
"60 filmes séries e documentários para despertar uma nova consciência..."	0,076

Este texto possui uma pontuação de toxicidade muito baixa, o que é consistente com seu conteúdo neutro e informativo, indicando títulos de filmes e séries. A pontuação baixa reflete adequadamente a ausência de linguagem prejudicial.

Texto mais insultante:

Texto	Pontuação
"homofóbico fascista nazista racista burro mediocre arrogante broxa manipulado obtuso filho da puta corno nazista e racista"	93,461

A pontuação de insultos para este texto é muito alta, o que é apropriado dado o uso intenso de termos pejorativos e ofensivos. Este texto é carregado de insultos direcionados, justificando a alta pontuação.

Texto menos insultante:

Texto	Pontuação
"lista de brasília atual presencial e telemedicina 2 hercilia pimenta telemedicina..."	0,510

Este texto tem uma pontuação de insultos muito baixa, refletindo seu conteúdo neutro e informativo sobre contatos de profissionais de saúde. A pontuação baixa é apropriada dada a ausência de linguagem ofensiva.

Em resumo, as médias de 15,364 para toxicidade e 8,409 para insultos indicam que, em geral, o conjunto de textos tende a ser relativamente pouco tóxico e insultante. Isso sugere que a maioria das mensagens possui um tom moderado, sem grande incidência de linguagem prejudicial. O texto com a maior pontuação de toxicidade (98,176) contém linguagem altamente ofensiva, justificando a pontuação extrema. Este tipo de análise é crucial para identificar mensagens que podem necessitar de moderação para manter um ambiente saudável nas plataformas de mídia social. O texto com a maior pontuação de insultos (93,461) também é extremamente ofensivo, com múltiplos insultos direcionados. Identificar esses textos é importante para entender o impacto negativo potencial sobre os usuários e tomar ações corretivas. Textos com pontuações muito baixas de toxicidade e insultos confirmam a presença de comunicações neutras e informativas, como esperado. Estes textos ajudam a balancear a visão geral do ambiente comunicacional, indicando que nem todas as interações são negativas.

Amostras de maior força para toxicidade:

Texto	Pontuação
“vai se fuder bicha do caralho”	98,1
“vai tomar no cu idiota do caralho”	97,4
“sua bicha do caralho”	96,8

Amostras de maior força para insultos:

Texto	Pontuação
“vai se fuder bicha do caralho”	91
“sua bicha do caralho”	90,2
“que bicho escroto putaquepariu”	89,8

BERT/BERTimbau

Além das ferramentas avançadas como LIWC, SentiStrength e Perspective, também foi testado a aplicação de um modelo BERTimbau e a ferramenta BERTopic para uma extração de tópicos eficaz, a fim de um aprimoramento das análises linguísticas e qualitativas realizadas até este ponto.

O BERTimbau, um modelo BERT pré-treinado para o português, foi escolhido para melhorar a qualidade das representações textuais, capturando nuances linguísticas e contextuais com maior precisão. Já o BERTopic tinha como objetivo extrair tópicos latentes dos textos, fornecendo uma visão estruturada das discussões nos grupos. Contudo, a aplicação desses modelos enfrentou sérios desafios computacionais. Mesmo com o uso quase que integral dos recursos da máquina utilizada para este projeto, o processamento com o BERTimbau foi extremamente lento e custoso, tornando inviável a conclusão da análise e, portanto, as dificuldades técnicas impediram a extração eficiente de tópicos com o BERTopic.

Embora não tenha alcançado os resultados esperados, a tentativa de usar BERTimbau e BERTopic destacou a importância dessas ferramentas para análises futuras. As dificuldades encontradas indicam a necessidade de melhores recursos computacionais e técnicas de otimização para integrar plenamente esses modelos.

6. Conclusão

Este estudo espera fornecer insights sobre como os grupos do Telegram lidam com conteúdo radicalizado e extremista por meio de diálogos textuais relacionados ao envio de vídeos com conteúdo potencialmente extremista. Utilizando ferramentas avançadas de análise linguística e qualitativa, como LIWC, SentiStrength e Perspective, além da análise de sentimentos com um modelo Naive Bayes realizada na primeira parte deste trabalho, foram identificados padrões de discurso que revelam emoções, comportamentos e interações sociais dos participantes.

A análise categorizou palavras em funções e contextos diversos, revelando emoções positivas e negativas, questões sociais e intenções dos emissores. Isso permitiu uma compreensão detalhada das interações sociais e temáticas discutidas nos grupos. A maioria das mensagens foi classificada como neutra, com uma leve tendência a sentimentos positivos. A intensidade dos sentimentos expressos foi geralmente baixa, sugerindo interações emocionais moderadas.

Não obstante, a análise focou-se em níveis de toxicidade e insultos que, em média, apresentaram níveis elevados em contraposição aos baixos níveis gerais da base de dados utilizada, sugerindo uma comunicação relativamente neutra ou moderada. Algumas mensagens, no entanto, continham linguagem altamente ofensiva, destacando a necessidade de moderação contínua.

O uso de modelos avançados de processamento de linguagem natural, como o Perspective, demonstrou ser eficaz na análise das reações e percepções dos participantes em relação a vídeos externos com conteúdo potencialmente extremista. Este estudo fornece uma base promissora para futuras pesquisas, contribuindo para o entendimento do conteúdo e de sua dinâmica, além da identificação de discursos prejudiciais nas mídias sociais.

As limitações inerentes, especialmente em relação à representatividade das mensagens avaliadas, indicam a necessidade de explorar técnicas mais avançadas de PLN e outras estratégias para expandir essa análise. Em pesquisas futuras, deve-se incorporar uma gama mais ampla de informações e perspectivas sobre os contextos das mensagens, além de um refinamento sistemático do tratamento dos dados e das análises já realizadas, garantindo uma compreensão mais abrangente e coesa sobre as dinâmicas de compartilhamento de conteúdo potencialmente extremista no Telegram.

Em suma, este trabalho propõe uma contribuição razoável para a análise das interações em grupos de Telegram e da comunicação no discurso de ódio, fornecendo insights sobre o comportamento dos usuários e os diferentes discursos no compartilhamento de conteúdo extremista.

Bibliografia

Castells, M. (2010). The Rise of the Network Society: The Information Age: Economy, Society, and Culture. Wiley.

Berger, J. M. (2019). Extremism. The MIT Press.

Winter, C. (2020). Telegram: the 'app of choice' for sharing extremist content. BBC News. Disponível em: https://www.bbc.com/news/technology-54581938

Jang, Y., & Kim, Y. (2020). Aspect-Based Sentiment Analysis of Movie Reviews on Twitter with BERT. IEEE Access, 8, 90080-90087.

McCauley, C., & Moskalenko, S. (2011). Toward a Profile of Lone Wolf Terrorists: What Moves an Individual from Radical Opinion to Radical Action.

Silke, A. (2008). Holy Warriors: Exploring the Psychological Processes of Jihadi Radicalization. European Journal of Criminology, 5(1), 99-123.

Neumann, P. R. (2013). Joining al-Qaida: Jihadist Recruitment in Europe. International Security, 37(4), 93-131.

Russell, S. J., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach. Pearson.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Bidirectional Encoder Representations from Transformers. arXiv preprint arXiv:1810.04805.

van der Plas, E., & Grootendorst, M. (2022). BERTimbau: A Pre-Trained BERT Model for Brazilian Portuguese. In Proceedings of the 31st International Conference on Computational Linguistics (COLING 2022)

Grootendorst, M. (2022). BERTopic. Disponível em: https://maartengr.github.io/BERTopic/index.html.

Perspective API. (s.d.). Recuperado em 02 de setembro de 2023, de https://perspectiveapi.com

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems (pp. 3111-3119).

Radford, A., Karthik, N., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-training.

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., Bagheri, M., Lample, G., ... & Usunier, N. (2016). FastText.zip: Compressing text classification models.

Vapnik, V. N. (1995). The nature of statistical learning theory. Springer Science & Business Media.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net.

Ferrara, E. (2020). COVID-19, conspiracy theories, and the nexus of online hate. Social Media+ Society, 6(3), 2056305120948165.