Revelando a dinâmica das discussões em grupos públicos do Telegram: um estudo sobre as eleições no Brasil em 2022

Revelando a dinâmica das discussões em grupos públicos do Telegram: um estudo sobre as eleições no Brasil em 2022

Daniel N. Junqueira[1], **Carlos H. G. Ferreira**[2,4], **Giordano Paoletti**[3], **Jussara M. Almeida**[4]

[1] Departamento de Engenharia Elétrica - Universidade Federal de Minas Gerais
[2]Departamento de Computação e Sistemas - Universidade Federal de Ouro Preto
[3]*Dipartimento di Automatica e Informatica - Politecnico di Torino*
[4]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais

danijnog@ufmg.br, chgferreira@ufop.edu.br,

giordano.paoletti@polito.it, jussara@dcc.ufmg.br

Abstract. *This study presents a new methodology to analyze the dynamics of discussions in online groups, as well as its application to investigate the dynamics of political discourse in Brazilian public groups on Telegram during the 2022 presidential election. Our methodology involves the use of: (1) a large language model (LLM) to summarize the messages shared in a group during a time interval; (2) embeddings to represent the summaries; and (3) an unsupervised learning algorithm to cluster the representations, thus identifying large themes of discussion. The discussion themes are then characterized around the main topics and terms that make up the discussions, and the temporal evolution of the groups around each theme is analyzed. Our results reveal the main contents of the discussions and how they evolve over time and in response to political events and happenings, such as January 8, highlighting how social networks can influence the formation of political discourse in society.*

Resumo. *Esse estudo apresenta uma nova metodologia para analisar a dinâmica de discussões em grupos online, bem como sua aplicação para investigar a dinâmica do discurso político em grupos públicos brasileiros no Telegram durante a eleição presidencial de 2022. Nossa metodologia envolve o uso de: (1) um grande modelo de linguagem (LLM) para sumarizar as mensagens compartilhadas em um grupo durante um intervalo de tempo, (2) embeddings para representação dos sumários e (3) um algoritmo de aprendizado não supervisionado para agrupar as representações, identificando assim grandes temas de discussão. Os temas de discussão são então caracterizados ao redor dos principais tópicos e termos que compõem as discussões, e a evolução temporal dos grupos em torno de cada tema é analisada. Nossos resultados revelam os conteúdos principais das discussões e como os mesmos evoluem ao longo do tempo e em resposta a eventos e acontecimentos políticos, como o 8 de Janeiro, evidenciando como as redes sociais podem influenciar a formação do discurso político na sociedade.*

1. Introdução

As redes sociais exercem notável influência sobre diversos aspectos da sociedade, especialmente na forma como a comunicação se estabelece atualmente. Diferentemente

de sites de notícias ou jornais online, esse tipo de mídia digital tem como principais características a interatividade e a participação dos usuários, transformando-se de um espaço meramente voltado ao compartilhamento de informações em um ambiente de produção, modificação e transformação do conteúdo informacional [Malagoli et al. 2021; Araujo et al. 2023; G. Da Fonseca et al. 2024].

Considerando os aplicativos de mensagens instantâneas muito populares atualmente, o Telegram se destaca por recursos vantajosos para disseminação de informações, como canais de transmissão sem limites de membros e supergrupos com maior capacidade que o WhatsApp. No entanto, essa ampla disseminação, aliada à pouca moderação de conteúdo, facilita a propagação de conteúdos problemáticos como notícias falsas, conspiratórias e narrativas hiperpartidárias [Bär et al. 2023; Hoseini et al. 2024].

Em um contexto político, o Telegram emerge como uma plataforma significativa para disseminação de informações e debates políticos, refletindo as opiniões de seus usuários. De fato, vários trabalhos anteriores já analisaram propriedades dos usuários e do conteúdo compartilhado em grupos públicos do Telegram, tanto no contexto político [Júnior et al. 2022; Bovet and Grindrod 2022], como em outros domínios [Perlo et al. 2025; Xu and Livshits 2019]. Especificamente, os autores de [Venâncio et al. 2024b] analisaram a coordenação de usuários para disseminar conteúdos relacionados aos movimentos sociais que culminaram com a invasão das sedes dos Três Poderes em Brasília em 8 de janeiro de 2023. No geral, estes trabalhos, assim como outros que analisaram a disseminação de informação em grupos do WhatsApp [Nobre et al. 2020; Nobre et al. 2022], apresentam uma análise estática, considerando todo o conteúdo compartilhado durante um dado intervalo de tempo, ou, na melhor das hipóteses, uma análise comparativa em alguns poucos intervalos diferentes.

Diferentemente dos trabalhos anteriores, mais do que apenas revelar os principais tópicos de discussão, nossa pesquisa é motivada pelo interesse em analisar a *dinâmica* das discussões em grupo e o papel que estas discussões online exercem sobre a formação do discurso político na sociedade. Para tanto, nós propomos uma metodologia de análise da evolução temporal das discussões em grupos públicos, tendo o Telegram e as eleições presidenciais brasileiras de 2022 como estudo de caso. Nossa metodologia contempla o uso de um grande modelo de linguagem (*Large Language Model* – LLM) para sumarizar a discussão em um grupo durante uma janela de tempo e o uso de *embeddings* para representar os vários sumários (vários grupos, várias janelas de tempo) em um espaço comum de alta dimensão. Esta representação dos sumários permite a posterior utilização de técnicas de aprendizado não supervisionado (*clustering*) para agrupar as discussões em grandes temas (*clusters*). Os temas de discussão correspondentes a cada *cluster* são então caracterizados a respeito dos principais tópicos e termos que compõem as discussões, oferecendo uma visão mais detalhada das mesmas. Por fim, a evolução temporal dos grupos em torno de cada grande tema/*cluster* é analisada, a fim de entender a dinâmica das discussões ao longo do tempo.

Nosso estudo analisa mais de um milhão de mensagens compartilhadas em 211 grupos públicos brasileiros do Telegram entre o período de 25/09/2022 a 15/01/2023. Nós investigamos a evolução das discussões nestes grupos diariamente ao longo deste período. Nossos resultados revelam os principais temas de discussão (*clusters*), refletindo a situação política do país, e mostram, com granularidade diária, a dinâmica da discussão

nos grupos em torno destes temas, com destaque para períodos em torno de datas específicas, como o 1º e o 2º turno das eleições e o 8 de janeiro. Mais precisamente, mostramos a convergência de muitos grupos públicos em torno de temas específicos fortemente relacionados aos movimentos em Brasília, mostrando o papel importante da plataforma na formação do discurso político que marcou a sociedade brasileira naquele período.

2. Trabalhos Relacionados

Nos últimos anos, vários trabalhos analisaram propriedades das discussões online em diferentes redes sociais, incluindo o X [Guimarães et al. 2022; Pinto et al. 2024] e o Instagram [Ferreira et al. 2021; Thorgren et al. 2024]. Considerando o Instagram, por exemplo, Ferreira *et al.* adotaram uma modelagem baseada em redes para analisar as discussões políticas em torno de eleições no Brasil e na Itália. Já em Thorgren *et al.*, os autores focaram no engajamento dos usuários em diferentes tipos de conteúdos. Considerando plataformas de mensagens instantâneas, como o WhatsApp e o Telegram, vários trabalhos analisaram os conteúdos compartilhados em grupos públicos. Por exemplo, propriedades dos conteúdos compartilhados em grupos do WhatsApp durante eleições brasileiras foram analisadas em [Maros et al. 2020; Resende et al. 2019]. Já em [Chagas 2022], os autores focaram na disseminação de propagandas durante as eleições de 2018, enquanto em [Nobre et al. 2022] os autores exploraram uma modelagem de redes, detecção de comunidades e extração de backbones para identificar campanhas de disseminação de notícias falsas.

Mais recentemente, o Telegram tem despertado o interesse de vários pesquisadores, dada a sua crescente popularidade no mundo inteiro. Alguns estudos focaram nas funcionalidades da plataforma [Anglano et al. 2017] ou no seu uso por grupos de usuários específicos (e.g., imigrantes [Nikkhah et al. 2018], terroristas [Alrhmoun et al. 2023] ou grupos extremistas [Kloo et al. 2024]). Alguns trabalhos estudaram propriedades do conteúdo textual e os padrões de uso da plataforma [Naseri and Zamani 2019; Wich et al. 2022]. Em [Paoletti et al. 2025], os autores analisaram os principais fatores que estimulam os usuários a participarem ativamente de discussões em grupos do Telegram, e como estes fatores variam dependendo do principal tema de discussão associado a cada grupo.

Especificamente focando em grupos políticos, os autores de [Júnior et al. 2022] propuseram o sistema *"Telegram Monitor"* para monitorar o debate político na plataforma e analisar o conteúdo compartilhado entre grupos e usuários. Mais recentemente, um estudo explorou modelagem de redes, técnicas de extração de *backbone* de redes e modelagem de tópicos para revelar ações coordenadas de usuários da plataforma para espalhar conteúdos anti-democráticos durante as eleições brasileiras de 2022 [Venâncio et al. 2024a; Venâncio et al. 2024b].

Em comum, a maioria desses trabalhos analisou o conteúdo de forma estática, considerando de maneira agregada todas as mensagens do período estudado ou, na melhor das hipóteses, realizaram comparações entre poucos intervalos. Em contraste, nosso interesse está em compreender a dinâmica das discussões com granularidade temporal diária, analisando como elas evoluem ao longo do tempo. Para isso, propomos uma metodologia que combina o uso de LLMs, *embeddings* e algoritmos de agrupamento, aplicada ao estudo da evolução das discussões em grupos públicos do Telegram. Esta metodologia é descrita a seguir.

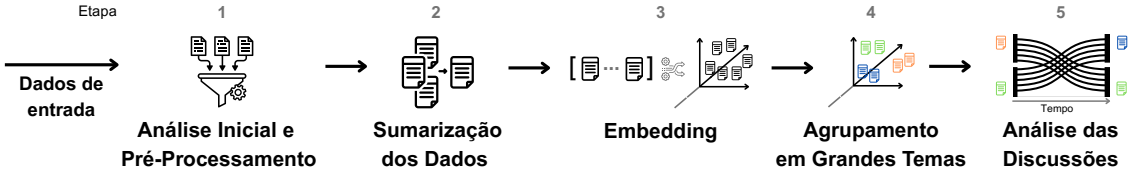


Figura 1. Visão geral da metodologia usada.

3. Metodologia

Nesta seção, descrevemos a metodologia proposta para analisar a dinâmica das discussões em grupos, a partir de um conjunto de dados de entrada. A abordagem é genérica e pode ser aplicada em diferentes contextos. Neste artigo, a utilizamos para investigar a evolução das discussões em grupos públicos do Telegram durante as eleições presidenciais brasileiras de 2022. Para isso, empregamos o mesmo conjunto de dados analisado em [Venâncio et al. 2024a], que reúne mais de 1,8 milhão de mensagens coletadas em 287 grupos públicos distintos, entre setembro de 2022 e janeiro de 2023. Como aquele trabalho já havia identificado campanhas de desinformação associadas aos movimentos que culminaram na invasão das sedes dos Três Poderes em Brasília, optamos por utilizar a mesma base de dados visando oferecer uma visão mais refinada e detalhada, com ênfase na evolução temporal das discussões que fomentaram tais campanhas[1]. A metodologia é composta por cinco etapas principais, conforme ilustrado na Figura 1 e detalhado nas próximas subseções.

3.1. Análise Inicial e Pré-Processamento

Os dados são compostos por um conjunto de mensagens previamente anonimizadas, cada uma representada por: *ID* da mensagem, *ID* do grupo/canal, conteúdo textual com data e hora de envio, e *ID* do autor. A fim de estudar a dinâmica das discussões com granularidade diária, separamos as mensagens compartilhadas em cada grupo por dia, analisando separadamente os conteúdos de cada par dia-grupo. Como nosso interesse está centrado em discussões ativas entre os usuários, filtramos os pares dia-grupo com número muito reduzido de mensagens, indicando pouca interação. A partir da análise da distribuição do número de mensagens por par dia-grupo, foi identificado um limiar mínimo de 25 mensagens — correspondente ao "cotovelo" da curva.

Em seguida, as mensagens passaram por um processo de limpeza, removendo emojis, URLs, menções a usuários, espaços extras, quebras de linhas repetidas, caracteres e palavras consecutivas repetidas. Esta limpeza foi feita para melhorar o processo de geração de sumários pela LLM (próxima etapa). Há um limite máximo de *tokens* que a LLM processa por vez (16384 para a LLM usada neste trabalho). Entradas mais longas são necessariamente truncadas. A remoção dos itens acima foi feita, pois evita a perda de conteúdo informativo devido ao truncamento. Também observamos uma melhor qualidade dos sumários (menos erros) produzidos após a remoção feita.

A Tabela 1 apresenta um sumário estatístico dos dados, tanto em sua forma original quanto após os filtros. No total, analisamos mais de 8 mil pares dia-grupo, com discussões que envolvem mais de 1,4 milhão de mensagens e mais de 67 mil usuários.

[1]Os autores gentilmente nos concederam acesso ao conjunto de dados, previamente anonimizado.

Tabela 1. Sumário da coleção de dados.

Período (25/09/2022 - 15/01/2023)	Dados originais	Dados filtrados e limpos
# Mensagens	1.854.157	1.442.312
# Pares dias-grupo	18.443	8.342
# Grupos	287	211
# Usuários distintos	75.043	67.462
Média diária de mensagens	100	129
Média de mensagens por grupo	6460	5117

3.2. Sumarização dos Dados

Vários trabalhos analisam discussões online aplicando técnicas de processamento de linguagem natural (PLN) - (e.g., análise de tópicos) diretamente na coleção de mensagens sob análise [Malagoli et al. 2024; Perlo et al. 2025]. Porém, a identificação de tópicos claros, representativos e não redundantes pode ser desafiadora, especialmente em ambientes mais abertos e sem moderação, como os grupos de Telegram, onde é comum observar várias mensagens que desviam do conteúdo central de uma discussão. Tais mensagens acabam por introduzir ruído e atrapalhar a detecção dos tópicos mais relevantes. Para lidar com esta questão, nós propomos utilizar grandes modelos de linguagem (*Large Language Models* – LLMs) para sumarizar as mensagens compartilhadas em um grupo em cada dia de análise. *Um sumário foi gerado para cada conjunto de mensagens correspondente a um par dia-grupo*. A premissa desta abordagem é que as LLMs são capazes de identificar os pontos principais em um conjunto de mensagens, sendo robustos a possíveis mensagens ruidosas. De fato, estudos recentes demonstram que as LLMs são muito promissoras para a tarefa de sumarização de texto [Sarmento and de Oliveira 2024].

A fim de escolher uma LLM, nós realizamos estudos preliminares com três modelos: Llama 3.1[2] e gpt-4o-mini[3], amplamente usados na literatura, e Sabiá-2 [Pires et al. 2023], uma LLM para língua portuguesa. Nós avaliamos manualmente a qualidade dos sumários produzidos por cada LLM para vários conjuntos de mensagens (cada um representando um par dia-grupo). Para cada LLM, experimentamos com diferentes prompts, inclusive impondo restrições no número máximo de palavras no sumário a fim de gerar sumários informativos, mas que não sejam muito extensos. Nossos resultados sugerem um melhor desempenho do modelo da OpenAI (gpt-4o-mini), com o seguinte prompt: *"Você receberá mensagens de conversas do Telegram, e sua tarefa é gerar um sumário conciso de no máximo 150 palavras das mensagens"*. Os resultados apresentados na Seção 4 foram gerados por este modelo.

3.3. Embedding

A representação (*embedding*) de textos é uma técnica usada em PLN para representar um texto de entrada como um vetor em um espaço de alta dimensão, a fim de capturar suas características semânticas e textuais. Esta representação permite calcular a similaridade dos textos a partir da distância entre suas representações no espaço de *embeddings*. Existem diferentes técnicas de representação de conteúdo textual por *embeddings*. Nós aqui testamos dois modelos: (a) BERTimbau [Souza et al. 2020], modelo BERT (i.e., baseado na arquitetura de Transformers) e pré-treinado para a língua portuguesa, que produz

[2]https://www.llama.com/

[3]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

representações vetoriais em um espaço de 1024 dimensões; e (b) text-embedding-3-large[4], modelo pertencente à OpenAI[5], que produz vetores em um espaço de 3072 dimensões.

Nós utilizamos ambos métodos para produzir representações para cada sumário, gerado para cada par dia-grupo analisado, normalizando os vetores ao final utilizando a norma L2. Na Seção 4, nós discutimos os resultados para os dois métodos de *embeddings*.

3.4. Agrupamento em Grandes Temas

Esta etapa consiste em aplicar um algoritmo de aprendizado não supervisionado, especificamente o algoritmo de *clustering* K-Means [Mining 2006], para agrupar os *embeddings* dos sumários em *clusters*, de forma que sumários semanticamente similares fiquem no mesmo *cluster*. Assim, cada *cluster* representa um grande tema de discussão identificado nos sumários. É conhecido que algoritmos de agrupamento sofrem em espaços de alta dimensionalidade. Assim, antes de aplicar o K-Means, utilizamos o algoritmo UMAP [McInnes et al. 2018] para reduzir a dimensionalidade dos *embeddings* para 10, valor comumente usado na literatura. Como será discutido na Seção 4, a escolha do melhor número de *clusters* foi feita avaliando duas métricas de qualidade de *clusters*, o Coeficiente de Silhueta [Mining 2006] e o Método do Cotovelo [Mining 2006].

3.5. Análise das Discussões

Uma vez agrupados (os *embeddings*) dos sumários associados a cada par dia-grupo, a última etapa consiste na análise destes *grandes temas* (i.e., *clusters*) a respeito das propriedades do conteúdo (i.e., sumários) associados a cada tema assim como da dinâmica dos dias-grupo do Telegram em torno de cada tema.

Em termos do conteúdo, nós analisamos cada *cluster* utilizando uma nuvem de palavras para representar os sumários correspondentes, ponderando as palavras pelo produto entre o TF (*Term Frequency*) e o IDF (*Inverse Document Frequency*), abordagem comumente usada que contempla conjuntamente os poderes descritivo (TF) e discriminativo (IDF) das palavras. Nós também analisamos as propriedades psicolinguísticas dos sumários em cada *cluster* utilizando o Linguistic Inquiry and Word Count Dictionary (LIWC) [Boyd et al. 2022]. O LIWC é um léxico que categoriza palavras em atributos derivados da gramática e psicologia. Exemplos de atributos incluem *relig*, *anger*, *sad*.

Cada *cluster* representa um grande tema de discussão, que por si só pode ser caracterizado em termos de diferentes tópicos mais específicos. Assim, a fim de melhor caracterizar as discussões, nós utilizamos um método de modelagem de tópicos para identificar os tópicos mais relevantes associados aos sumários da base inteira (em todos os *clusters*). A intenção é prover uma análise hierárquica, em que os grandes temas (*clusters*) são apresentados em um primeiro nível, enquanto os tópicos específicos associados a cada tema são caracterizados em um segundo nível.

A modelagem de tópicos foi feita utilizando o BERTopic [Grootendorst 2022], que combina um modelo de representação vetorial em alta dimensão com uma redução de dimensionalidade seguida de agrupamento, permitindo extrair os principais tópicos dos sumários. Configuramos a ferramenta com o BERTimbau como forma de representação

[4]https://platform.openai.com/docs/guides/embeddings

[5]https://openai.com/

Tabela 2. Coeficiente de Silhueta para diferentes números de clusters.

# Clusters	2	3	4	5	6
OpenAI (text-embedding-3-large)	0,45	0,36	0,44	0,37	0,42
BERTimbau	0,31	0,32	0,32	0,32	0,32

vetorial, eficaz na captura semântica do português. Reduzimos a dimensionalidade com UMAP e aplicamos HDBSCAN para identificar agrupamentos. Testamos também o K-Means ao invés do HDBSCAN, mas como não tínhamos um número fixo de *clusters* (tópicos) bem distribuídos, optamos pelo HDBSCAN que não necessita dessa parametrização. A relevância dos tópicos encontrados foi determinada via c-TF-IDF, extraindo palavras-chave de cada documento. Os hiperparâmetros do BERTopic, UMAP e HDBSCAN foram escolhidos conforme a literatura e a documentação do *framework* [Grootendorst 2022; Venâncio et al. 2024b][6]. Por exemplo, para o UMAP, utilizamos número de componentes igual a 10, número de vizinhos igual a 2, a mínima distância entre os pontos no espaço como 0 e a métrica como similaridade de cosseno [Collell and Moens 2021]. Para o HDBSCAN, o tamanho mínimo dos agrupamentos foi 25 e a quantidade mínima de amostras foi 40. Além disso, o modelo utiliza o *embedding* previamente gerado durante a metodologia, garantindo a extração de tópicos relevantes dos sumários.

4. Análises e Resultados

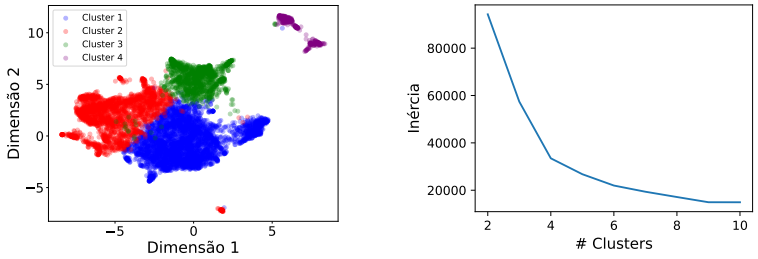
Nesta seção, apresentamos e discutimos os principais resultados obtidos com base na metodologia descrita anteriormente.

4.1. Análise e Validação dos Agrupamentos

Para validar os agrupamentos gerados, comparamos os resultados obtidos a partir dos dois modelos de *embedding* utilizados — BERTimbau e *text-embedding-3-large* (modelo da OpenAI) — utilizando o Coeficiente de Silhueta como métrica de avaliação. Inicialmente, foram identificados nove *clusters* principais. Após a remoção de outliers, o processo de *clusterização* foi refeito, resultando nos valores apresentados na Tabela 2. Os valores da métrica para diferentes quantidades de agrupamentos possibilitam uma comparação direta entre os dois modelos. Observa-se que o modelo da OpenAI apresenta coeficientes de silhueta consistentemente superiores aos do BERTimbau, indicando uma separação mais clara entre os grupos gerados.

A Figura 2(a) apresenta a visualização da *clusterização* final, com redução de dimensionalidade aplicada ao *embedding* gerado pelo modelo da OpenAI com as duas primeiras dimensões do UMAP. Já a Figura 2(b) complementa a análise por meio da métrica inércia, sugerindo a existência de um platô em torno de quatro agrupamentos, conforme discutido na metodologia. Com base na combinação dessas evidências, isto é, valores das métricas de silhueta e inércia, bem como a inspeção visual do agrupamento, definimos o número ideal de *clusters* como sendo igual a 4, representando os principais temas das discussões analisadas.

[6] `https://maartengr.github.io/BERTopic/getting_started/quickstart/quickstart.html`



(a) Clusters formados (b) Análise do número de clusters

Figura 2. Representação 2D do agrupamento final com UMAP.

(a) Cluster 1 (b) Cluster 2 (c) Cluster 3 (d) Cluster 4

Figura 3. Nuvem de palavras dos clusters obtidos.

4.2. Identificação de Grandes Temas de Discussões

Para compreender o conteúdo dos agrupamentos gerados, utilizamos nuvens de palavras para representar visualmente os principais termos presentes nos sumários de cada *cluster*, conforme ilustrado na Figura 3.

As nuvens de palavras oferecem uma visão geral dos temas predominantes em cada agrupamento. No *Cluster* 1, destacam-se termos como *suposto*, *ministro*, *patriótico*, *manifestações*, *quartéis*, *intervenções* e *militares*, sugerindo discussões associadas a temas institucionais, cívico-militares e ações políticas. O *Cluster* 2 apresenta palavras como *ministro*, *esquerda*, *eleitores*, *mídia*, *manipulação*, *alegações* e *decisões*, indicando conteúdos relacionados a processos eleitorais, disputas políticas e narrativas sobre interferência institucional.

No *Cluster* 3, observam-se termos como *suposto*, *elite*, *ataque*, *teorias*, *manipulação*, *globais*, *China*, *Rússia*, *mortes* e *conspiração*, o que aponta para a presença de discussões envolvendo temas geopolíticos, narrativas conspiratórias e alegações sobre eventos globais. Já o *Cluster* 4 é caracterizado por palavras como *comunhão*, *camaradagem*, *salmo*, *acolhedor*, *rotina*, *descontraído*, *brincadeiras* e *preferências*, refletindo interações de cunho religioso, afetivo e descontraído entre os participantes.

Visando entender características psicolinguísticas associadas aos grupos de discussão, aplicamos a ferramenta LIWC nos sumários, classificando-os em categorias relacionadas a processos psicológicos, sociais e linguísticos. Para destacar os atributos mais discriminativos entre os *clusters*, ranqueamos as categorias com base no Coeficiente de Gini e selecionamos as 16 mais relevantes [Yitzhaki 1979]. A Figura 4 apresenta um mapa de calor com os valores normalizados (z-score) para cada categoria por *cluster*. As colunas representam as categorias do LIWC, e as linhas, os quatro *clusters* obtidos anteriormente. Cada célula mostra a intensidade relativa de determinada categoria naquele *cluster* em comparação com os demais. Valores positivos indicam maior ocorrência relativa daquela categoria no *cluster* correspondente, enquanto valores negativos indicam menor presença.

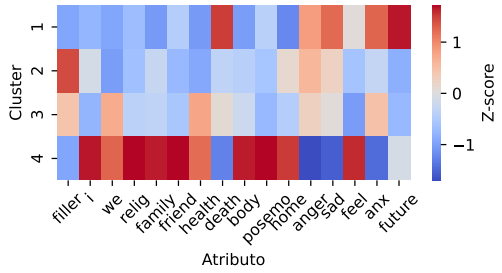


Figura 4. Distribuição dos atributos psicolinguísticos nos clusters.

No *Cluster* 1, observamos destaque para as categorias *death* (morte), *anger* (raiva), *sad* (tristeza), *anx* (ansiedade) e *future* (futuro), sugerindo um conteúdo com forte carga emocional negativa, associado a preocupações e tensões políticas ou sociais. Este padrão está alinhado à nuvem de palavras previamente analisada, que continha termos como *suposto*, *intervenções* e *manifestações*. O *Cluster* 2 apresenta valores elevados principalmente na categoria *filler*, relacionada ao uso de palavras ou expressões que preenchem pausas em conversas informais. Isso sugere a presença de trocas discursivas mais espontâneas ou interativas, compatíveis com o conteúdo político e eleitoral observado nas nuvens de palavras desse *cluster*. Já o *Cluster* 3 não apresenta elevações expressivas em nenhuma das categorias analisadas, indicando uma distribuição mais homogênea dos atributos linguísticos, sem predominância clara de um perfil psicológico específico. O *Cluster* 4, por sua vez, mostra alta concentração em categorias como *relig* (religião), *family* (família), *friend* (amizade), *posemo* (emoções positivas), *home* (lar) e *feel* (sentimentos). Esses atributos reforçam o caráter afetivo, espiritual e acolhedor das mensagens nesse agrupamento, conforme também sugerido pela nuvem de palavras correspondente.

4.3. Caracterização da Dinâmica e Evolução das Discussões

Após a caracterização dos agrupamentos em grandes temas, analisamos sua evolução temporal ao longo do período estudado. A Figura 5 apresenta dois aspectos dessa dinâmica. A Figura 5(a) mostra a distribuição semanal dos pares dia-grupo entre os quatro *clusters* identificados. Observa-se uma mudança marcante a partir da semana de 30/10/2022 (semana do 2º turno das eleições presidenciais), com uma diminuição acentuada da proporção de pares classificados no *Cluster* 2 e um aumento correspondente no *Cluster* 1. Esse padrão sugere uma possível mudança de foco das discussões, saindo de temas ligados ao processo eleitoral (*Cluster* 2) e migrando para conteúdos mais reativos ou mobilizadores, associados ao *Cluster* 1. Os *Clusters* 3 e 4 apresentam menor variação temporal. O *Cluster* 3 mostra presença moderada e relativamente constante ao longo do período, enquanto o *Cluster* 4, associado a conteúdo mais afetivo e religioso, mantém uma participação estável, porém com volume reduzido em relação aos demais.

A Figura 5(b) apresenta a matriz de transição temporal entre os *clusters*, considerando dias consecutivos. Cada grupo de quatro linhas da matriz representa as probabilidades de transições com destino fixo em um dos *clusters* (1 a 4). Essas transições são calculadas a partir da análise da permanência ou mudança de dias-grupo entre os clusters em certo dia comparando-os com o dia subsequente. Nota-se, que antes do segundo turno, há maior probabilidade de migração para o *Cluster* 2, enquanto no período posterior às eleições observa-se maior intensidade de transições para o *Cluster* 1, sugerindo

(a) Distribuição dos pares dia-grupo (b) Matriz de transição temporal entre clusters

Figura 5. Análises temporais das discussões.

uma reorganização das discussões em torno de temas associados a esse agrupamento e reforçando a dinâmica temporal observada na Figura 5(a).

A seguir, focamos na análise de tópicos a fim de revelar melhor o conteúdo e dinâmica dos temas obtidos a partir de cada agrupamento, como descrito na seção 3. A Tabela 3 mostra os cinco principais tópicos obtidos por *cluster*, totalizando 17 tópicos finais. Ao fazermos essa seleção, perdemos apenas 10,6% dos pares dias-grupo que temos, capturando o contexto geral desejado dos tópicos.

Diversos tópicos estão relacionados aos eventos políticos ocorridos na época, destacando-se o tópico 1 em maior quantidade, com sumários referentes a posse, manifestação e tomada de poder, fortemente relacionados ao movimento que culminou na invasão de edifícios federais no dia 8 de Janeiro. O tópico 6, por sua vez, reflete a situação em Brasília depois dos ataques, em que golpistas detidos associam detenção pela Polícia Federal a campos de concentração desumanos.

A Tabela 3 também apresenta a distribuição dos tópicos por *cluster*. No *cluster* 1, o tópico 1 mencionado predomina, representando mais de 60% dos sumários por dia-grupo, indicando um conteúdo fortemente relacionado a manifestações e infiltrações em Brasília. O *cluster* 2 é dominado pelo tópico 2 (43% dos sumários) que trata de campanhas eleitorais e votações políticas, embora contenha também cerca de 18% dos sumários representados pelo tópico 1, sem, no entanto, ser predominante. Isso evidencia ainda mais que o *cluster* 1 possui um conteúdo incisivo referente a manifestações e posse em Brasília, envolvendo uma tensão política maior, enquanto o segundo compreende uma dinâmica eleitoral e discussão política entre os participantes. O *cluster* 3, por sua vez, possui uma incidência de tópicos relacionados a vacinas, saúde, políticas mundiais e teorias da conspiração, complementando a nuvem de palavras discutida. O *cluster* 4 predomina os tópicos referentes à religião e experiência entre os participantes.

A Figura 6 apresenta a evolução temporal dos 17 tópicos identificados ao longo das semanas analisadas. A frequência dos tópicos foi normalizada por z-score por linha (tópico), ou seja, para cada tópico, foi calculada a média e o desvio padrão de sua frequência semanal. Assim, células em vermelho indicam semanas em que aquele tópico apareceu com frequência acima da média, enquanto tons de azul indicam frequência abaixo da média.

Tabela 3. Descrição dos tópicos e predominância em cada cluster.

ID	Palavras Características	% Cluster 1	% Cluster 2	% Cluster 3	% Cluster 4
1	posse, infiltrações, permanecer, determinação	65%	18%	0	0
2	campanha, eleitores, voto, votação, turno	4%	43%	0	0
3	efeitos, vacina, colaterais, mortes, pandemia	0	0	18%	0
4	Deus, bíblicos, versículos, encorajamento	0	0	0	63%
5	youtube, especulando, fatos, natural, artigo	3%	0	0	0
6	detidos, concentração, humanos, campos	5%	0	0	0
7	globais, ucrânia, despertar, conscientização	0	0	13%	0
8	bitcoins, criptomoedas, serviços, privacidade	0	6%	0	1%
9	economia, repressão, econômicas, boicote	3%	0	0	0
10	rússia, ucrânia, putin, energia, mortes, ataque, europa	0	0	9%	0
11	damares, regime, américa, latina, drogas	0	0	0	0.2%
12	experiências, interações, humor, leve, piadas	0	0	0	34%
13	maçonaria, efeitos, indústria, ivermectina	0	0	10%	0
14	provocações, ironia, humor, piadas, ofensas	0	0	0	1%
15	bilhões, milhões, infraestrutura, obras	0	5%	0	0
16	ambos, lados, ironia, gastos, comparações	0	5%	0	0
17	terra, tecnologia, históricos, gates	0	0	9%	0

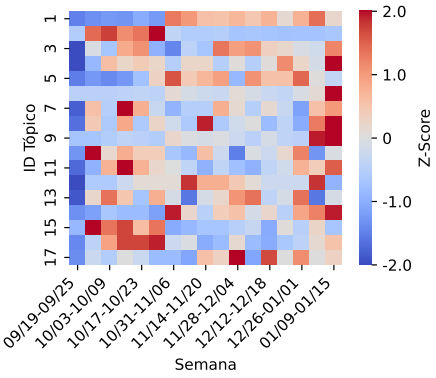


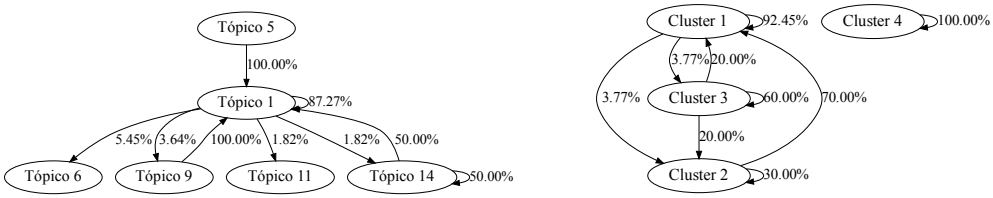
Figura 6. Evolução temporal dos tópicos identificados.

É possível observar que o Tópico 1, relacionado à *posse* e *infiltrações*, teve destaque especialmente nas semanas seguintes ao segundo turno das eleições (30/10/2022), reforçando sua associação com as manifestações políticas do período. O Tópico 6, por sua vez, relacionado a *detidos* e *campos*, tem um pico claro na semana de 08/01/2023, refletindo a repercussão das prisões ocorridas após os atos antidemocráticos em Brasília. Outros exemplos relevantes incluem o Tópico 3 (relacionado a *vacinas* e *efeitos colaterais*), que mostra variações pontuais ao longo do tempo, e o Tópico 12 (conteúdos afetivos e humorísticos), com picos isolados que refletem interações menos políticas em momentos específicos. Esses resultados evidenciam a dinâmica dos temas abordados nos grupos ao longo do tempo e reforçam a relação entre os tópicos identificados e eventos políticos e sociais de alta relevância no contexto das eleições e seus desdobramentos.

Para analisar como as discussões evoluíram durante o evento crítico de 8 de janeiro de 2023, investigamos as transições de tópicos e de *clusters* nesse intervalo. A Figura 7 apresenta dois grafos: (a) mostra a transição de tópicos dentro do *Cluster* 1 entre os dias 08/01 e 09/01, e (b) mostra as transições entre *clusters* no intervalo de 07/01 para 08/01. No *Cluster* 1, observa-se que a maioria dos dias-grupo manteve-se no Tópico 1 (87,27%),

fortemente associado a manifestações e ocupações em Brasília. Pequenas proporções migraram para tópicos como o Tópico 6 (5,45%), relacionado a *detidos* e *campos*, e Tópico 9 (3,64%), ligado a economia e repressão. Há também registros pontuais de transição para os Tópicos 11 e 14.

Simultaneamente, como mostrado no grafo de transição de *clusters* (Figura 7b), verifica-se que 70% dos dias-grupo que estavam no *Cluster* 2 em 07/01 migraram para o *Cluster* 1 em 08/01. Essa movimentação reforça a predominância do Tópico 1 no *Cluster* 1, uma vez que os grupos envolvidos nas discussões passaram a se concentrar em torno dos eventos do 8 de Janeiro. De modo geral, esses grafos ilustram como os temas mais mobilizadores, especialmente aqueles ligados a manifestações em Brasília, passaram a dominar as discussões nos grupos públicos do Telegram nesse período específico. Essa visualização complementa as análises anteriores ao revelar o comportamento dinâmico das discussões frente a eventos políticos marcantes.



(a) Transição de tópicos no Cluster 1 entre 08/01 e 09/01.

(b) Transição de dias-grupo entre os clusters entre 07/01 e 08/01.

Figura 7. Transições observadas no período de 8 de janeiro de 2023.

5. Conclusões e Trabalhos Futuros

Neste trabalho, propusemos uma metodologia para a análise da dinâmica de discussões em grupos públicos de mensagens, com foco na evolução temporal de temas e tópicos debatidos online. Aplicamos essa abordagem a um conjunto de mais de 1,8 milhão de mensagens oriundas de grupos públicos do Telegram e combinando estratégias de sumarização com LLMs, representação vetorial por *embeddings*, agrupamento temático e análise psicolinguística, conseguimos identificar grandes temas de discussão, seus principais tópicos e sua evolução ao longo do tempo. Os resultados revelaram fortes correlações entre os eventos políticos marcantes — como o segundo turno das eleições e os atos de 8 de janeiro — e a reorganização das discussões nos grupos analisados, evidenciando o potencial da abordagem para entender processos de mobilização discursiva em ambientes digitais.

Como trabalhos futuros, pretendemos: (i) estender a metodologia para outras plataformas com características distintas, como WhatsApp e X (Twitter), visando comparar padrões discursivos entre redes; e (ii) explorar aplicações em tempo real da metodologia, com foco no monitoramento contínuo da dinâmica informacional em cenários sensíveis. Para fins de replicabilidade do projeto e acesso à base de dados, favor entrar em contato com os autores da pesquisa.

Agradecimentos. Os autores agradecem ao CNPq, à CAPES, à Fundação de Amparo à FAPEMIG e ao Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional e Tratamento e Disseminação de Informação (INCT-TILD-IAR) pelo apoio financeiro e suporte à realização desta pesquisa.

Referências

[Alrhmoun et al. 2023] Alrhmoun, A., Winter, C., and Kertesz, J. (2023). Automating terror: The role and impact of telegram bots in the islamic state's online ecosystem. *Terrorism and Political Violence*.

[Anglano et al. 2017] Anglano, C., Canonico, M., and Guazzone, M. (2017). Forensic analysis of telegram messenger on android smartphones. *Digital Investigation*, 23:31–49.

[Araujo et al. 2023] Araujo, M. M., Ferreira, C. H., Reis, J. C., Silva, A. P., and Almeida, J. M. (2023). Identificação e caracterização de campanhas de propagandas eleitorais antecipadas brasileiras no twitter. In *Brazilian Workshop on Social Network Analysis and Mining (BrasNAM)*, pages 67–78. SBC.

[Bär et al. 2023] Bär, D., Pröllochs, N., and Feuerriegel, S. (2023). New threats to society from free-speech social media platforms. *Communications of the ACM*, 66(10):37–40.

[Bovet and Grindrod 2022] Bovet, A. and Grindrod, P. (2022). Organization and evolution of the uk far-right network on telegram. *Applied Network Science*.

[Boyd et al. 2022] Boyd, R. L., Ashokkumar, A., Seraj, S., and Pennebaker, J. W. (2022). The development and psychometric properties of liwc-22. Technical report, University of Texas at Austin, Austin, TX.

[Chagas 2022] Chagas, V. (2022). Whatsapp and digital astroturfing: A social network analysis of brazilian political discussion groups of bolsonaro's supporters. In *International Journal of Communication*.

[Collell and Moens 2021] Collell, G. and Moens, M.-F. (2021). How do simple transformations of text and image features impact cosine-based semantic match? In Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., and Sebastiani, F., editors, *Advances in Information Retrieval*.

[Ferreira et al. 2021] Ferreira, C. H. G., Murai, F., Silva, A. P. C., Almeida, J. M., Trevisan, M., Vassio, L., Mellia, M., and Drago, I. (2021). On the dynamics of political discussions on instagram: A network perspective. *Online Social Networks and Media*, 25.

[G. Da Fonseca et al. 2024] G. Da Fonseca, L. G., Gomes Ferreira, C. H., and Soares Dos Reis, J. C. (2024). The role of news source certification in shaping tweet content: Textual and dissemination patterns in brazil's 2022 elections. In *Proceedings of the 20th Brazilian Symposium on Information Systems*, pages 1–10.

[Grootendorst 2022] Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[Guimarães et al. 2022] Guimarães, S., Silva, M., Caetano, J., Araújo, M., Santos, J., Reis, J. C. S., Silva, A. P. C., Benevenuto, F., and Almeida, J. M. (2022). Análise de propagandas eleitorais antecipadas no twitter. *Brazilian Workshop on Social Network Analysis and Mining*.

[Hoseini et al. 2024] Hoseini, M., de Freitas Melo, P., Benevenuto, F., Feldmann, A., and Zannettou, S. (2024). Characterizing information propagation in fringe communities on telegram. In *International AAAI Conference on Web and Social Media*.

[Júnior et al. 2022] Júnior, M., Melo, P., Kansaon, D., Mafra, V., Sá, K., and Benevenuto, F. (2022). Telegram monitor: Monitoring brazilian political groups and channels on telegram. *arXiv:2202.04737 [cs.SI]*.

[Kloo et al. 2024] Kloo, I., Cruickshank, I. J., and Carley, K. M. (2024). A cross-platform topic analysis of the nazi narrative on twitter and telegram during the 2022 russian invasion of ukraine. In *International AAAI Conference on Web and Social Media*.

[Malagoli et al. 2024] Malagoli, L., Piorino, G., Ferreira, C. H., and da Silva, A. P. C. (2024). Twitter and the 2022 brazilian elections portrait: A network and content-driven analysis. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 283–291. SBC.

[Malagoli et al. 2021] Malagoli, L., Stancioli, J., Ferreira, C. H., Vasconcelos, M., da Silva, A. P. C., and Almeida, J. (2021). Caracterizaçao do debate no twitter sobre a vacinaçao contra a covid-19 no brasil. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 55–66. SBC.

[Maros et al. 2020] Maros, A., Almeida, J. M., Benevenuto, F., and Vasconcelos, M. (2020). Analyzing the use of audio messages in whatsapp groups. *Proceedings of The Web Conference 2020*.

[McInnes et al. 2018] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

[Mining 2006] Mining, W. I. D. (2006). Data mining: Concepts and techniques. *Morgan Kaufinann*.

[Naseri and Zamani 2019] Naseri, M. and Zamani, H. (2019). Analyzing and predicting news popularity in an instant messaging service. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[Nikkhah et al. 2018] Nikkhah, S., Miller, A. D., and Young, A. L. (2018). Telegram as an immigration management tool. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 345–348.

[Nobre et al. 2020] Nobre, G., G. Ferreira, C., and Almeida, J. (2020). Beyond groups: Uncovering dynamic communities on the whatsapp network of information dissemination. In *International Conference on Social Informatics*.

[Nobre et al. 2022] Nobre, G. P., Ferreira, C. H., and Almeida, J. M. (2022). A hierarchical network-oriented analysis of user participation in misinformation spread on whatsapp. *Information Processing & Management*.

[Paoletti et al. 2025] Paoletti, G., Almeida, J. M., Vassio, L., Gonçalves, M. A., and Mellia, M. (2025). Join the chat: How curiosity sparks participation in telegram groups. International AAAI Conference on Web and Social Media.

[Perlo et al. 2025] Perlo, A., Paoletti, G., Jha, N., Vassio, L., Almeida, J., and Mellia, M. (2025). A topic-wise exploration of the telegram group-verse. *International AAAI Conference on Web and Social Media*.

[Pinto et al. 2024] Pinto, S. L., Campolina, J. J., Sena, J. P. M., Félix, G., Ferreira, L. N., and Reis, J. C. S. (2024). Caracterização e predição de usuários tóxicos no twitter/x durante as eleições brasileiras de 2022. *Brazilian Workshop on Social Network ANalysis and Mining*.

[Pires et al. 2023] Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*.

[Resende et al. 2019] Resende, G., CS Reis, J., Vasconcelos, M., Almeida, J. M., and Benevenuto, F. (2019). Analyzing textual (mis) information shared in whatsapp groups. In *ACM conference on web science*.

[Sarmento and de Oliveira 2024] Sarmento, M. and de Oliveira, H. (2024). Sumarização automática de artigos de notícias em português: Da extração à abstração com abordagens clássicas e modelos de neurais. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

[Thorgren et al. 2024] Thorgren, E., Mohammadinodooshan, A., and Carlsson, N. (2024). Temporal dynamics of user engagement on instagram. In *ACM Web Science Conference*.

[Venâncio et al. 2024a] Venâncio, O. R., Ferreira, C. H., Almeida, J. M., and da Silva, A. P. C. (2024a). Unraveling user coordination on telegram: A comprehensive analysis of political mobilization during the 2022 brazilian presidential election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1545–1556.

[Venâncio et al. 2024b] Venâncio, O. R., Gonçalves, G. H. S., Ferreira, C. H. G., and da Silva, A. P. C. (2024b). Evidências de disseminação de conteúdo no telegram durante o ataque aos órgãos públicos brasileiros em 2023. In *Brazilian Symposium on Multimedia and Web*.

[Wich et al. 2022] Wich, M., Gorniak, A., Eder, T., Bartmann, D., Çakici, B. E., and Groh, G. (2022). Introducing an abusive language classification framework for telegram to investigate the german hater community. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[Xu and Livshits 2019] Xu, J. and Livshits, B. (2019). The anatomy of a cryptocurrency Pump-and-Dump scheme. In *28th USENIX Security Symposium (USENIX Security 19)*.

[Yitzhaki 1979] Yitzhaki, S. (1979). Relative deprivation and the gini coefficient. *The quarterly journal of economics*, pages 321–324.