UNIVERSIDADE FEDERAL DO CEARÁ

INSTITUTO UFC VIRTUAL

CURSO DE GRADUAÇÃO EM SISTEMAS E MÍDIAS DIGITAIS

MORGANA ARAÚJO DOS SANTOS

UM ESTUDO SOBRE A REPERCUSSÃO DA ELEIÇÃO PRESIDENCIAL BRASILEIRA DE 2022 NO TWITTER USANDO BERTOPIC

FORTALEZA

2022

MORGANA ARAÚJO DOS SANTOS

UM ESTUDO SOBRE A REPERCUSSÃO DA ELEIÇÃO PRESIDENCIAL BRASILEIRA DE 2022 NO TWITTER USANDO BERTOPIC

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas e Mídias Digitais do Instituto UFC Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas e Mídias Digitais.

Orientadora: Profa. Dra. Ticiana Linhares Coelho da Silva.

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S236e Santos, Morgana Araújo dos.
Um estudo sobre a repercussão da eleição presidencial brasileira de 2022 no Twitter usando BERTopic / Morgana Araújo dos Santos. – 2022.
43 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Instituto UFC Virtual, Curso de Sistemas e Mídias Digitais, Fortaleza, 2022.
Orientação: Profa. Dra. Ticiana Linhares Coelho da Silva.

1. Eleições Presidenciais. 2. Twitter. 3. Processamento de Linguagem Natural. 4. BERTopic. I. Título.
CDD 302.23

MORGANA ARAÚJO DOS SANTOS

UM ESTUDO SOBRE A REPERCUSSÃO DA ELEIÇÃO PRESIDENCIAL BRASILEIRA DE 2022 NO TWITTER USANDO BERTOPIC

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas e Mídias Digitais do Instituto UFC Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas e Mídias Digitais.

Aprovada em: 08/12/2022.

BANCA EXAMINADORA

Profa. Dra. Ticiana Linhares Coelho da Silva (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Alysson Diniz dos Santos
Universidade Federal do Ceará (UFC)

Profa. Dra. Lívia Almada Cruz
Universidade Federal do Ceará (UFC)

Aos que pensam em recomeçar, dá certo.

AGRADECIMENTOS

Começar outra faculdade de novo? Do zero? Certeza? Então, vamos lá! E todos eles apoiaram meu desafio. Todos os meus agradecimentos a eles:

Aos meus pais por sempre terem acreditado que eu podia chegar onde eu quisesse e por terem me proporcionado os ensinamentos pra trilhar esses caminhos. Amo vocês!

Ao meu companheiro, Alex, que é amor, calma, objetividade, compreensão, abraço apertado e alicerce. Obrigada pelo apoio em tudo e por não me deixar duvidar do que consigo alcançar.

As minhas tias, tios, primos que entre as conversas animadas na varanda da casa da minha avó sempre perguntaram sobre os estudos e afastaram meus receios de não me encontrar nesse novo caminho "das tecnologias". A minha avó Mundinha, que me recebe sempre com sorrisos e beijos mesmo eu estando mais ausente.

À família que a vida me permitiu escolher, meus amigos Aline Conde, Aline Ramos, Aurélio, Bianca, Cristiano, Elane, Juliane, Mariana e Matheus por todo o carinho, as comemorações, as conversas e os ouvidos atentos. Eu amo muito poder compartilhar a vida com vocês!

Aos meus amigos teens mais incríveis, divertidos e companheiros. Sem eles a ideia que eu tinha de turma de faculdade e de amigos pra vida toda não teria sentido. Obrigada, Thaís, Pedro, Mariana, Vicente, Vitor, Lívia, Lucas, Jefferson, Rodrigo, Felipe, Dimas, João Lucas e Gustavo.

Aos meus professores Alysson Diniz, por fazer do SMD um curso muito mais divertido e instigante que qualquer outro; e Ticiana Linhares, minha professora orientadora, pela calma, pela disponibilidade, pelas orientações valiosas, pelas aulas divertidas de APS e de testes. Obrigada por terem apresentado esse universo de possibilidades do SMD e por tornarem esse curso e minha passagem por ele muito melhor.

"E com o bucho mais cheio comecei a pensar
que eu me organizando posso desorganizar
que eu desorganizando posso me organizar
que eu me organizando posso desorganizar."
(Chico Science & Nação Zumbi, 1994)

RESUMO

Um dos temas mais repercutidos em toda a sociedade brasileira no ano de 2022 foi a eleição presidencial. Esta, que é tida por muitos como uma das mais importantes da história, movimentou as discussões em praticamente todos os meios de comunicação, sejam eles online ou não. Neste trabalho buscou-se identificar os assuntos mais discutidos na rede social Twitter sob a temática das eleições presidenciais. Para alcançar esse objetivo, foram aplicadas técnicas de Aprendizagem de Máquina e de Processamento de Linguagem Natural a um conjunto de 1,4 milhões de postagens em língua portuguesa sobre o tema das eleições e especificamente aos quatro principais candidatos ao Palácio do Planalto, no período de 28/09/2022 a 05/10/2022. Foram utilizados em conjunto algoritmos de *sentence embeddings*, clusterização via HDBSCAN e TF-IDF, com suporte do BERTopic, a fim de agrupar e extrair automaticamente os tópicos mais discutidos. Ao final do processo, os 12 tópicos mais relevantes foram analisados e comparados manualmente com as notícias apresentadas pelos meios de comunicação tradicionais nos mesmos dias. Dessa forma, pôde ser observado que o BERTopic apresentou resultados úteis ao entendimento das discussões do Twitter.

Palavras-chave: eleições presidenciais; assuntos; Twitter; linguagem natural; BERTopic.

ABSTRACT

One of the most reverberated themes throughout Brazilian society in 2022 was the presidential election. This election is considered by many to be one of the most important in history and moved discussions in virtually all media, whether online or not. In this work, we sought to identify the most discussed subjects under the theme of presidential elections on Twitter. To achieve this objective, Machine Learning and Natural Language Processing techniques were applied to a set of 1.4 million posts in Portuguese on the subject of the elections and specifically to the four main candidates for the *Palácio do Planalto*, in the period of 09/28/2022 to 10/05/2022. Techniques such as sentence embedding, clustering through HDBSCAN and TF-IDF, were used together with support from BERTopic, in order to group and automatically extract the most discussed topics. At the end of the process, the 12 most relevant topics were manually analyzed and compared with the news presented by the traditional media on the same days. Thus, it could be observed that BERTopic presented useful results for understanding Twitter discussions.

Keywords: presidential elections; subjects; Twitter; natural language; BERTopic.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE CÓDIGOS-FONTE

LISTA DE ABREVIATURAS E SIGLAS

DBSCAN	Density-Based Spatial Clustering of Applications With Noise
GPU	Graphics Processing Unit
HDBSCAN	Hierarchical Density-based Spatial Clustering of Applications with Noise
PLN	Processamento de Linguagem Natural
TF-IDF	Term Frequency - Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection

LISTA DE SÍMBOLOS

DF_w	Número de documentos que contém uma palavra
$TF_{w,d}$	Frequência de uma palavra em um documento
$TFIDF_{w,d}$	TF-IDF de uma palavra em um documento

SUMÁRIO

1 INTRODUÇÃO

As redes sociais são, além de um lugar para encontrar amigos e relatar as atividades diárias, um espaço de informação em que tanto os meios tradicionais de comunicação quanto as pessoas comuns podem compartilhar notícias e acontecimentos relevantes para a sociedade. Atualmente 53% da população mundial está conectada às redes sociais, sendo, apenas no Twitter, 436 milhões de usuários ativos. Assim muitos assuntos distintos são discutidos na rede que tem o potencial tanto de reverberar notícias quanto de produzi-las.

No período eleitoral, as redes sociais tornaram-se mais um espaço de campanha, pois grande parcela dos eleitores possui alguma rede social em que despendem parte do tempo no consumo de diversos conteúdos produzidos e compartilhados (VERGEER, 2015). Para os candidatos, a rede é um meio de difusão de suas propostas e um canal direto com seu eleitor (CERVI; MASSUCHIN, 2012).

No entanto, acompanhar os assuntos, enxergar a relevância e identificar a evolução deles nas redes é uma tarefa difícil, em razão da quantidade de dados gerados e transmitidos nesses veículos. Ademais, com a facilidade de acesso, há disseminação de notícias falsas, o que aumenta a dificuldade de acompanhar a evolução de fatos relevantes da eleição, pela dificuldade do usuário associar e correlacionar fatos e acontecimentos (VOSOUGHI *et al.*, 2018).

Há algum tempo pesquisadores de Ciência de Dados têm encontrado meios de extrair informações de um volume dados. As pesquisas desenvolvidas na área buscam tornar essa extração de informações dos dados automatizada, dessa forma o risco de perda de informações e mesmo a classificação dos dados seria estritamente impessoal, guardando relações mais precisas com a realidade dos fatos apresentados.

Algumas técnicas tornaram-se bastante populares, entre eles o conceito de clusterização que corresponde à formação de agrupamentos que guardam similaridades. Há também a formação desses clusters de maneira automatizada por meio do algoritmos como o K-Means (FORGY, 1965), *Density-Based Spatial Clustering of Applications With Noise (DBSCAN)* (ESTER *et al.*, 1996), o *Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCAN)* (CAMPELLO *et al.*, 2013), e tantos outros. Cada um desses algoritmos pode ajudar a entender grupos de documentos ou mesmo textos de redes sociais. Esses algoritmos podem demonstrar os novos assuntos presentes em um documento, os assuntos que se repetem ao longo dos dias, os assuntos desaparecem de um dia para outro, e dessa forma auxiliar no entendimento da evolução dos assuntos nas redes sociais.

É nesse contexto que surge o BERTopic, ferramenta que se utiliza do contexto em que as palavras estão inseridas para realizar a classificação e extração dos tópicos mais representativos. O BERTopic realiza a clusterização dos tópicos assim como outras ferramentas, porém adiciona uma complexidade quando considera o contexto, aproximando espacialmente palavras que possuam um contexto similar e por transitividade significado (GROOTENDORST, 2022).

Dado a necessidade de acompanhar a evolução dos assuntos no Twitter, o BERTopic se mostra uma ferramenta possivelmente adequada para explorar o conjunto de dados coletado na rede social. A ferramenta permite obter os tópicos mais discutidos e relevantes no conjunto de dados de forma automatizada, além de permitir o acompanhamento desses tópicos ao longo do tempo, inclusive por meio de visualizações.

O trabalho busca então apresentar a evolução dos assuntos durante o período eleitoral de 2022, com foco na eleição presidencial, utilizando o BERTopic como ferramenta. Com ele espera-se identificar os tópicos mais relevantes dos assuntos do período e validar o resultado por meio da identificação dos tópicos em outros veículos de comunicação da imprensa como os jornais, com a apresentação de notícias relacionadas.

1.1 Objetivos

De forma sucinta, o trabalho tem os objetivos listados abaixo.

1.1.1 Objetivo Geral

Apresentar a evolução dos tópicos discutidos no Twitter durante o período da eleição presidencial brasileira de 2022, por meio da extração automática de tópicos provida pelo BERTopic.

1.1.2 Objetivos Específicos

a) Visualizar os principais tópicos durante o período eleitoral;
b) Acompanhar a evolução dos tópicos mais relevantes durante o período eleitoral;
c) Estabelecer uma relação entre os tópicos mais relevantes e as notícias da mídia.

1.2 Estrutura do Trabalho

Este trabalho está organizado da seguinte forma: o Capítulo 2 expõe os conceitos fundamentais para o desenvolvimento da metodologia utilizada. O Capítulo 3 apresenta trabalhos recentes que se relacionam com o presente trabalho, seja quanto à metodologia ou tecnologias utilizadas. O Capítulo 4 detalha a metodologia empregada para a obtenção dos resultados, que são descritos no Capítulo 5. Finalmente o Capítulo 6 desenvolve conclusões sobre o trabalho e seus resultados e aponta direções de trabalhos futuros.

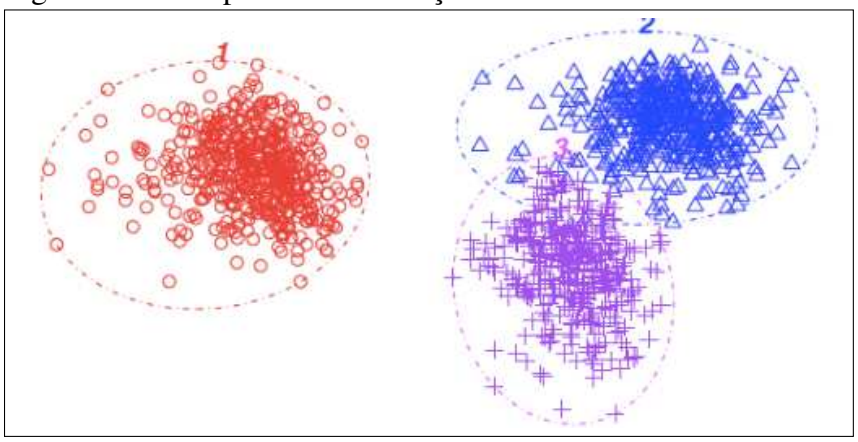
2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conceitos chave para o desenvolvimento desta pesquisa. Todos os conceitos utilizados são inerentes às áreas de Ciência de Dados, onde os algoritmos utilizados são amplamente aplicados em problemas relativos a Processamento de Linguagem Natural (PLN).

2.1 Clusterização

No campo da Ciência de Dados, a clusterização se mostra como uma ferramenta poderosa na análise de dados. A área ocupa papel de destaque, dado a relevância e o grande número de possíveis aplicações das técnicas desenvolvidas (EZUGWU *et al.*, 2022).

Figura 1 – Exemplo de clusterização realizada com o K-Means



Fonte: Columbia Public Health (2022).

A clusterização é o processo de agrupar um conjunto de dados que tenham alguma similaridade entre si, de forma automática. Os algoritmos para a execução dessa tarefa são classificados como algoritmos de aprendizagem de máquina não supervisionada, pois para o funcionamento desses algoritmos não é necessário que exista o indicativo de quais informações os algoritmos devem encontrar ou tomar como verdade. O processo de clusterização, nesse caso, leva em consideração os dados e suas características, que são exploradas à medida que os algoritmos evoluem.

Na Figura 1 é exemplificado o resultado de uma clusterização com o algoritmo K-Means, um dos vários disponíveis na literatura. Nele define-se o número de clusters que se deseja obter e, aleatoriamente, os centroides de cada cluster. A formação dos três clusters na Figura 1 se dá após a identificação das distâncias entre cada ponto e o centroide mais próximo. O

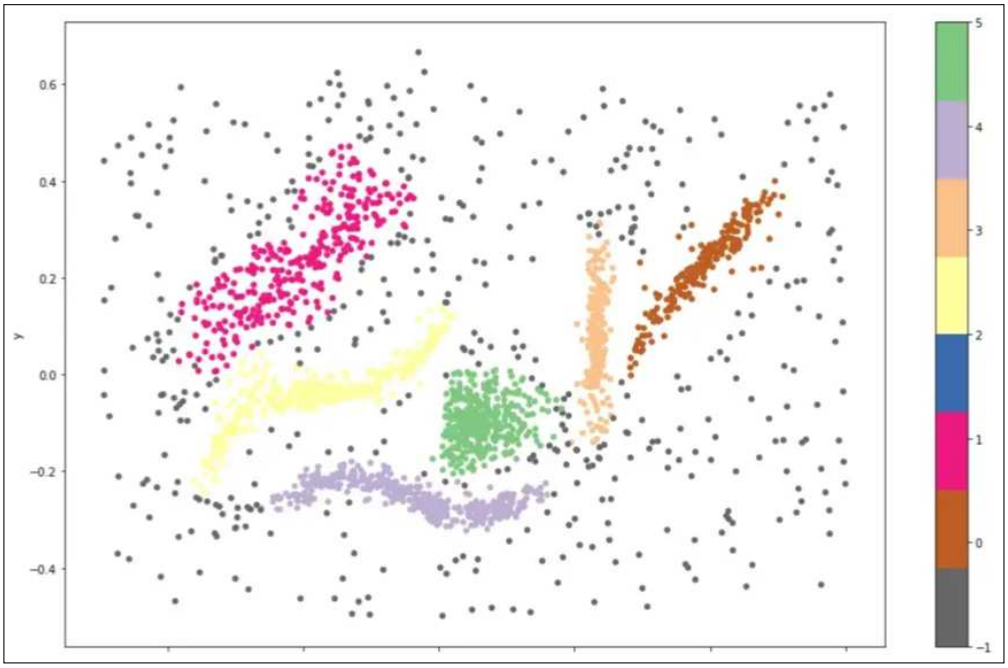
centroide pode ser reposicionado de modo que, ao final, corresponda a média de todos os pontos (FORGY, 1965).

Como todo processo na Ciência de Dados, os algoritmos de clusterização podem ter desempenhos diferentes de acordo com cada situação. Dessa forma o conjunto de dados que se deseja clusterizar pode fazer com que um algoritmo tenha um desempenho melhor ou pior do que outro, a depender das características dos dados (EZUGWU *et al.*, 2022).

2.2 HDBSCAN

O HDBSCAN é um algoritmo de clusterização que tem sua origem decorrente do DBSCAN. O HDBSCAN utiliza a densidade dos dados como um meio para identificar clusters de forma hierárquica (CAMPELLO *et al.*, 2013). A Figura 2 mostra como os clusters seguem a densidade dos dados como critério de formação dos clusters.

Figura 2 – Resultado de uma clusterização de dados feita pelo HDBSCAN



Fonte: Alberto (2020).

O método utiliza a ideia de densidade para encontrar clusters, portanto é necessário verificar os lugares em que os dados estão mais próximos uns dos outros, a região com maior densidade. Para isso, escolhe-se um ponto aleatoriamente e define-se um número mínimo necessário de dados que devem estar em seu raio de alcance. Todos os demais pontos serão avaliados com base na distância do ponto central do dado até o ponto mais distante localizado dentro da mesma região, essa medida é chamada *core distance* (CAMPELLO *et al.*, 2013).

O *core distance* será utilizado na descoberta da *mutual reachability distance*, medida que é calculada a partir de todas os valores de *core distance* dos pontos existentes ou da distância euclidiana entre dois pontos. O *mutual reachability distance* será o maior valor dentre esses demais, correspondendo a distância entre dois clusters (CAMPELLO *et al.*, 2013).

Com essa medida é possível chegar a uma árvore hierárquica de dados que será construída utilizando as menores distâncias observadas no conjunto de dados. Assim, pode-se visualizar o caminho em que os clusters se unificam com outros mais próximos e assim sucessivamente, até o momento em que exista apenas um conjunto estável de clusters. A árvore também pode ser pensada de forma contrária, um só cluster dividido entre vários outros (CAMPELLO *et al.*, 2013).

O HDBSCAN possui parâmetros que permitem ajustar a quantidade de clusters a serem obtidos dos dados. O *min_cluster_size*, por exemplo, permite delimitar a quantidade mínima de dados que compõem um cluster. Esse parâmetro permite definir a granularidade dos possíveis clusters encontrados no conjunto de dados, pois se o valor for definido como baixo, será permitido encontrar clusters de tamanho pequeno. Se o valor for alto, então há uma boa chance dos clusters menores serem agrupados em um único grande cluster.

2.3 TF-IDF

O *Term Frequency - Inverse Document Frequency (TF-IDF)* pode ser definido como uma medida que informa a importância de um termo em um documento ou em um conjunto de documentos (QAISER; ALI, 2018). Essa medida é feita a partir do produto da Frequência do Termo com a Frequência Inversa do Documento, descrita por

$$TFIDF_{w,d} = TF_{w,d} \times log(\frac{N}{DF_w}). \tag{2.1}$$

A Frequência do Termo (TF) é o número de vezes que uma palavra "w"aparece em um documento "d". Já a Frequência Inversa do Documento (IDF) é a identificação de um termo como raro ou comum em um conjunto de textos. Para calcular o IDF, é calculado o logaritmo da divisão do número total de documentos (N) pelo número de documentos que contém "w"(DF_W). Quando o termo é considerado comum o valor da equação tende a 0, quando raro, aproxima-se de 1.

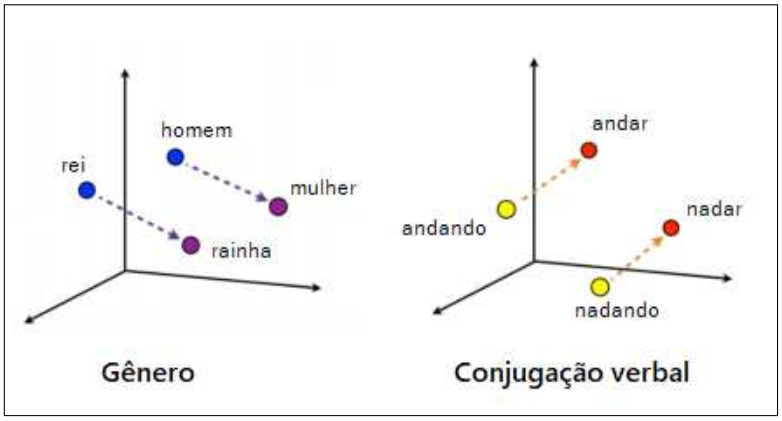
Assim, um termo que aparece bastante em um documento pode ser considerado importante, porém torna-se comum quando esse mesmo termo também é encontrado em outros documentos pertencentes ao conjunto. Quando a palavra é comum ao texto de todos os

documentos analisados, sinaliza que ele possivelmente não é relevante para a compreensão do documento.

2.4 Representação de Sentenças via Embeddings

As *word embeddings* têm sido largamente utilizadas em diferentes ferramentas como google translate, sugestão de palavras no teclado do celular, dentre outros. Essa é uma forma de representação das palavras em que um vetor possui múltiplas informações que auxiliam no estabelecimento do significado da palavra.

Figura 3 – Representação de *Word Embeddings*



Fonte: Fonseca (2021).

A ideia é que as *word embeddings* obtenham o significado das palavras utilizadas na linguagem natural de forma que consigam captar, identificar e reproduzir os sentido das palavras em seus contexto como ocorre na linguagem natural. As *word embedding* buscam um sentido semântico para uma palavra a partir das definições das palavras próximas no espaço vetorial.

A Figura 3 mostra duas *word embeddings* de gênero e conjugação verbal. Em ambos os casos, pode-se visualizar a relação semântica entre as palavras no espaço vetorial, onde palavras com significados mais próximos possuem menor distância.

A representação dessas palavras se dá por meio dos vetores. Os modelos de aprendizagem de máquina não reconhecem palavras, mas valores numéricos. Assim, para utilizar palavras é necessário realizar um processo de vetorização do texto que pode ser realizado por *embeddings*.

Para construir as *word embeddings* o texto é padronizado, removendo-se pontuação e tornando todos os caracteres minúsculos, e dividido em partes, processo conhecido como tokenização. Sobre as *word embeddings* descreve Oliveira *et al.* (2022, p. 32):

> *word embeddings* como um espaço vetorial semântico. Partindo do princípio de que palavras formam um espaço estruturado, isto é, que compartilham informações entre si, a hipótese distributiva (distributional hyphothesis) define que palavras que frequentemente aparecem em contextos semelhantes (que possuem um relacionamento semântico) também possuem significados semelhantes. Essa hipótese estabelece que capturar significado e capturar contextos são inerentemente a mesma coisa.

As *word embeddings* contém então informações sobre o contexto de uma palavra de forma mapeada em um vetor. Isso possibilita conferir significado às palavras e diferenciá-las umas das outras. Aqui se utiliza a distância euclidiana para estabelecer a relação entre os termos, logo, quanto maior a distância, mais distintas as palavras (OLIVEIRA *et al.*, 2022). As *word embeddings* podem ser classificadas como estáticas, onde as palavras apresentam o mesmo significado independente do contexto em que aparecem, e as contextuais, cujas palavras são representadas junto de seu contexto (OLIVEIRA *et al.*, 2022).

Assim como podemos representar palavras por meio de vetores, também é possível fazer o mesmo trabalho para sentenças inteiras. Esse processo, chamado de *sentence embedding* segue o mesmo princípio: representar sentenças num espaço vetorial n-dimensional de forma que sentenças semanticamente ou contextualmente similares tenham valores igualmente similares nos vetores (JUNIOR *et al.*, 2021). Dois algoritmos amplamente utilizados para a construção desse tipo de embedding são o *Universal Sentence Encoder* (CER *et al.*, 2018) e o *Doc2Vec* (LE; MIKOLOV, 2014). A Figura 4 exemplifica três *sentence embeddings* de duas frases relativamente similares e uma terceira com pouca similaridade com as demais.

Figura 4 – Representação de *Sentence Embeddings*

Fonte: Jindal (2019).

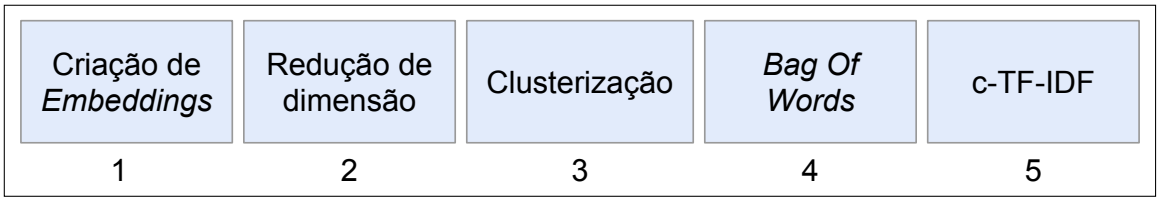
2.5 BERTopic

Como já abordado anteriormente, vive-se um momento em que muito cotidiano é convertido em algum dado e, a partir deles, pode-se obter informações diversas sobre as pessoas, o contexto em que estão inseridas e muitos outros. Grande parte desses dados encontra-se na forma escrita e muito deles perdem o sentido em razão da grande quantidade disponível e não utilizada.

A área da Ciência de Dados tem buscado formas de dar sentido aos dados subutilizados, proporcionando que se encontrem informações relevantes para diferentes grupos nos dados. Assim, diversos modelos de aprendizagem de máquina têm-se ocupado a compreender a linguagem natural e traduzi-la para que um computador possa ser capaz de compreender e até mesmo sugerir palavras.

É nesse contexto que o BERTopic (GROOTENDORST, 2022) se apresenta como um modelo de identificação de tópicos. Criado em 2022, o BERTopic utiliza alguns conceitos consolidados na aprendizagem de máquina como embeddings, clusterização e outros para extrair os tópicos mais relevantes de um corpo de texto. Para isso, o BERTopic realiza uma série de passos, descritos na Figura 5.

Figura 5 – Processos executados durante a extração de tópicos com o BERTopic



Fonte: A própria autora.

No primeiro passo, a geração de *sentence embeddings*, como descrito anteriormente é necessário realizar a conversão dos documentos em representações numéricas para que possam ser analisados. O BERTopic utiliza os *sentence-transfomers* que realizam a vetorização das sentenças, porém o usuário é livre para utilizar quaisquer modelos para representação vetorial. No presente trabalho, utilizou-se um *sentence-transfomer* já disponibilizado pela implementação padrão, o *paraphrase-multilingual-MiniLM-L12-v2*, que possui tradução para mais de 50 línguas (REIMERS; GUREVYCH, 2019).

Com a representação vetorial das palavras obtida no passo anterior, realiza-se uma diminuição da dimensão desses dados. Os modelos de cluster tem dificuldade em trabalhar com dados que possuem muitas dimensões, assim utiliza-se por padrão o Uniform Manifold

Approximation and Projection (UMAP), que é uma técnica de redução de dimensões que mantém estruturas locais e globais.

Com a redução das dimensões, utiliza-se o HDBSCAN que permite a identificação de clusters de diferentes formatos e também a identificação de pontos que não pertencem a nenhum cluster, os chamados *outliers*. O BERTopic usa o HDBSCAN como padrão, mas assim como nos embeddings é possível utilizar outro algoritmo de clusterização.

Após a clusterização, utiliza-se a técnica *bag-of-words* que consiste na extração das palavras mais relevantes de cada cluster por meio da sua frequência (OLIVEIRA *et al.*, 2022). Ressalta-se que as representações em *bag-of-words* consideram o tamanho de cada cluster existindo assim uma normalização.

De posse das palavras mais relevantes de cada cluster, verifica-se a ocorrência de cada uma nos diferentes clusters. Para isso, utiliza-se o TF-IDF que permite a comparação das palavras em um conjunto de documentos, conferindo a elas relevância, ou não, dentro de cada texto ou cluster. No BERTopic, o TF-IDF é modificado para que todos os documentos de uma classe sejam agrupados em apenas um documento, fazendo com que os termos mais importantes sejam extraídos com maior relevância para um determinado cluster. Essa técnica é chamada de c-TF-IDF.

3 TRABALHOS RELACIONADOS

Neste capítulo serão abordados brevemente os trabalhos relacionados a este estudo.

Em seu trabalho, (RODRIGUES, 2016) aborda a dificuldade dos usuários de redes sociais em acompanhar de maneira não automatizada o grande volume de conteúdos levantados nesses espaços. A autora apresenta formas de facilitar esse processo por meio do algoritmo de clusterização DBSCAN, com intuito de agrupar conjuntos de dados similares automaticamente.

A proposta de (RODRIGUES, 2016) consiste em identificar os assuntos populares repercutidos no Twitter, além de identificar a evolução dos clusters ao longo do tempo. Para isso, ela utilizou duas medidas de similaridade, Jaccard e Fading, realizando comparações para analisar qual das medidas proporcionaria melhores resultados no processo de clusterização.

Os assuntos publicados no Twitter também foram objeto de análise no trabalho de (LYU *et al.*, 2021). Aqui observou-se as discussões, os temas e os sentimentos dos usuários a partir das publicações sobre a pandemia de COVID-19. O estudo contribui na apresentação das preocupações da população com a eficácia e segurança da vacina.

Para isso (LYU *et al.*, 2021) utilizaram uma abordagem de modelagem de tópicos, em específico o algoritmo *Latent Dirichlet Allocation* (LDA). A técnica permitiu o agrupamento de 16 diferentes tópicos em 5 temas abrangentes, sendo o maior deles relativos às opiniões e as emoções sobre a vacinação. Os autores observaram a mudança dos tópicos com o avanço no desenvolvimento das vacinas, quando as instruções sobre tomar a vacina ocuparam o grupo de assuntos mais discutidos. O estudo indica que esse movimento foi impulsionado pelas notícias na mídia, tornando o sentimento a respeito da vacinação cada vez mais positivo.

Assim como no trabalho anterior, (SOUSA; BECKER, 2022) também analisam a movimentação no Twitter sobre a vacinação contra a COVID-19. No entanto, eles procuram fazer um comparativo entre os posicionamentos dos cidadãos norte-americanos e dos cidadãos brasileiros, além de tentarem garantir uma maior qualidade dos dados analisados. Diferente de (LYU *et al.*, 2021), (SOUSA; BECKER, 2022) utilizaram o BERTopic para identificação dos principais tópicos entre os grupos favoráveis e contrários à vacina.

Dessa forma, o presente trabalho aproxima-se de (RODRIGUES, 2016) quanto a proposição de uma forma automatizada de acompanhar o grande volume de assuntos discutidos no Twitter em determinado período, mas utilizando outra ferramenta para análise. Aqui o trabalho aproxima-se de (SOUSA; BECKER, 2022) quanto a utilização do BERTopic e análise dos resultados.

O Quadro 1 apresenta e sumariza algumas diferenças entre os trabalhos relacionados, comparando-os sob a perspectiva do contexto dos dados, do conjunto de dados e da estratégia aplicada.

Quadro 1 – Comparativo com os trabalhos relacionados

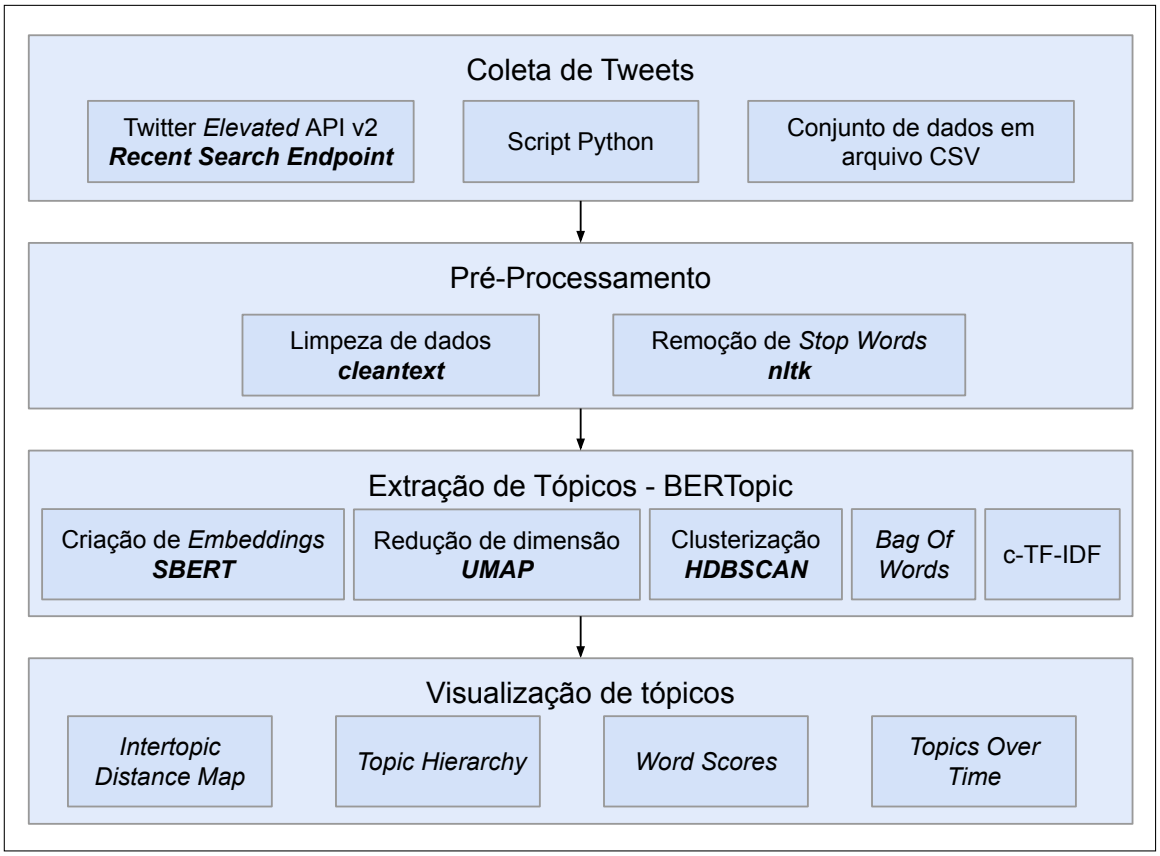
Trabalho	**Contexto**	**Conjunto de Dados**	**Estratégia**
Rodrigues (2016)	Impeachment de 2016 Movimento Gay	672.551 tweets 798.362 tweets	DBSCAN
Lyu *et al.* (2021)	Pandemia de COVID-19	1.499.421 tweets	LDA
Sousa e Becker (2022)	Vacinação da COVID-19	893.985 tweets dos EUA 139.131 tweets do Brasil	BERTopic
Este trabalho	Eleições presidenciais de 2022	1.249.987 tweets	BERTopic

Fonte: elaborado pela autora.

4 METODOLOGIA

O trabalho corresponde a uma pesquisa de natureza qualitativa, em razão da classificação e análise dos resultados obtidos no decorrer do trabalho. Como o objetivo é apresentar a evolução dos assuntos discutidos no Twitter no período eleitoral, a metodologia contará com um conjunto de 4 etapas, ilustradas na Figura 6 e descritas a seguir.

Figura 6 – Etapas da metodologia utilizada no trabalho



Fonte: a própria autora.

4.1 Coleta de Tweets

O Twitter é uma rede social que existe desde 2006. Diversas mudanças já ocorreram desde sua criação, mas o objetivo principal da rede permanece: disponibilizar um espaço para que as pessoas comentem sobre o que está acontecendo com elas ou com o mundo. E assim o microblog de 280 caracteres mantém relevância e destaque nas redes.

O Twitter possui uma plataforma para desenvolvedores que pode ser utilizada por quem deseja conhecer e desenvolver soluções com a ajuda dos recursos da rede. A plataforma possui três diferentes produtos: *Twitter API*, que possui um conjunto de *endpoints* utilizados para

compreensão ou construção de uma conversa no Twitter; *Twitter Ads API*, voltada para melhorar a experiência da visualização de anúncios; e *Twitter for Websites*, que permite a inclusão do Twitter no em páginas web com o objetivo de aumentar o engajamento de seguidores, entre outros.

Para este trabalho, somente o Twitter API será utilizado, pois permite a busca e análise de dados apresentados pelos usuários no Twitter. Essa ferramenta possui quatro diferentes níveis de acesso (*Essential, Elevated, Elevate+, Academic Research*), onde a cada nível são apresentadas mais possibilidades de busca, aumento do limite de requisições suportadas, dentre outros. O presente trabalho utilizou o nível *Elevated*, que é gratuito e é fornecido após o preenchimento de algumas questões realizadas pelo Twitter.

O acesso ao nível *Elevated* permite a coleta de 2 milhões de tweets por mês, sendo possível o desenvolvimento de um projeto por conta. A API utilizada foi a "API v2", recomendada pelo Twitter, que contém *features* e *endpoints* voltados para uma melhor experiência do desenvolvedor na coleta e análise de dados. Essa versão permite ao desenvolvedor especificar os campos que deseja obter nas respostas das requisições, como filtros de contexto, retorno de 100% dos conteúdos pesquisados disponíveis publicamente no Twitter e muitos outros.

Dentre as funcionalidades disponíveis na API v2, estão as *Annotations* que apresentam uma forma de compreender a informação do contexto existente em um tweet. O Twitter realiza filtragens por contas, hashtags, palavras-chave ou frases para realizar a compreensão e classificação do contexto. O Twitter faz uma distinção sobre *Entity Annotations* em que são apresentados todas as menções explícitas ao texto do tweet. Nessa categoria estão pessoas, lugares, produtos ou organizações que são diretamente referenciados nos tweets. Já o *Context Annotation* deriva de uma análise do contexto do tweet e aqui ela se utiliza de entidades estabelecidas para agrupar os tweets. O Twitter fornece uma lista das entidades de contexto disponíveis no GitHub. A partir delas, é possível fazer uma busca direcionada por um assunto específico.

Existem dois *endpoints* de coletas de dados no Twitter: o *Recent Search* e o *Full-Archive Search*. Esses *endpoints* permitem a pesquisa por tweets a partir de seus atributos, palavras-chave, *hashtags*, além de combiná-los para tornar a busca mais precisa. A diferença entre eles está na limitação temporal e na permissão de acesso.

O *Recent Search* busca todos os tweets publicados na última semana. Para acessá-los, é necessárias a chave do token do projeto criado. O *Recent Research* coleta 100 tweets por requisição. Já o *Full-Archive Search* está disponível apenas para os acessos acadêmicos e

enterprise, permitindo a coleta dos dados desde a data de início da rede social, onde podem ser coletados até 500 tweets a cada requisição. No trabalho, utilizou-se o *Recent Search* em razão do nível de acesso concedido pela rede social.

A ideia inicial para realizar a pesquisa foi a criação de um conjunto de dados para o posterior estudo dos dados coletados. Precisava-se de uma ferramenta que pudesse proporcionar tweets relativos aos assuntos eleitorais. A API do Twitter foi pensada como recurso para a coleta, uma vez que é uma ferramenta disponibilizada pela própria rede, possui características de pesquisa que contribuem para recortes de assuntos específicos de forma bem definida e de fácil acesso (como a presença de tutoriais exemplificando o passo a passo necessário). Assim, o conjunto de dados seria construído já com o aspecto eleitoral como temática.

O início do registro e arquivamento dos tweets ocorreu 15 dias antes da realização do primeiro turno (02 de outubro de 2022). Criou-se um script que, passada a chave de acesso, definiu-se os parâmetros utilizados na coleta dos tweets. Foram definidos dois domínios de busca dentre os apresentados pelo Twitter na *Feature Annotations* para categorizar os tweets, os domínios 35 (*Politicians*) e 38 (*Political Race*). A partir desses domínios, realizou-se uma busca pelos códigos das entidades do *Context Annotation* já mapeadas pelo Twitter. Assim, a consulta foi construída a partir dos filtros relativos às entidades expressas na Tabela 1.

Tabela 1 – Filtros aplicados na coleta de tweets

Entidade	**Descrição**	**Domínio**	**Identificador**
Luiz Inácio Lula da Silva	Político	35	862070591737675776
Ciro Gomes	Político	35	912370288968458240
Jair Bolsonaro	Político	35	912697101083041792
Simone Tebet	Político	35	1091083297654886400
Eleições Brasileiras de 2022	Corrida Política	38	1411038381841084418

Fonte: elaborada pela autora.

Além da definição da consulta, estabeleceu-se que dado o período de um dia, seriam coletados uma média de 80 tweets a cada minuto, desse modo seria possível acompanhar o assunto ao longo de um dia, caso fosse necessário. Também foi acrescida a restrição aos tweets em português e a não coleta de retweets, em razão da duplicação da informação. O Código-fonte 1 mostra como esse processo foi implementado.

Os tweets foram endereçados para arquivos *csv*, organizados em colunas de id, data de criação, autor e texto (representando o próprio tweet). Esse processo foi acionado a partir do dia 18 de outubro e continuou até o fim do período eleitoral no segundo turno, dia 30 de outubro de 2022.

Código-fonte 1 – Procedimento de consulta a API do Twitter, executado de minuto a minuto

```
query_params = {
    'query': 'context:38.1411038381841084418 (context
    :35.912697101083041792 OR context:35.912370288968458240 OR context
    :35.1091083297654886400 OR context:35.862070591737675776) -is:
    retweet lang:pt',
    'tweet.fields': 'author_id,created_at,text',
    'start_time': start,
    'end_time': end,
    "max_results": 80
}

twitter_api_response = requests.get("https://api.twitter.com/2/tweets
    /search/recent", auth=authentication, params=query_params)
```

4.2 Pré-Processamento

Enquanto ocorria a coleta para o conjunto de dados, iniciava-se a fase de pré-processamento dos dados já coletados. O pré-processamento é o momento em que pontuações, símbolos ou mesmo palavras são retiradas do conjunto de dados com a finalidade de deixar no corpo do texto apenas os dados que agreguem para a extração de uma informação. Assim, importou-se a uma biblioteca *clean-text*[1] do Python para a retirada da pontuação, urls e emojis dos tweets analisados.

Foram removidas pontuações e emojis, além de substituídas as URLs exibidas nos tweets por espaços vazios. Para isso, utilizou-se a biblioteca *clean-text* que apresenta uma série de argumentos que podem ser habilitados na manipulação do texto. O Código-fonte 2 mostra como esse processo foi implementado.

Código-fonte 2 – Código de limpeza dos Tweets

```
from cleantext import clean
dataset.tweet = dataset.apply(lambda row: clean(row.tweet, no_punct=
    True, no_urls=True, replace_with_url="", no_emoji=True), 1)
```

[1] Disponível em https://pypi.org/project/clean-text/0.6.0/

No pré-processamento ocorreu ainda a remoção de palavras consideradas "não úteis"à compreensão do texto, as chamadas *stop words*. Classificadas como *stop words* encontram-se as preposições que costuram palavras para a formação do texto, elas, no entanto, não adicionam sentido semântico ao contexto, podendo ser dispensadas. No trabalho foi utilizada a biblioteca *nltk*[2] que possui um conjunto de palavras pré-definidas na língua portuguesa que são consideradas *stop words*. O Código-fonte 3 mostra como esse processo foi implementado.

Código-fonte 3 – Código de limpeza das stop words

```
def remove_stop_words_from_tweet(stopwords, tweet):
  tweet_words = tweet.lower().split(" ")
  filtered_words = filter(lambda word: word not in stopwords,
   tweet_words)
  return " ".join(filtered_words)

import nltk
nltk.download('stopwords')
stopwords = nltk.corpus.stopwords.words('portuguese')
dataset.tweet = dataset.apply(lambda row:
   remove_stop_words_from_tweet(stopwords, row.tweet), 1)
```

4.3 Extração de tópicos

O conjunto de dados contendo o texto normalizado foi submetido à ferramenta BERTopic. A ideia era apresentar os tópicos encontrados ao longo de um período, por isso o BERTopic foi executado no conjunto de dados em diferentes períodos de tempo. Inicialmente, esse período foi definido como um dia.

Criou-se uma instância do BERTopic em que foram passados dois parâmetros: *language* e *min_topic_size*. O primeiro foi definido com o valor "portuguese", enquanto o segundo com valor 200. Com isso, o algoritmo irá utilizar os modelos de *embeddings* pré-treinados com dados em português, assim como especifica que os clusters devem ter pelo menos 200 tweets. Ainda que a coleta tenha sido de 120.000 tweets diários por média, o valor mínimo definido para a formação dos clusters foi de 200. Esse valor foi escolhido empiricamente para garantir que não tenhamos clusters com pouquíssimos documentos. O Código-fonte 4 mostra

[2] Disponível em https://pypi.org/project/nltk/3.7/

como esse processo foi implementado.

Assim, estabelecidos os argumentos, é chamada a função *fit_transform* e o algoritmo é executado. Neste trabalho todo o processo foi executado com base nos parâmetros padrão do BERTopic, mesmo havendo a possibilidade de personalização.

Código-fonte 4 – Código de execução do Bertopic sobre um conjunto de dados

```
from bertopic import BERTopic

tweets = dataset.tweet.to_list()
topic_model = BERTopic(language="portuguese", min_topic_size=200)
topics, _ = topic_model.fit_transform(tweets)
```

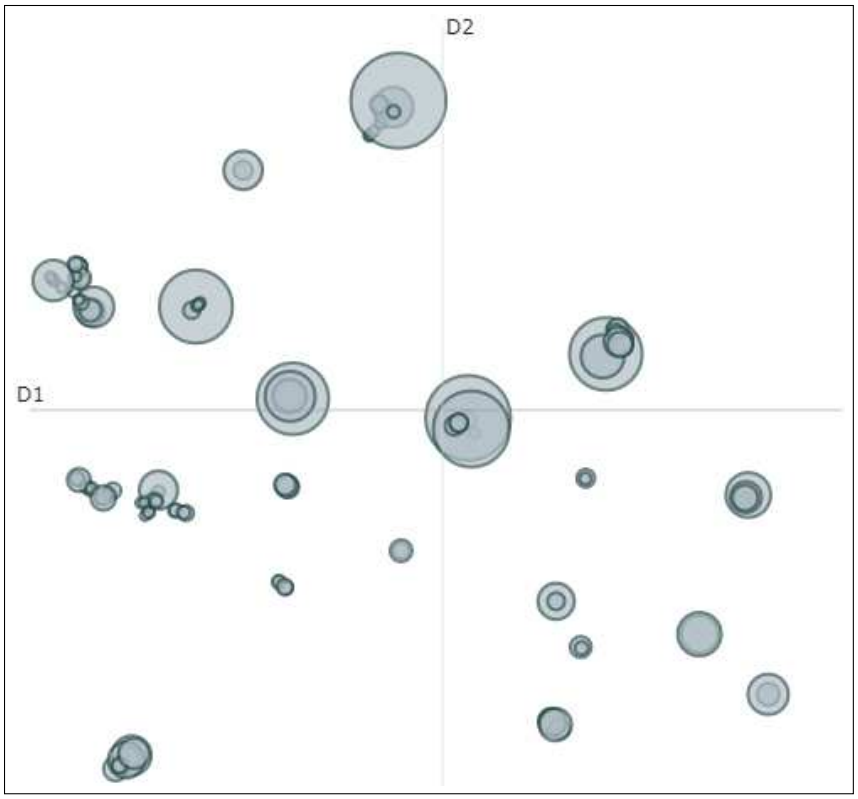
4.4 Visualização de tópicos

O BERTopic possui diferentes tipos de visualização dos dados. A *Intertopic Distance Map* dispõe em um plano cartesiano os clusters gerados, que podem estar distantes, próximos ou sobrepostos uns aos outros. Essa visualização é exemplificada na Figura 7, onde cada cluster é representado por um círculo cujo diâmetro é definido pela quantidade de documentos. A partir dessa visualização definiu-se o tamanho dos clusters que seriam gerados, pois ficava explícita grande quantidade de clusters pequenos gerados, além da sobreposição entre eles, o que indicava que estes poderiam pertencer a um mesmo assunto.

Outra visualização é a *Topic Hierarchy* que apresenta a hierarquia entre os clusters. É possível observar quais clusters possuem maior proximidade de assuntos e que poderiam ser unificados, processo que é repetido até a formação de um único grande cluster. Um exemplo de visualização é mostrado na Figura 8. Vale ressaltar que essa característica hierarquizada dos tópicos é decorrente da estrutura de hierarquias extraída durante a execução do HDBSCAN.

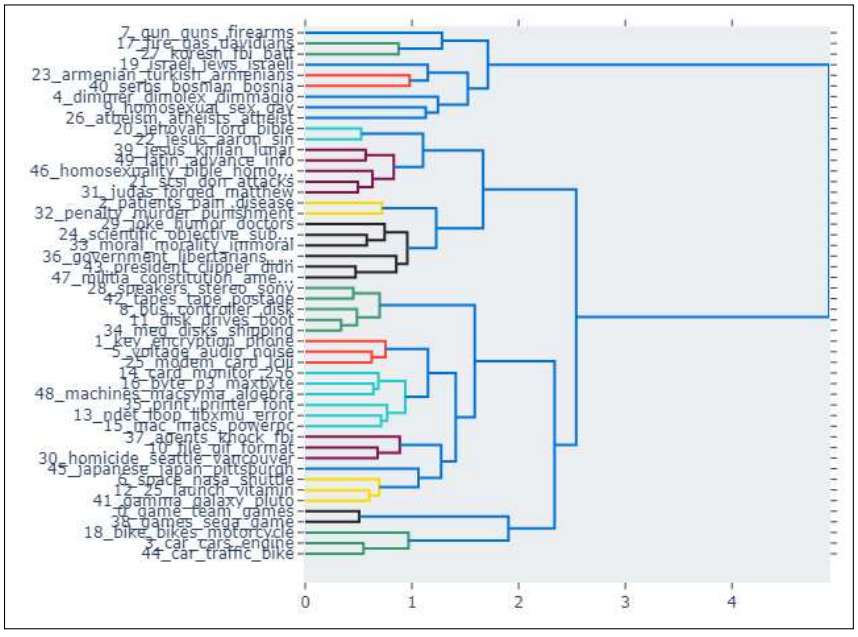
Também utilizou-se o *Topic Word Scores* ou *Barchart*, em que são dispostos os termos com maior pontuação no TF-IDF. Assim, tem-se a apresentação dos termos mais relevantes em cada um dos clusters extraídos. A Figura 9 ilustra essa visualização. No trabalho, utilizou-se as 10 palavras com as maiores pontuações. Por fim, utilizou-se *Topics over Time* que indica a quantidade de documentos associados aos clusters ao longo do tempo, exemplificado na Figura 10. No trabalho é possível acompanhar o surgimento e oscilação dos assuntos ao longo de um dia.

Figura 7 – Exemplo da visualização *Intertopic Distance Map*



Fonte: adaptado de Grootendorst (2021).

Figura 8 – Exemplo da visualização *Topic Hierarchy*

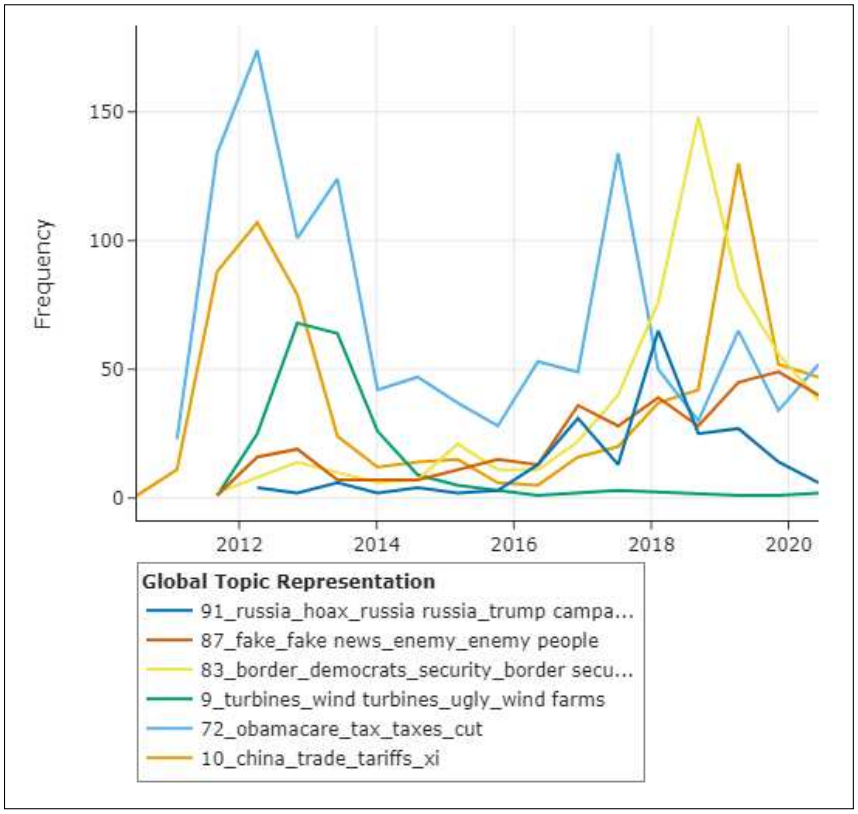


Fonte: adaptado de Grootendorst (2021).

Figura 9 – Exemplo da visualização *Topic Word Scores*

Fonte: adaptado de Grootendorst (2021).

Figura 10 – Exemplo da visualização *Topics Over Time*



Fonte: adaptado de Grootendorst (2021).

5 RESULTADOS

O trabalho tem como objetivo apresentar a evolução dos tópicos discutidos no Twitter durante o período da eleição presidencial brasileira de 2022, por meio da extração automática de tópicos provida pelo BERTopic. A coleta para construção do conjunto de dados teve início em 19/09/2022, 13 dias antes do primeiro turno e se estendeu até logo após o final do segundo turno, em 01/11/2022. O conjunto de dados conta com 5.615.277 tweets, mas apenas uma parte deles será analisada.

O período escolhido para análise vai do dia 28/09/2022 ao dia 05/10/2022, intervalo que engloba o primeiro turno das eleições. Durante os 8 dias foi utilizada a metodologia de coleta descrita anteriormente utilizando a API do Twitter e os filtros de buscas pelas eleições, resultando em 1.249.987 tweets efetivamente analisados.

Foi realizada uma tentativa frustrada de extração de tópicos com toda a massa de dados coletada, causada pela insuficiência de recursos de computação para tratar todo o conjunto de dados. O ambiente de execução utilizado foi o Google Colaboratory[1], que disponibiliza gratuitamente 12.68 GB de memória RAM e acesso a *Graphics Processing Unit (GPU)*, que não é especificada pelo Google. Por esta razão, toda a metodologia de extração de tópicos foi executada dia a dia, considerando o período de estudo.

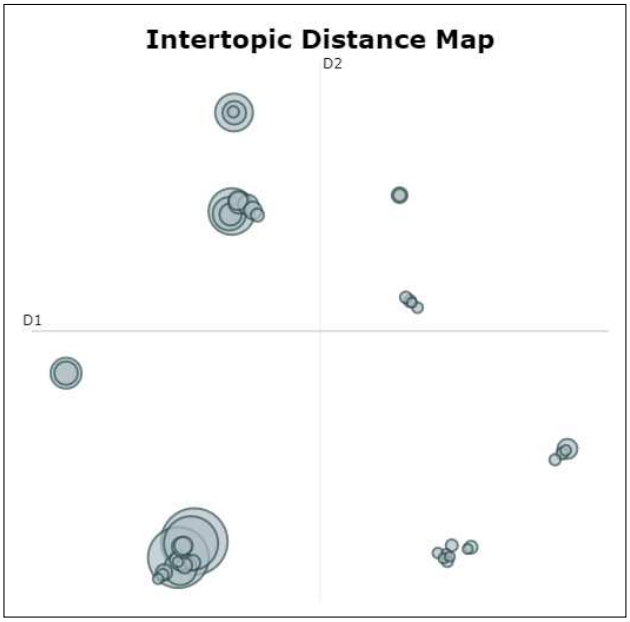
A fim de garantir a mesma expressividade dos resultados e manter a capacidade de analisar a evolução dos assuntos ao longo do período, foi feito um processo manual para identificar tópicos que se repetiam ao longo dos 8 dias analisados. O processo consiste em identificar quais tópicos surgem em mais de 1 dia, permitindo que seja feito um agrupamento de todas as palavras identificadas de cada um desses tópicos. Criou-se assim conjuntos de palavras únicas de um determinado assunto que perdurou ao longo dos dias.

Durante o processo de análise dos dados alguns ajustes foram realizados a fim de conseguir maior expressividade nos resultados. Um dos exemplos foram as decisões a respeito do tamanho dos clusters a serem analisados. A partir da visualização *Intertopic Distance Map*, na Figura 11, pode-se perceber o tamanho e a distância dos clusters e decidir pela realização da união dos clusters.

O BERTopic proporciona observar os tópicos mais relevantes de um cluster de diferentes maneiras, permitindo conclusões mais elaboradas. O gráfico *Topic Word Scores* na Figura 12, por exemplo, apresenta as palavras mais relevantes do dia 04/10/2022 de cada tópico.

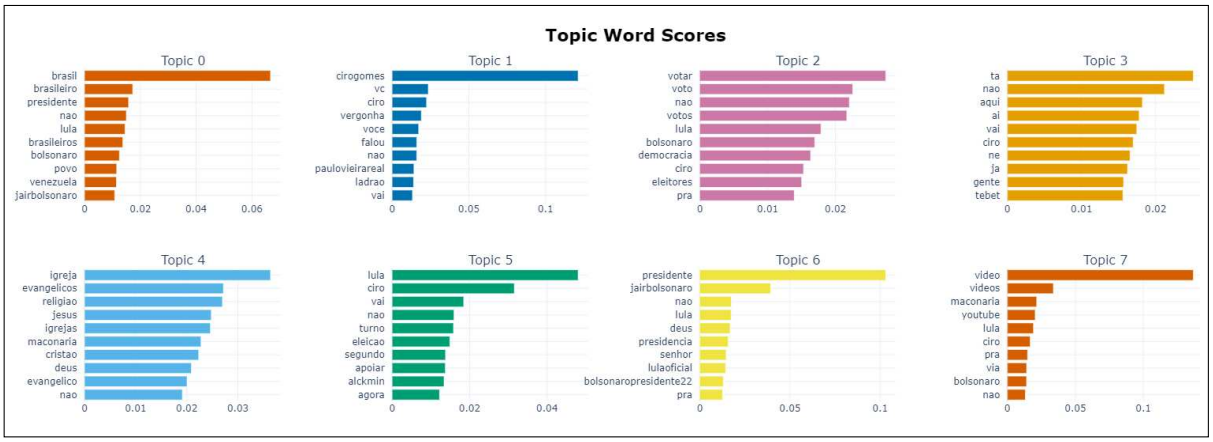
[1] https://colab.research.google.com/

Figura 11 – *Intertopic Distance Map* dos tópicos observados no dia 04/10/2022



Fonte: a própria autora.

Figura 12 – *Topic Word Scores* dos tópicos observados no dia 04/10/2022



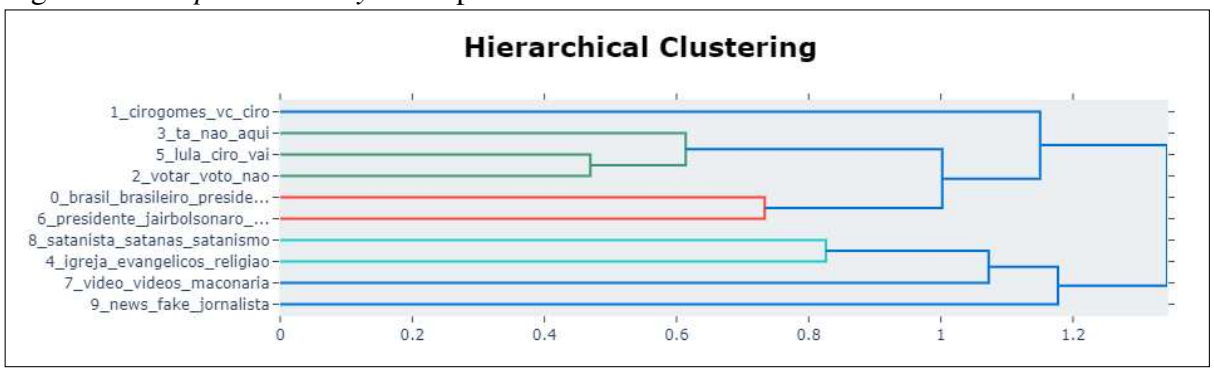
Fonte: a própria autora.

Já as visualizações como a do gráfico *Intertopic Distance Map*, ilustrado na Figura 11, junto com o gráfico *Topic Hierarchy*, ilustrado na Figura 13, ambos exemplos do dia 04/10/2022, contribuíram para a escolha do tamanho dos clusters, em razão da similaridade de diferentes tópicos representada pela sobreposição dos clusters e menor distância entre eles.

Após todo o processo de ajuste, identificação e agrupamento manual dos tópicos, criou-se o Quadro 2 que contém os tópicos mais relevantes do período, evidenciando os assuntos que perduraram por mais de um dia.

Analisando-se o Quadro 2, observa-se que, em média, os assuntos debatidos no Twitter encontravam-se inseridos entre 5 dos 12 assuntos identificados no período. Também é possível notar que a maioria dos assuntos teve uma média de duração de 3 dias, o que demonstra

Figura 13 – *Topic Hierarchy* dos tópicos observados no dia 04/10/2022



Fonte: a própria autora.

Quadro 2 – Dispersão dos tópicos ao longo do período analisado

Tópicos / Datas	**28/09**	**29/09**	**30/09**	**01/10**	**02/10**	**03/10**	**04/10**	**05/10**
Tópico 0	X	X	X	X	X	X	X	X
Tópico 1	X	X	X	X	X	X	X	
Tópico 2	X	X						
Tópico 3		X	X					
Tópico 4			X	X				
Tópico 5				X	X			
Tópico 6					X	X		X
Tópico 7						X	X	X
Tópico 8						X	X	X
Tópico 9						X	X	X
Tópico 10						X	X	X
Tópico 11							X	X

Fonte: elaborado pela autora.

a volatilidade das informações que adentram a rede social. Assim, tem-se exposição dos assuntos mais relevantes do período a partir da aplicação da ferramenta BERTopic. A ideia é que uma ferramenta como o BERTopic auxilie na identificação desses assuntos e assim permita que as pessoas possam transformar esses dados em informações que lhes possam ser úteis.

Por entender o Twitter como um meio onde os assuntos do cotidiano reverberam, buscou-se estabelecer uma relação entre os tópicos que apareceram como mais relevantes no período definido e as notícias veiculadas na mídia, de forma manual. Para cada tópico apresentado no Quadro 2, buscou-se pelo menos uma notícia publicada que tivesse relação com o assunto, com suporte da função *get_representative_docs*, que tem o objetivo de retornar os 3 documentos mais representativos de cada cluster encontrado pelo BERTopic. Esses tweets seriam aqueles possuem maior similaridade com os demais documentos dos seus respectivos clusters.

Ao conseguir relacionar os pontos apresentados, pode-se chegar à conclusão de que os assuntos são relevantes, pois a publicação de notícias na grande mídia só corrobora a existência e importância de determinado assunto. O inverso também pode ocorrer, como o exemplo dos

vídeos sobre a Maçonaria divulgados em diferentes redes sociais, que deram início as notícias sobre a relação dos candidatos e questões religiosas. Esses vídeos ocuparam os assuntos mais relevantes dos dias 03, 04 e 05 de outubro de 2022 e tornaram-se notícia tamanha a repercussão do caso. A Figura 14 resume as palavras e as notícias associadas a cada assunto identificado.

Na Figura 14 pode-se observar que existe alguma similaridade entre alguns tópicos. Tomando como exemplo os Tópicos 8, 9 e 10, é possível observar que estes tratam de assuntos relacionados a religião. Enquanto no Tópico 8 é recorrente o uso de palavras ligadas a *igreja* e *evangélicos*, no Tópico 9 é recorrente o uso de palavras como *vídeo* e *maçonaria*, e no Tópico 10 tem-se *satanista*. Todos esses tópicos estão ligados a viralização de um vídeo onde um dos candidatos foi associado a Maçonaria, o que desencadeou reações de eleitores que associam a organização a práticas não compatíveis com suas crenças. Assim, os tópicos, apesar de terem sido desencadeados por um mesmo evento, se apresentam de forma distinta pelo BERTopic dado o significado das palavras associadas.

Figura 14 – Palavras e notícias mais relevantes dos tópicos identificados no período

Tópico 0	
Palavras	brasil, brasildaesperanca, brasileiro, presidente, povo, nao, lula, brasileiros, bolsonaro, brasilvota22, brasileira, melhor, vai, vamos, bolsonaronoprimeiroturno, pais, venezuela
Notícia 1	Zeca Pagodinho promove campanha 'Brasil da Esperança' para Lula
Notícia 2	Apoio de artistas e influenciadores faz Lula ter recorde de engajamento nas redes em setembro
Tópico 1	
Palavras	votar, voto, nao, lula, vou, vai, pra, ciro, bolsonaro, votos, democracia, ciro,vota, lulaoficial, presidente, amanha, turno, eleitores, bolsonarismo
Notícia 1	Eleições no Brasil: um voto pela democracia
Notícia 2	O voto útil favorece o processo democrático
Tópico 2	
Palavras	cirogomes, anapaulamatosba, tamirfelipe, felipeneto,ciro, vc, nao, voce, ta, meupaiseursal, tvglobo, sistema, contra, ticostacruz, 12
Notícia 1	Campanha de Lula espera voto útil na hora H e aliados de Bolsonaro apostam em abstenção
Notícia 2	Vice de Ciro, Ana Paula atribui violência política à economia e nega acenos à direita
Tópico 3	
Palavras	estadual, federal, governador, senador, dep, 13, deputado, 400, deputada, colinha, 13, deputada
Notícia 1	Cola eleitoral pode facilitar votação
Notícia 2	Na véspera da eleição, candidatos ao governo ainda correm atrás dos votos
Tópico 4	
Palavras	padre, igreja, falso, fake, debate, kelmon, nao, lula, catolicos, fariseu, video, youtube, via, videos, tv, globo, padre
Notícia 1	'Padre de festa junina' foi termo mais falado nas redes sociais durante debate na Globo
Notícia 2	No "Todos contra Lula", padre de mentirinha é escada de moralista de araque
Tópico 5	
Palavras	amanaha, presidente, lula, pronto, intensifies, amanhalula, aa, saimos, bb, media, hoje, lulapresidente, lulanoprimeiroturno13, dia, bom, presidencia, lulano1oturno
Notícia 1	A estratégia final de Lula para vencer no 1º turno e o plano para um eventual 2º
Tópico 6	
Palavras	nordeste, norte, nordestino, sul, sudeste, estados, amo, pais, nordestinos, salvar, nordestinos, povo, lula, xenofobia, nao, aqui, votos, analfabetismo, bolsonaro
Notícia 1	VÍDEO: Bolsonaro associa vitória de Lula no Nordeste a "analfabetismo"e "falta de cultura"
Notícia 2	Bolsonaro diz que Nordeste tem mais analfabetos porque é governado pelo PT
Tópico 7	
Palavras	cirogomes, paulovieirareal, simonetebetbr, sorayathronicke, rogeriomorgado, jlivres, pdtnacional, ticostacruz, ciro, lulaoficial, vergonha, voce, falou, nao, ladrao, vai, vc, pdtnacional, cabodaciolo, engzin
Notícia 1	PDT, de Ciro Gomes, e Cidadania decidem apoiar Lula
Tópico 8	
Palavras	igreja, evangelicos, religiao, jesus, igrejas, maconaria, cristao, deus, evangelico, nao, lula, jesus
Tópico 9	
Palavras	video, videos, maconaria, youtube, lula, ciro, pra, via, bolsonaro, nao, veja
Notícia 1	Pastores minimizam possíveis efeitos de vídeo de Bolsonaro na maçonaria
Tópico 10	
Palavras	satanista, satanas, satanismo, maconaria, bolsonaro, satanistas, lula, diabo, cristao, deus, nao, video
Notícia 1	Por que discurso de Bolsonaro na maçonaria pode irritar evangélicos?
Notícia 2	Segundo turno começa com fake news associando Lula ao satanismo e polêmica sobre Bolsonaro na maçonaria
Tópico 11	
Palavras	news, fake, jornalista, fakenews, nao, estao, lula, jornalistas, jornalismo, campanha, tsejusbr, adocevida3, noticia
Notícia 1	Segundo turno começa com fake news associando Lula ao satanismo e polêmica sobre Bolsonaro na maçonaria
Notícia 2	Lista falsa de propostas do programa de governo de Lula circula nas redes

Fonte: a própria autora.

6 CONCLUSÕES E TRABALHOS FUTUROS

O trabalho tem como objetivo apresentar a evolução dos assuntos mais discutidos no Twitter ao longo do período eleitoral de 2022, com foco na eleição presidencial. Para identificar os assuntos mais comentados na rede social, utilizou-se uma ferramenta que realiza a extração de tópicos por meio de clusterização chamada BERTopic.

O BERTopic é uma ferramenta que aponta os termos mais relevantes de um documento utilizando algumas técnicas de aprendizagem de máquina como a criação dos embeddings, representando vetorialmente os termos para que pudessem ser identificados; a clusterização, agrupado os assuntos de semelhantes; e o TF-IDF, observando a relevância dos termos para cada cluster. Assim, verifica-se que o BERTopic agrega um conjunto de técnicas que lhe permitiria auxiliar na extração de tópicos, além de apresentar formas de visualização que facilitam a compreensão dos dados extraídos.

Com os resultados obtidos, pode-se observar que existem tópicos muito próximos e por isso poderiam fazer parte de um só grupo, a exemplo do Tópicos 8, 9 e 10 descrito nos resultados. Esses tópicos, no entanto, ocupam grupos distintos por haver pequenas nuances que são transformadas em agrupamentos diferentes. Isso indicaria que independente do desempenho do BERTopic, ainda é necessário que um especialista faça uma otimização dos modelos construídos pelo BERTopic.

O BERTopic utiliza-se do contexto para capturar os termos mais relevantes e essa é uma forma de tornar mais precisa a identificação dos assuntos. Essa organização, apesar de possuir um sentido e de conseguir identificar diferentes grupos, pode gerar situações em que os grupos parecem tratar do mesmo ponto. Essa sensação pode ser causada, como no trabalho atual, por se tratar de uma temática exclusiva, em que todos os assuntos possuem alguma ligação direta: eleições presidenciais de 2022. Isso só reforça a necessidade de um especialista no objeto de estudo.

Há também os *representative_docs*, que são os tweets dos documentos analisados que mais sintetizariam a ideia de cada cluster. Nesse ponto, alguns resultados apresentados não guardavam tantas relações com os 10 termos mais relevantes revelados pelo *get_topics*. Assim é necessário a realização de outros testes e respectivas análises para observar e identificar o comportamento.

Também é importante ressaltar que o BERTopic é uma ferramenta que pode ser utilizada para analisar um período completo. No presente trabalho, no entanto, não foi possível a

execução do período completo entre os dias 28/09/2022 a 05/10/2022 por limitações quanto ao uso de memória RAM. Assim, as análises dos tópicos do período foram realizadas manualmente conforme descritas nos resultados.

A execução em grandes períodos fica como sugestão para pesquisas futuras, o que será útil para comparar se a abordagem manual é compatível com os resultados obtidos a partir da execução do BERTopic. Além disso, outro possível trabalho futuro seria explorar as possibilidades de configuração do próprio BERTopic, que é implementado de forma a aceitar diferentes modelos de embeddings, redução de dimensão, clusterização e outros passos relevantes para obtenção de resultados.

REFERÊNCIAS

ALBERTO, R. **Clustering con DBSCAN y HDBSCAN con Python y sus hiperparámetros en SKlearn**. Medium, 2020. Disponível em: https://rubialesalberto.medium.com/clustering-con-dbscan-y-hdbscan-con-python-y-sus-hiperparÃąmetros-en-sklearn-8728283b96ac. Acesso em: 12 dez. 2022.

CAMPELLO, R. J.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: SPRINGER. **Pacific-Asia conference on knowledge discovery and data mining**. [*S. l.*], 2013. p. 160–172.

CER, D.; YANG, Y.; KONG, S.-y.; HUA, N.; LIMTIACO, N.; JOHN, R. S.; CONSTANT, N.; GUAJARDO-CESPEDES, M.; YUAN, S.; TAR, C. *et al.* Universal sentence encoder. **arXiv preprint arXiv:1803.11175**, 2018.

CERVI, E. U.; MASSUCHIN, M. G. Redes sociais como ferramenta de campanha em disputas subnacionais: análise do twitter nas eleições para o governo do paraná em 2010. **Sociedade e Cultura**, v. 15, n. 1, p. DOI: 10.5216/sec.v15i1.20670, out. 2012. Disponível em: https://revistas.ufg.br/fcs/article/view/20670.

Columbia Public Health. **K-Means Cluster Analysis**. United States: Columbia Public Health, 2022. Disponível em: https://www.publichealth.columbia.edu/research/population-health-methods/k-means-cluster-analysis. Acesso em: 24 nov. 2022.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: **kdd**. [*S. l.*: *s. n.*], 1996. v. 96, n. 34, p. 226–231.

EZUGWU, A. E.; IKOTUN, A. M.; OYELADE, O. O.; ABUALIGAH, L.; AGUSHAKA, J. O.; EKE, C. I.; AKINYELU, A. A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. **Engineering Applications of Artificial Intelligence**, v. 110, p. 104743, 2022. ISSN 0952-1976. Disponível em: https://www.sciencedirect.com/science/article/pii/S095219762200046X.

FONSECA, C. **Word Embedding: fazendo o computador entender o significado das palavras**. Medium, 2021. Disponível em: https://medium.com/turing-talks/word-embedding-fazendo-o-computador-entender-o-significado-das-palavras-92fe22745057. Acesso em: 24 nov. 2022.

FORGY, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. **biometrics**, v. 21, p. 768–769, 1965.

GROOTENDORST, M. **Topic Visualization**. Github.IO, 2021. Disponível em: https://maartengr.github.io/BERTopic/getting_started/visualization/visualization.html. Acesso em: 24 nov. 2022.

GROOTENDORST, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. **arXiv preprint arXiv:2203.05794**, 2022.

JINDAL, D. **Fine-Grained Analysis of Sentence Embeddings**. Towards Data Science, 2019. Disponível em: https://towardsdatascience.com/fine-grained-analysis-of-sentence-embeddings-a3ff0a42cce5. Acesso em: 12 dez. 2022.

JUNIOR, V. O. D. S.; BRANCO, J. A. C.; OLIVEIRA, M. A. D.; SILVA, T. L. C. D.; CRUZ, L. A.; MAGALHãES, R. P. A natural language understanding model covid-19 based for chatbots. In: **2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)**. [*S. l.*: *s. n.*], 2021. p. 1–7.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: PMLR. **International conference on machine learning**. [*S. l.*], 2014. p. 1188–1196.

LYU, J. C.; HAN, E. L.; LULI, G. K. Covid-19 vaccine–related discussion on twitter: Topic modeling and sentiment analysis. **Journal of Medical Internet Research**, v. 23, n. 6, p. e24435, Jun 2021. ISSN 1438-8871. Disponível em: https://doi.org/10.2196/24435.

OLIVEIRA, B. S. N.; RÊGO, L. G. C. do; PERES, L.; SILVA, T. L. C. da; MACÊDO, J. A. F. de. Processamento de linguagem natural via aprendizagem profunda. **Sociedade Brasileira de Computação**, 2022.

QAISER, S.; ALI, R. Text mining: use of tf-idf to examine the relevance of words to documents. **International Journal of Computer Applications**, Foundation of Computer Science, v. 181, n. 1, p. 25–29, 2018.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019. Disponível em: http://arxiv.org/abs/1908.10084.

RODRIGUES, P. R. F. **Dinâmica de temas abordados no Twitter via evolução de clusters**. Monografia (Graduação em Engenharia de Software) – Universidade Federal do Ceará, Campus Quixadá, Quixadá, 2016.

SOUSA, A.; BECKER, K. Comparando os posicionamentos a favor/contra a vacinação covid nos estados unidos da américa e no brasil. In: **Anais do XXXVII Simpósio Brasileiro de Bancos de Dados**. Porto Alegre, RS, Brasil: SBC, 2022. p. 65–77. ISSN 2763-8979. Disponível em: https://sol.sbc.org.br/index.php/sbbd/article/view/21796.

VERGEER, M. Twitter and political campaigning. **Sociology Compass**, v. 9, n. 9, p. 745–760, 2015. Disponível em: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12294.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. **Science**, v. 359, n. 6380, p. 1146–1151, 2018. Disponível em: https://www.science.org/doi/abs/10.1126/science.aap9559.