Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Caracterização de tweets de senadores, governadores e convidados durante a CPI da Pandemia: uma análise através das principais hashtags

Guilherme Libardi Gonçalves

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Helen de Cassia Sousa da Costa Lima

COORIENTAÇÃO:
Filipe Nunes Ribeiro

Junho, 2022
João Monlevade–MG

Guilherme Libardi Gonçalves

Caracterização de tweets de senadores, governadores e convidados durante a CPI da Pandemia: uma análise através das principais hashtags

Orientador: Helen de Cassia Sousa da Costa Lima

Coorientador: Filipe Nunes Ribeiro

Monografia apresentada ao curso de Engenharia de Computação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Junho de 2022

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

G635c Gonçalves, Guilherme Libardi.
Caracterização de tweets de senadores, governadores e convidados durante a CPI da Pandemia [manuscrito]: uma análise através das principais hashtags. / Guilherme Libardi Gonçalves. - 2022.
58 f.: il.: color., gráf., tab..

Orientadora: Profa. Dra. Helen de Cassia Sousa da Costa Lima.
Coorientador: Prof. Dr. Filipe Nunes Ribeiro.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Computação .

1. Comissões parlamentares de inquérito. 2. COVID-19. 3. Pandemias. 4. Política. 5. Twitter (Rede social on-line). I. Lima, Helen de Cassia Sousa da Costa. II. Ribeiro, Filipe Nunes. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 004.775

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Guilherme Libardi Gonçalves

Caracterização de *tweets* de senadores, governadores e convidados durante a CPI da Pandemia: uma análise através das principais *hashtags*

Monografia apresentada ao Curso de Engenharia da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia da Computação

Aprovada em 22 de junho de 2022

Membros da banca

Dra. Helen de Cássia Sousa da Costa Lima - Orientadora (Universidade Federal de Ouro Preto)
Dr. Filipe Nunes Ribeiro - Coorientador (Universidade Federal de Ouro Preto)
Me. Alexandre Magno de Sousa (Universidade Federal de Ouro Preto)
Dr. Carlos Henrique Gomes Ferreira (Universidade Federal de Ouro Preto)

Helen de Cássia Sousa da Costa Lima, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 27/06/2022

Documento assinado eletronicamente por **Helen de Cassia Sousa da Costa Lima**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 27/06/2022, às 14:59, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0352324** e o código CRC **8E96F708**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: (31)3808-0819 - www.ufop.br

Este trabalho é dedicado à minha avó Isis e minha mãe Idalira

Agradecimentos

Agradeço primeiramente a Deus, por me guiar durante toda a graduação. À minha amada avó Isis, pela sabedoria que me foi passada em momentos difíceis. À minha mãe, Idalira, por ser mais do que uma mãe em todos os momentos da minha vida. Ao meu pai, Rodolfo, pelos ensinamentos que me apresentou durante a vida. Agradeço ao Guilherme, pelo apoio e por acreditar em mim desde o começo. Ao meu avô Miro, pela sabedoria que me foi passada. Ao meu avô Emídio, que lá de cima tenho certeza que gostaria de participar deste momento tão especial. Ao meu irmão Rhuan, que me ensinou a usar o primeiro computador que tivemos. Aos meus irmãos de república Felipe e Lorran, os quais levarei para o resto da vida. À Helen, minha orientadora, pela paciência, incentivos e direcionamentos, sem os quais este trabalho não seria possível. Ao Filipe, pela participação fundamental na construção da pesquisa realizada neste trabalho. À Vanessa, por me apoiar em todos os momentos. Aos profissionais que pude conhecer na CEMIG e Altasnet, Valmar e Wanderson, cuja colaboração e compreensão tornaram este trabalho possível de ser realizado. Aos meus amigos de IFF, Gabriel, Alexandre e Charles.

"A verdadeira viagem de descobrimento não consiste em procurar novas paisagens, mas em ter novos olhos."

— Marcel Proust (1871 – 1922),
No Caminho de Swann.

Resumo

A pandemia de COVID-19 foi o principal fator que ocasionou a crise de oxigênio que ocorreu na cidade de Manaus em janeiro de 2021. Além disso, uma má gestão do consumo de produtos hospitalares também contribuiu para a falta do recurso na cidade. A crise foi objeto de investigação pelo Senado Federal, através da instalação da CPI da Pandemia em abril de 2021. Grande parte da discussão popular sobre os fatos investigados e revelados pela CPI ocorreu nas redes sociais, em especial no Twitter. Neste trabalho é analisado o conteúdo dos *tweets* dos principais participantes do debate público, através de uma caracterização do conteúdo postado pelos senadores e governadores da república, bem como os convidados e testemunhas que compareceram à CPI e que possuam uma conta no Twitter. É feita a identificação, categorização por contexto e uma análise de correlação entre as principais categorias de *hashtags*. Também é realizada uma análise de tópicos com o modelo LDA para cada uma das principais categorias identificadas. Por fim é feita uma correlação entre os períodos de maior e menos intensidade de publicações e as pessoas que estiveram na CPI nesses períodos. Assim, é possível identificar uma correlação entre *hashtags* que referenciam a CPI da Pandemia e a COVID-19, além de uma correlação de assuntos como uso de máscara, vacinação e COVID-19, sendo frequentemente citados juntos em um mesmo *tweet*. A *hashtag* mais utilizada pelos usuários analisados neste trabalho foi a #cpidacovid, indicando que este foi o principal assunto debatido. Com a modelagem de tópicos foi possível visualizar que o presidente Bolsonaro foi citado em todos os assuntos identificados nas *hashtags* que compõem a categoria CPI da Pandemia. Senadores falaram mais sobre a CPI da Pandemia, assim como os Governadores falaram mais sobre vacinação em seus estados, desta forma, os atores políticos abordaram, em suas contas no Twitter, os temas que eram de sua responsabilidade. O Ministro da Saúde, Marcelo Queiroga, principal autoridade de saúde no Brasil, foi a pessoa que mais mencionou *hashtags* ligadas à COVID-19, abordando assuntos como eficácia das vacinas, campanhas de vacinação e notícias sobre a pandemia. O evento da CPI que ocorreu no período de maior volume de postagens pelos usuários analisados foi a semana em que Luciano Hang, dono das lojas Havan e apoiador do presidente Jair Bolsonaro, esteve na CPI. A medida com que as investigações da CPI sobre os possíveis casos de corrupção no ministério da saúde avançaram, é possível perceber uma tendência de aumento do volume de postagens.

Palavras-chaves: Twitter, política, COVID-19, CPI, Pandemia.

Abstract

The COVID-19 pandemic was the principal factor that caused the oxygen crisis occurred in Manaus in January 2021. Furthermore, poor management of hospital products also contributed to the lack of this resource in the city. The crisis was investigated by the Federal Senate, through the installation of the Pandemic PCI in April 2021. Much of the popular discussion about the facts investigated and revealed by the PCI took place on social media, especially on Twitter. In this work, the content of the tweets from the main participants of the public debate is analyzed, through a characterization of the content posted by the senators and governors, as well as the guests and witnesses who attended the CPI and have a Twitter account. Identification, categorization by context and a correlation analysis between the main categories of hashtags are performed in this work. A topic analysis is also performed using the LDA model for each of the main identified categories. Finally, a correlation is made between the periods of greater and lesser intensity of publications and the people who were at the PCI in these periods. Thus, it is possible to identify a correlation between hashtags that reference the Pandemic PCI and COVID-19, as well as a correlation of subjects such as mask use, vaccination and COVID-19, which are often cited together in the same tweet. The most used hashtag by the users analyzed in this work was #cpidacovid, indicating that this was the main topic discussed. With the topic modeling, it was possible to visualize that President Bolsonaro was mentioned in all the subjects identified in the hashtags that make up the CPI category of the Pandemic. Senators tweeted more about the Pandemic CPI, just as Governors spoke more about vaccination in their states, in this way, political actors addressed, on their Twitter accounts, the topics that were their responsibility. The Minister of Health, Marcelo Queiroga, the main health authority in Brazil, was the person who most mentioned hashtags linked to COVID-19, addressing issues such as vaccine effectiveness, vaccination campaigns and news about the pandemic. The PCI event that took place in the period with the highest volume of posts by users analyzed was the week in which Luciano Hang, owner of Havan stores and supporter of President Jair Bolsonaro, was at the PCI. As the PCI's investigations into possible cases of corruption in the Ministry of Health advanced, it is possible to perceive a trend towards an increase in the volume of posts.

Key-words: Twitter, Politics, COVID-19, PCI, Pandemic.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ANVISA Agência Nacional de Vigilância Sanitária

API *Application Programming Interface*

CV Coeficiente de Variação

COVID-19 *Coronavirus Disease 2019*

CNN *Convolutional Neural Network*

CPI Comissão Parlamentar de Inquérito

LDA *Latent Dirichlet Allocation*

LSTM *Long short-term memory*

OMS Organização Mundial de Saúde

OPAS Organização Pan-Americana da Saúde

STF Supremo Tribunal Federal

TCU Tribunal de Contas da União

TF-IDF *Term Frequency-Inverse Document Frequency*

WEB *World Wide Web*

Sumário

1 Introdução

Em 31 de dezembro de 2019 a Organização Mundial de Saúde (OMS) foi alertada sobre vários casos de pneumonia na cidade de Wuhan, na China. Uma semana depois, as autoridades chinesas confirmaram que se tratava de um novo tipo de coronavírus, causador da doença conhecida hoje por *Coronavirus Disease 2019* (COVID-19). O vírus de Wuhan se espalhou por vários países, até que em 11 de março de 2020, a OMS caracteriza a doença como uma pandemia (SAÚDE, 2020b).

Após a caracterização da doença como pandemia, a Organização Pan-Americana da Saúde (OPAS) alertou para uma infodemia como consequência da pandemia de COVID-19: um excesso de informações, algumas precisas e outras não, que tornam difícil encontrar fontes idôneas e orientações confiáveis quando se precisa (SAÚDE, 2020a). No Brasil, durante a pandemia de COVID-19, o Governo Federal foi acusado de propagar desinformação nas redes sociais e incentivar a compra e o uso de medicamentos sobre os quais não haviam comprovação científica (CARVALHO; GUIMARÃES, 2020).

Em janeiro de 2021, a cidade de Manaus no estado do Amazonas sofreu um aumento repentino no consumo de oxigênio, ultrapassando a produção da região. A crise de oxigênio resultou na morte por asfixia de pelo menos 30 pessoas, e a transferência de mais de 500 pacientes para outros estados (GAZEL; CRUZ, 2022; LAVOR et al., 2021). Na época, o governo federal foi acusado de criar tendas de atendimento na cidade, para distribuição de medicamentos sem eficiência comprovada contra a COVID-19 (NACIONAL, 2021).

Para investigar possíveis omissões do governo federal no combate à pandemia de COVID-19, e a causa do agravamento da crise sanitária em Manaus, em 27 de abril de 2021, foi instalada a Comissão Parlamentar de Inquérito (CPI) da Pandemia (SENADO, 2021e). A CPI foi amplamente discutida nas redes sociais, gerando um 11,8 milhões de publicações no Twitter. O grande volume de publicações foi associado à records de audiência em outras plataformas, como o Youtube (ZACHARIAS, 2021).

No Brasil, políticos vêm adotando o *Twitter* como parte relevante de suas comunicações, desde que a rede social foi utilizada pelos candidatos às eleições do Parlamento Europeu em 2009, a primeira utilização da rede como ferramenta de comunicação e na política brasileira, nas eleições gerais de 2010 (VERGEER; HERMANS; SAMS, 2013; CREMONESE, 2012). Mas foi na corrida presidencial vitoriosa de Barack Obama em 2012, que a rede se mostrou ser uma valiosa ferramenta de comunicação com o eleitorado. Políticos de todo o mundo tem voltado suas atenções ao uso de novas tecnologias, em especial as redes sociais (AMARAL; PINHO, 2018).

No *Twitter*, os usuários podem postar pequenas mensagens, de até 280 caracteres,

contendo ou não uma opinião sobre qualquer assunto, essas mensagens são denominadas *tweets*. Os assuntos discutidos podem ser centralizados em torno de palavras-chave chamadas de *hashtag*, basta que o usuário mencione uma *hashtag* em seu *tweet* para participar de um assunto (WANG et al., 2012). As *hashtags* são responsáveis por organizar uma discussão em tópicos específicos, e seu uso para fins políticos se tornou relevante em eventos como as eleições iranianas de 2009, onde a *hashtag* mais utilizada no mundo foi a *#iranelection* (SMALL, 2011).

A compreensão do conteúdo postado nas redes sociais dos políticos tem sido cada vez mais objeto de pesquisa ao redor do mundo, sua análise permite descobrir padrões de comportamento e conhecimentos que podem passar despercebidos no momento em que estão sendo publicados, mas que podem ser descobertos através da mineração de texto. Desta forma, este trabalho buscar caracterizar as publicações dos Senadores da República, dos Governadores, bem como de convidados e testemunhas que tenham conta no Twitter, durante a CPI da Pandemia.

1.1 Objetivos

O principal objetivo deste trabalho é identificar os principais assuntos discutidos durante a CPI da Pandemia, através de uma análise das *hashtags* mais utilizadas no Twitter, pelos governadores, senadores, convidados e testemunhas.

Para alcançar o objetivo geral, o trabalho busca realizar os seguintes objetivos específicos:

- Coletar os *tweets* dos senadores, governadores, convidados e testemunhas da CPI, no período da CPI da Pandemia;
- Identificar as principais *hashtags* utilizadas pelos senadores, governadores, convidados e testemunhas da CPI, dentre os *tweets* coletados;
- Agrupar as diferentes *hashtags* que falam de um mesmo assunto em categorias;
- Descobrir os tópicos discutidos nas categorias mais relevantes;
- Relacionar os principais períodos de maior e menor intensidade de postagens de *tweets* com os convidados e testemunhas que estiveram presentes na CPI da Pandemia nesses períodos.

1.2 Estrutura do documento

Neste Capítulo 1 foram contextualizados os problemas de pesquisa, o objetivo geral e os específicos, bem como a estrutura do trabalho. No Capítulo 2 serão abordados os

referenciais utilizados na fundamentação teórica do trabalho, os trabalhos relacionados e as tecnologias utilizadas na pesquisa. O Capítulo 3 sumariza e descreve em detalhes o processo de construção da base de dados utilizada para extrair os resultados da pesquisa. O Capítulo 4 apresenta os resultados da pesquisa, através da descrição do conjunto de dados, uma análise postagens em relação aos acontecimentos na CPI da Pandemia, o capítulo também descreve como as *hashtags* mais utilizadas se relacionam e finaliza realizando uma análise de tópicos. Por fim, o Capítulo 5 conclui o presente trabalho e apresenta possibilidades para trabalhos futuros.

2 Revisão bibliográfica

Este capítulo busca embasar os conceitos utilizados durante o desenvolvimento deste trabalho, através de uma revisão da bibliografia existente sobre esses conceitos, bem como utilizar pesquisas já existentes que também enforçam o tema proposto no presente trabalho. Na Seção 2.1 é apresentada a CPI da Pandemia, seus objetivos e motivações. A Seção 2.2 argumenta sobre o uso do Twitter por atores políticos. A Seção 2.3 apresenta os principais conceitos de mineração de texto, base para a realização deste trabalho. De forma semelhante, a Seção 2.4 discute os principais métodos de pré-processamento de texto para análise da mineração de texto. A Seção 2.5 apresenta a técnica de modelagem de tópicos utilizada neste trabalho. Por fim, a Seção 2.6 relaciona as contribuições de outros trabalhos relacionados com este trabalho.

2.1 Comissão Parlamentar de Inquérito (CPI)

Em abril de 2021 uma Comissão Parlamentar de Inquérito (CPI) foi instaurada no Senado brasileiro para investigar a atuação do governo federal na resposta brasileira à pandemia (bem como o repasse de recursos federais aos governos estaduais) (SENADO, 2021e). De acordo com dados oficiais, o Brasil tinha no momento mais de 14 milhões de casos confirmados e quase 400 mil mortes, com um consenso de que estes números são sub-relatados (G1, 2021a; ORELLANA et al., 2021).

A resposta da COVID-19 do Brasil foi considerada a pior do mundo pelo ranking do *Lowy Institute* de janeiro de 2021, e também por organizações não governamentais relevantes no cenário global da saúde, como “Médicos Sem Fronteiras”, que reconheceu a existência de uma catástrofe humanitária no país. (INSTITUTE, 2021; REUTERS, 2021).

A CPI é baseada em um banco de dados de mais de 3.000 instrumentos legais federais coletados em 2020; a jurisprudência do Supremo Tribunal Federal (STF) e do Tribunal de Contas da União (TCU), além de documentos e discursos oficiais, vídeos, postagens e notícias que veiculam entrevistas e declarações de autoridades brasileiras (MACHADO, 2021). De acordo com Senado (2021e), em suas conclusões, o relatório identifica três eixos da estratégia institucional para a disseminação do vírus:

1. Atos normativos federais: incluindo a promulgação de regras por autoridades e órgãos federais e vetos presidenciais - por exemplo, decretos que definiam os serviços religiosos, a indústria da construção, salões de beleza, barbearias e academias como serviços essenciais para que pudessem permanecer abertos mesmo sob confinamento;

e veta uma série de medidas legislativas para limitar a disseminação da COVID-19, como o uso obrigatório de máscaras faciais em espaços fechados, inclusive em prisões.

2. Atos de obstrução aos esforços dos governos estaduais e municipais para responder à pandemia: principalmente a "guerra" contra governantes que adotaram medidas de contenção da doença, que incluem a demora injustificável de verbas emergenciais a estados e municípios, e ação judicial perante o Supremo Tribunal Federal contra três governadores, que suspenderam temporariamente as atividades comerciais, o que foi indeferido por inconsistências jurídicas básicas.

3. Propaganda contra a saúde pública: aqui definida como o discurso político que utiliza argumentos econômicos, ideológicos e morais, além de notícias falsas e informações técnicas sem comprovação científica, com o objetivo de desacreditar as autoridades sanitárias, fragilizando a adesão popular à ciência. Para promover o ativismo político contra as medidas de saúde pública necessárias para conter a disseminação da COVID-19.

2.2 A relação de políticos e o uso do Twitter

Twitter é uma plataforma social, que permite que seus usuários se comuniquem de maneira pública. Ministérios das Relações Exteriores utilizam o Twitter para expandir sua presença online. Redes diplomáticas digitais e funcionários do governo são incentivados a interagir com o público por meio desta ferramenta (SILVA, 2012). No mundo político, o ator mais comunicativo no Twitter é o Governo do Nepal, com 96% de seus *tweets* tendo respostas de outros usuários do Twitter (SETIAWAN et al., 2022).

Outra camada mais sutil da diplomacia do Twitter é o seguimento mútuo de pares entre chefes de estado oficiais, ministros e outras contas do governo. Em junho de 2020, o Ministério das Relações Exteriores da Islândia estava classificado em primeiro lugar, tendo 147 conexões mútuas com o mundo, líderes e ministros das Relações Exteriores no Twitter (BERNARDES, 2021). Muitos parlamentares usam o Twitter diariamente para fazer comentários sobre outros políticos, celebridades e notícias diárias, por vezes hostilizando outras pessoas com suas declarações polêmicas (ESTADO, 2022).

O Twitter é caracterizado por recursos específicos que o tornam uma ferramenta interessante para a comunicação política. Nesta rede social, é possível que os usuários divulguem mensagens curtas, de até 280 caracteres, dirigidas a um vasto público constituído principalmente por seus seguidores (OSMAN, 2021). Dos recursos do Twitter, tem-se, por exemplo, a inclusão de usuários (@ - menções), links para conteúdo externo (hiperlinks), e tópicos (*hashtags*). Esses recursos têm sido utilizados como ferramenta de comunicação entre políticos e seus eleitores (CREMONESE, 2012).

No contexto político, o Twitter trabalha com a promessa de uma forma mais direta de comunicação política onde os parlamentares falam pessoalmente para seus seguidores. No entanto, apesar do potencial de interação mais próxima com o eleitorado, não necessariamente essa comunicação ocorre pelos políticos que usam o Twitter, uma vez que alguns utilizam a rede para discutir temas sem relevância para a sociedade, com o objetivo de aumentar seu engajamento (MARTINI, 2022; ESTADO, 2022).

Com o início da pandemia de COVID-19, os líderes políticos tiveram que informar os cidadãos sobre as medidas que deveriam tomar para evitar a propagação do vírus e quais medidas estavam sendo tomadas pelo mundo político para minimizar os danos causados pela pandemia. As redes sociais oferecem a possibilidade de informar o público em tempo real sobre qualquer assunto, desta forma, diversos atores políticos utilizaram o Twitter para informar o público sobre as medidas de enfrentamento à pandemia e informações sobre a distribuição de vacinas (OSMAN, 2021).

O Twitter é uma plataforma conveniente para obter informações, pois os indivíduos precisam apenas de um celular ou tela de computador e acesso à internet. No ano de 2019, de acordo com IBGE (2021), o Brasil possuia 82,7% de domicílios com acesso à internet, e 99,5% das casas brasileiras com acesso à internet possuem pelo menos um aparelho celular. Os usuários que buscam informações sobre a pandemia podem descobrir facilmente quais medidas governamentais estão em vigor, como está o avanço da pandemia na sua região, entre outras informações.

Cientistas políticos e pesquisadores podem analisar o Twitter para obter *insights*, assim como as pesquisas de Amaral e Pinho (2018), Vergeer, Hermans e Sams (2013) e Cremonese (2012). Atualmente, muitos estudos estão sendo publicados buscando analisar e examinar a disseminação de desinformação nas mídias sociais durante a pandemia COVID-19, descobrindo, por exemplo, que informações falsas sobre a pandemia são mais publicadas, mas menos recebem menos repostagens do que *tweets* com base científica.

Recentemente, devido à pandemia, a alteração nas políticas de uso do Twitter permitiu uma melhor avaliação da plataforma sobre o uso de personalidades políticas no aplicativo, o que pode evitar, por exemplo, a disseminação de notícias falsas, conhecidas como *fake news*. Assim, o Twitter anunciou recentemente a expansão de suas regras para cobrir conteúdo que poderia ser contra as informações de saúde pública fornecidas por fontes oficiais e poderia colocar as pessoas em maior risco de transmitir COVID-19 (MARTINI, 2022). Os *tweets* excluídos agora redirecionam para a página de regras e políticas do Twitter.

No mês de março de 2020, o Twitter deletou dois tweets do presidente brasileiro Jair Bolsonaro porque continham informações falsas ou enganosas sobre a COVID-19, a doença causada pelo novo coronavírus (ESTADO DE SÃO PAULO, 2022). As exclusões marcaram a primeira vez que o site tomou medidas contra o conteúdo postado pelo

Presidente Bolsonaro, eleito em outubro de 2018 (G1, 2020). Não é a primeira vez que o site usa sua política de coronavírus para excluir uma postagem de um chefe de estado, que o Twitter oferece uma latitude mais ampla do que para a maioria dos usuários. Ainda no mesmo mês, o site deletou uma postagem do presidente venezuelano Nicolás Maduro por promover uma "poção natural" para curar COVID-19 (VEJA, 2020).

Já em julho de 2020, o ministro Alexandre de Moraes ordenou a retirada de 16 contas do Twitter e 12 contas do Facebook, decisão vinculada a uma investigação em andamento sobre a suposta divulgação de desinformação por partidários de direita. Um dos objetivos da investigação de *fake news*, é descobrir se desinformação e ameaças contra funcionários do Supremo Tribunal Federal estão sendo financiadas de forma ilícita. Entre os donos das contas suspensas estão Roberto Jefferson, ex-deputado e presidente do conservador PTB, além dos empresários Luciano Hang, Edgar Corona e Oscar Fakhoury, e a ativista Sara Giromini, mais conhecida como Sarah Winter (JURiDICO, 2020).

2.3 Mineração de texto

De acordo com Feldman (2006), a mineração de textos busca extrair informação útil de coleções de documentos, através da identificação e exploração de padrões interessantes, a partir de dados textuais desestruturados encontrados nesses documentos. Jo (2019) complementa esta definição dizendo que mineração de texto é o processo de encontrar conhecimento implícito a partir de dados textuais.

Um documento pode ser definido como uma unidade de dado textual em uma coleção, que pode ou não correlacionar com um documento real, por exemplo, um e-mail ou uma notícia (FELDMAN, 2006). Texto, por sua vez, pode ser definido como dados desestruturados composto por *strings* que são denominadas palavras, de acordo com Jo (2019).

Miner (2012) argumenta que a mineração de texto pode ser dividida em sete áreas, baseando-se nas características únicas de cada área. Apesar de distintas, essas áreas estão inter-relacionadas. Um projeto de mineração de texto pode utilizar técnicas de múltiplas áreas. As sete áreas de Miner (2012) são:

1. Aquisição de informações: processo de buscar, armazenar e adquirir documentos.
2. Clusterização de documentos: agrupamento e categorização de termos ou documentos, utilizando técnicas de clusterização de dados.
3. Classificação de documentos: agrupamento e categorização de termos ou documentos, utilizando técnicas de classificação de dados, baseando-se em modelos treinados com exemplos rotulados.

4. Mineração web: mineração de dados e texto na *Internet*, com foco específico na interconexão e escalabilidade da *World Wide Web* (WEB).

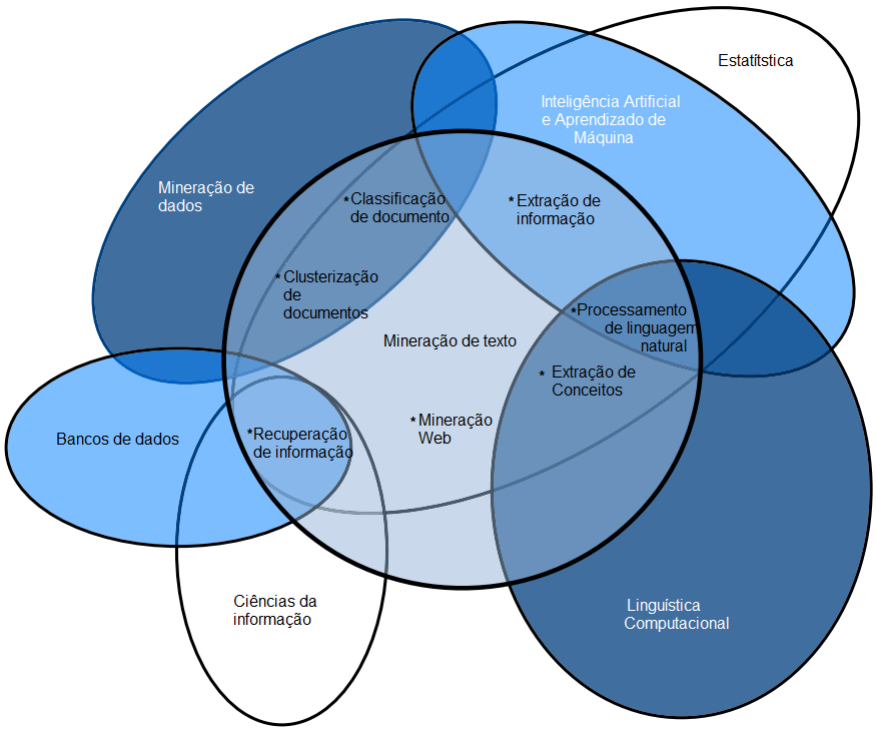
5. Extração de informações: identificação e extração de fatos relevantes e relacionamentos de texto desestruturado, ou o processo de construir dados estruturados a partir de texto desestruturado ou semi-estruturado.

6. Processamento de linguagem natural: processamento de baixo nível de linguagem e suas tarefas correlacionadas.

7. Extração de conceitos: agrupamento de palavras, termos ou frases em grupos semanticamente similares.

A Figura 1 ilustra como as sete áreas de Miner (2012) se relacionam entre si e com outras áreas do conhecimento.

Figura 1 – Áreas da mineração de texto



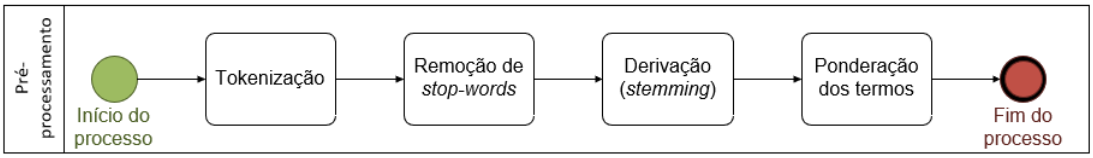
Fonte: Adaptado de Miner (2012)

2.4 Pré-processamento

De acordo com Feldman (2006), Miner (2012) e Jo (2019), a mineração de texto, deve ser precedido de um processo denominado pré-processamento. Para Vijayarani et

al. (2015), o pré-processamento consiste em quatro etapas fundamentais: tokenização, remoção de *stop-words*, derivação (*stemming*) e *TF-IDF*. Uma vez finalizado o processo de pré-processamento, o artefato que resulta deste processo está pronto para ser utilizado na mineração de texto. A ordem das etapas pode ser visualizada na Figura 2.

Figura 2 – Etapas de pré-processamento de texto



Fonte: Adaptado de Vijayarani et al. (2015)

2.4.1 Tokenização

Para Vijayarani, Janani et al. (2016), tokenização é o processo de quebrar um fluxo de conteúdos textuais em palavras, termos, símbolos, ou outro elemento que possua um significado, denominados *tokens*. De forma semelhante, Jo (2019) diz que a tokenização consiste no processo de transformar um ou mais textos em *tokens*, separando-os por espaços em branco, ou pontuações.

Figura 3 – Processo de tokenização

Fonte: Adaptado de Vijayarani, Janani et al. (2016)

2.4.2 Remoção de *stop-words*

Stop-words são uma divisão da linguagem natural e irrelevantes para o contexto, elas podem ser removidas para facilitar a análise e melhorar a eficiência dos algoritmos de mineração de texto (VIJAYARANI et al., 2015; JO, 2019).

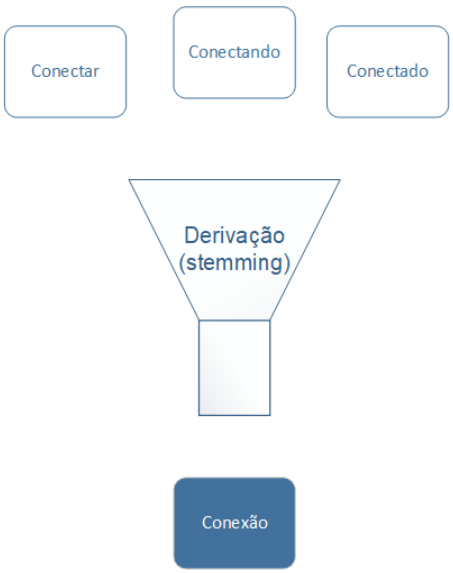
O processo de remoção das *stop-words* ocorre da seguinte maneira: a lista de *stop-words* é preparada como um arquivo que é carregado, e para cada palavra neste arquivo,

esta é removida da lista de *tokens*, obtida na etapa anterior. O que resulta normalmente é uma lista de nomes, verbos e adjetivos (JO, 2019).

2.4.3 Derivação (*Stemming*)

Derivação refere-se ao processo de mapear cada token resultante dos processos anteriores, em sua forma original, ou base (JO, 2019). Por exemplo, as palavras conectar, conectado, conectando, podem ser derivadas para a palavra conexão. O propósito deste método é remover os prefixos e sufixos das palavras, bem como reduzir o número total de palavras (VIJAYARANI et al., 2015).

Figura 4 – Processo de derivação (*stemming*)



Fonte: Adaptado de Vijayarani et al. (2015)

2.4.4 Ponderação dos termos

O processo de ponderação dos termos tem por objetivo calcular e atribuir o peso de cada palavra como grau de importância (JO, 2019). Vijayarani et al. (2015) explora em seu trabalho a técnica *Term Frequency-Inverse Document Frequency* (TF-IDF), uma estatística numérica que revela o quão importante é uma palavra em um documento. O valor do TF-IDF aumenta proporcionalmente ao número de vezes que uma palavra aparece em um documento, e decresce proporcionalmente quando a palavra aparece no conjunto de documentos como um todo, também denominado *corpus* (JO, 2019).

De acordo com Feldman (2006), o peso TF-IDF, de uma palavra, w, em um documento, d, é calculada pela Equação 2.1:

$$\text{TF-IDF}\,(w,d) = FrequenciaTermo\,(w,d) \times log\left(\frac{N}{FrequenciaDoc\,(w)}\right) \tag{2.1}$$

Onde $FrequenciaTermo\,(w,d)$ é o número de aparições da palavra w no documento d, N é o número total de documentos e $FrequenciaDoc(w)$ é o número de documentos que contém a palavra w (FELDMAN, 2006).

2.5 Modelagem de tópicos

A modelagem de tópicos refere-se ao processo de segmentar um grupo de documentos em subgrupos de documentos similares através de um modelo de tópicos (JO, 2019; SRIVASTAVA, 2009). Este processo se encaixa na sétima área de Miner (2012), extração de conceitos.

Um modelo de tópicos é um modelo probabilístico que tem como objetivo revelar a estrutura semântica de um *corpus*, baseando-se em uma análise hierárquica Bayesiana dos documentos originais (SRIVASTAVA, 2009). Um modelo não consegue entender a semântica de cada palavra que compõe um documento, ao invés disso ele supõe que qualquer parte do texto é possível de ser combinada selecionando palavras de grupos prováveis de palavras, onde cada grupo corresponde a um tópico. Um tópico então, é uma lista de palavras estatisticamente significantes (JELODAR et al., 2019).

O modelo é capaz de descobrir padrões de uso de palavras e como elas se relacionam entre os documentos que possuem padrões similares, por isso a técnica tem sido cada vez mais utilizada em mineração de texto (SRIVASTAVA, 2009).

2.5.1 *Latent Dirichlet Allocation*

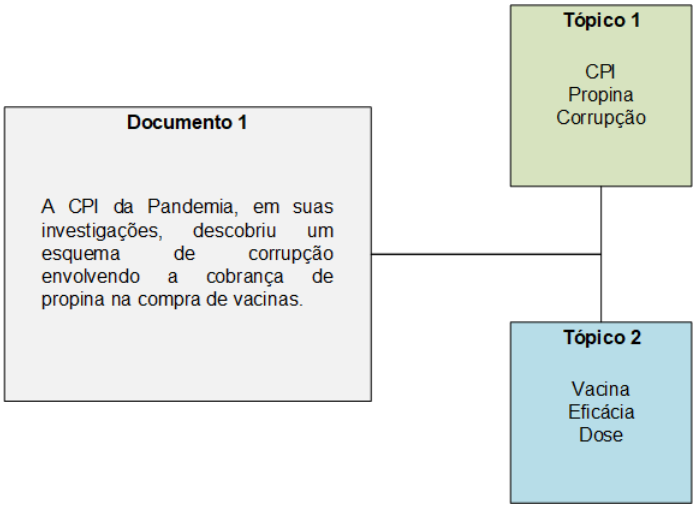
O *Latent Dirichlet Allocation* (LDA) é um modelo probabilístico generativo e Bayesiano hierárquico de três níveis para coleções de dados discretos, como um *corpus*, e foi introduzido por Blei, Ng e Jordan (2003) em 2003. De acordo com Jelodar et al. (2019), o modelo LDA considera que um documento é formado por combinações de tópicos latentes, onde cada tópico é caracterizado por uma distribuição sobre palavras, conforme a Figura 5.

Ainda de acordo com Blei, Ng e Jordan (2003) e Jelodar et al. (2019), considerando um *corpus* D, composto por M documentos, com o documento $d\,(d \in 1, ..., M)$ contendo N_d palavras, o LDA é capaz de modelar D, através do seguinte processo generativo:

1. Escolha uma distribuição multinomial ϕ_t para um tópico t, através de uma distribuição de *Dirichlet*[1], com parâmetro β.

[1] Mais informações sobre a derivação e as especificações da distribuição *Dirichlet* podem ser encontradas

Figura 5 – Modelo LDA



2. Escolha uma distribuição multinomial θ_d para um documento d, através de uma distribuição de *Dirichlet*, com parâmetro α.

3. Assim, para cada palavra w_n $(n \in 1, ..., N_d)$ no documento d, selecione um tópico z_n da distribuição θ_d e uma palavra w_n de $p(w_n|z_n, \beta)$, uma distribuição multinomial condicionada no tópico z_n.

Jelodar et al. (2019) argumenta que algumas palavras nos documentos são variáveis observadas, enquanto outras palavras são variáveis latentes (ϕ e θ) e hiperparâmetros do modelo (α e β). Bartholomew, Knott e Moustaki (2011) define que variáveis latentes são as variáveis aleatórias de um modelo estatístico que não são observáveis, que estão ocultas.

Para inferir as variáveis latentes e os hiper parâmetros, a distribuição de probabilidade de um *corpus* D é computada e maximizada seguindo a Equação 2.2:

$$p\left(D|\alpha,\beta\right) = \prod_{d=1}^{M} \int p\left(\theta_d|\alpha\right) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p\left(z_{dn}|\theta_d\right) p\left(w_{dn}|z_{dn},\beta\right) \right) d\theta_d \tag{2.2}$$

O LDA será empregado neste trabalho com o mesmo objetivo que descreve Srivastava (2009), descobrir tópicos correlacionados em um conjunto de *tweets*.

2.6 Trabalhos relacionados

Matos, Dourado e Mesquita (2017) examinou o uso do perfil da ex-presidenta Dilma Roussef (@dilmabr) durante seu processo de Impeachment. Para tal, foi realizada

em Lin (2016).

uma análise quantitativa relacionada à frequência de postagens e repostagens (*retweets*), relacionando-as com as etapas-chave do processo. Seu trabalho também analisou as *hashtags* mencionadas pelo perfil neste período, e por fim realizou uma categorização do conteúdo publicado. Sua pesquisa verificou que o conteúdo publicado pelo perfil da ex-presidenta no Twitter faz criticas ao processo de Impeachment, às empresas tradicionais de jornalismo e aos atores políticos envolvidos, além de prestar esclarecimentos sobre as causas da abertura do processo.

O trabalho de Malagoli et al. (2021) coletou mais de 9 milhões de *tweets* em português que citasse termos relacionados à vacinação, entre os meses de dezembro de 2020 e janeiro de 2021, período que marca o inicio da vacinação no Brasil, com o objetivo de caracterizar o debate sobre a vacinação contra COVID-19 no Brasil. Sua pesquisa relacionou os períodos de maior intensidade de postagens com os eventos que ocorreram, e concluiu que existiu uma forte correlação entre os eventos e os picos de postagem no período de coleta. Além disso, seu trabalho também realizou uma análise de sentimentos através da ferramenta *SentiStrenght*, e concluiu que existiu um sentimento negativo nas palavras *vacinação* e *vacina*, apesar de alguns termos específicos relacionados à vacinação, como *novavax*, *moderna* e *sputnikv* não terem o sentimento negativo tão evidente. Em uma análise de correlação, o trabalho mostra que as palavras relacionadas à vacinação, como *coronavac*, *vacina*, *pfizer* coocorrem com frequência nos *tweets* coletados.

O trabalho de Oliveira et al. (2018) coletou mais de 750 mil *tweets* de 692 deputados brasileiros, de outubro de 2013 até outubro de 2015 e construiu uma metodologia de classificação de *tweets*, sendo possível classifica-los como políticos e não-políticos, a classificação construída é utilizada na caracterização do conjunto de *tweets* coletados. Dentre outros modelos, os autores treinaram uma *Convolutional Neural Network* (CNN), para classificar os textos coletados, a CNN superou outros modelos como, *Long short-term memory* (LSTM) e *FastText*, e alcançou um *F1 score* de 99% no conjunto de dados de treinamento, e 97% nos dados de teste. Os autores identificaram também que o número de publicações dos deputados aumentou no período próximo à data das eleições de 2014. Por fim, o autor identifica que aproximadamente 40% dos deputados mudaram o comportamento de publicações, 24% postaram mais *tweets* políticos e 16% postaram menos conteúdo com essa característica.

Amaral e Pinho (2018) coletou *tweets* de parlamentares antes e depois das eleições de 2014, totalizando 897 entre ex-deputados, deputados novatos e reeleitos, entre dezembro de 2013 e abril de 2015. O autor mostra que a partir de março de 2009 os políticos brasileiros começaram a adotar o Twitter, e que em 2015, havia mais de 90% de adoção à plataforma. Os autores também mostram que, curiosamente, o uso da plataforma diminuiu no período pré-eleitoral, teve um pico durante as eleições e no período pós-eleitoral assumiu níveis menores do que antes das eleições, mostrando que eventos externos à plataforma

influenciam no volume de postagens.

De forma geral, todos os trabalhos relacionados abordados nesta Seção fizeram uma análise do conteúdo publicado no Twitter, além disso todos eles caracterizaram os *tweets* que foram coletados. Assim como Matos, Dourado e Mesquita (2017), Amaral e Pinho (2018) e Malagoli et al. (2021), que relacionaram os períodos de maior postagem com os eventos que ocorreram na vida real, o presente trabalho busca relacionar os períodos de maior e menor volume de publicações com os eventos que ocorreram na CPI da Pandemia.

Assim como o trabalho de Matos, Dourado e Mesquita (2017), este trabalho busca utilizar as *hashtags* com o objetivo de aprofundar a caracterização do *tweets* coletados. Amaral e Pinho (2018) e Oliveira et al. (2018) também contribui para este trabalho mostrando que cada vez mais políticos estão utilizando o Twitter para se comunicar com seu eleitorado, especialmente próximos a eventos políticos relevantes, como as eleições.

3 Metodologia

Neste capítulo é apresentada a metodologia utilizada na pesquisa. A Seção 3.1 aborda as principais tecnologias utilizadas no desenvolvimento do trabalho. A Seção 3.2 explica como foi realizada a coleta dos *tweets*. A Seção 3.3 mostra as etapas de pré-processamento dos *tweets* coletados para construir a base de dados final, utilizada para elaborar os resultados. A Seção 3.4 faz uma caracterização das *hashtags* encontradas na base de dados. Por fim, a Seção 3.5 realiza uma caracterização da base de dados.

3.1 Tecnologias

Para a realização deste trabalho, foram utilizadas algumas tecnologias específicas, que serão apresentadas nesta seção.

3.1.1 API do Twitter

A *Application Programming Interface* (API) que o Twitter oferece para consultar os dados históricos da plataforma não possui acesso aberto para o público em geral. Para obter uma permissão de acesso à API, é necessário realizar um cadastro extensivo e informar quais as intenções do utilizador ao consumir a API[1].

Uma vez que o cadastro é realizado e aprovado pelo Twitter, o utilizador recebe um código de uso da API, este código é denominado *bearer token*, e ele é comumente utilizado para realizar autenticações em APIs protegidas pelo protocolo *OAuth 2.0* (JONES; HARDT, 2012). Uma vez em posse do *bearer token*, é possível realizar chamadas à API do Twitter.

3.1.2 *Python*

O Python é uma linguagem de programação interpretada de alto nível que é utilizada em grande parte dos trabalhos descritos na seção 2.6. Desde sua criação, em 1991, o Python se tornou amplamente popular, junto com Perl, Ruby e outras. Essas linguagens são frequentemente chamadas de linguagens de *scripting*, pois permitem que o programador escreva pequenos programas ou *scripts* para automatizar tarefas.

Ao longo dos últimos anos, o Python construiu uma ampla comunidade de computação científica e análise de dados, e hoje é uma das linguagens mais relevantes para a análise de dados, principalmente pela grande variedade de bibliotecas disponíveis para essa

[1] A documentação da API do Twitter pode ser encontrada em https://developer.twitter.com/en.

tarefa (MCKINNEY, 2012; LUBANOVIC, 2014). Neste trabalho, o Python é utilizado para automatizar as coletas de *tweets* realizadas no Capítulo 3, bem como a construção dos gráficos e resultados discutidos no Capítulo 4.

3.1.2.1 NumPy

NumPy é uma biblioteca do Python, que fornece ao usuário que a utiliza uma variedade de estruturas de dados e algoritmos necessários para realizar diversas tarefas de análise de dados. De acordo com McKinney (2012), em sua implementação, o NumPy contém entre outras coisas:

- Uma implementação rápida e multidimensional de um vetor, chamado *ndarray*;
- Funções que auxiliam na operação elemento a elemento dentro desses vetores e operações matemáticas entre vetores, ambos executam muito rápido por serem extremamente otimizados;
- Funções de álgebra linear, transformada de Fourier, e até geração aleatória de números.

3.1.2.2 Pandas

O Pandas, assim como o NumPy, é uma biblioteca do Python desenvolvida para fornecer uma estrutura de dados robusta para trabalhar com dados tabulares, como uma planilha de Excel. A principal estrutura do Pandas se chama *DataFrame*, e é responsável por representar uma tabela em memória. McKinney (2012) argumenta que a biblioteca Pandas reúne a eficiência do NumPy com a facilidade de manuseio de dados de uma planilha, ou um banco de dados relacional.

3.1.2.3 Matplotlib

A biblioteca Matplotlib é utilizada neste trabalho para a construção de gráficos e representações de dados advindos de *DataFrames* da biblioteca Pandas, as duas bibliotecas se integram de forma fluida, o que permite uma construção fácil de diversos tipos de representação de dados que estejam sendo representados de forma tabular no Pandas.

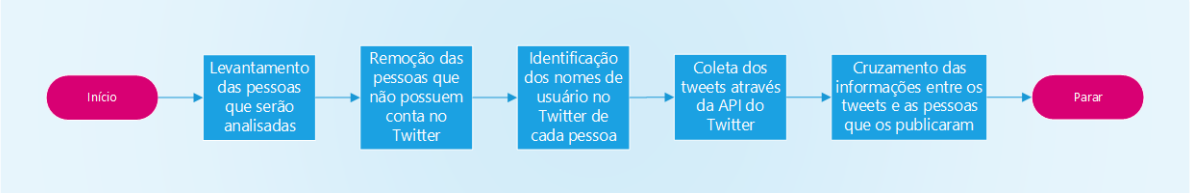
3.1.3 *UpSet Plot*

A biblioteca *UpSet Plot* foi desenvolvida por Lex et al. (2014), e possui o objetivo de resolver o problema de visualização de intersecções de conjuntos, quando existem mais de três conjuntos sendo analisados simultâneamente. O *UpSet* visualiza intersecções de conjuntos de dados em um formato matricial, através da agregação e agrupamento de cada ponto de dados.

3.2 Coleta de dados

O objetivo da coleta de dados neste trabalho é buscar e organizar todos os *tweets* que foram postados pelos senadores, governadores, convidados e testemunhas que participaram da CPI da Pandemia no Senado Federal. Apesar ter sido protocolada em 15/01/2021 e instalada somente no dia 24/04/2021, a CPI começa a aparecer no noticiário nacional a partir de uma determinação do Supremo Tribunal Federal (STF), através do ministro Luís Roberto Barroso, no dia 08/04/2021 (FALCÃO, 2021; BRASIL, 2021). O período de coleta destes dados, será então, a partir de 08/04/2021 até 05/11/2021, prazo final prorrogado da CPI (SENADO, 2021e). A Figura 6 ilustra através de um fluxograma a metodologia de coleta de dados.

Figura 6 – Fluxograma da coleta de dados



Uma vez identificados todos os usuários que terão seus *tweets* coletados, a metodologia de coleta para cada usuário é a seguinte:

1. Obter o nome de usuário, no Twitter desta pessoa, manualmente;
2. Obter o identificador de usuário deste usuário, através da API;
3. Obter os *tweets* do usuário, no período de coleta, através da API.

3.2.1 Identificação dos usuários da coleta

Para realizar a coleta, inicialmente foram armazenados todos os governadores e senadores, além de convidados e testemunhas da CPI da Pandemia. Tais informações foram recuperadas da página oficial do Senado Federal Senado (2021e), Caesar (2018) e Senado (2022).

A planilha construída neste trabalho possui três abas, uma para cada perfil de pessoa analisada: convidados e testemunhas, senadores e governadores. A divisão da planilha em abas será discutida na Seção 3.3.

O próximo passo na coleta foi identificar o nome de usuário no Twitter de cada pessoa da planilha. O nome de usuário identifica um perfil no Twitter de forma única, prefixado pelo símbolo @, e pode ser diferente do nome do usuário. A Figura 7 ilustra a diferença entre o nome de usuário e o nome do indivíduo.

Figura 7 – Identificando o nome de um usuário no Twitter

A coleta de todos os nomes de usuário foi realizada de forma manual pelo autor, pois alguns usuários não possuíam contas no Twitter, ou possuíam perfis privados, o que impede a coleta via API, a não ser que a conta da API seja aceita como seguidor pelo usuário dono do perfil privado. Por exemplo, a conta do atual presidente do Senado, Rodrigo Pacheco, está privado e, apesar de tentativas de solicitação para seguir o perfil, não foi possível adiciona-lo ao conjunto de dados coletado.

Alguns convidados e testemunhas que compareceram à CPI não possuíam, no momento de coleta dos dados, uma conta no Twitter, como por exemplo, o Ex-ministro da Saúde, o general Eduardo Pazuello. A Tabela 1 informa o nome, a data de comparecimento à CPI e o cargo ocupado por cada convidado e testemunha que não tinha uma conta no Twitter no momento de coleta dos dados.

3.2.2 Transformação do nome de usuário em um identificador de usuário

O endpoint da API do Twitter que permite a coleta de *tweets* requer um parâmetro desconhecido pelo público em geral, o identificador de usuário (*user id*). Este parâmetro é oculto, composto por números inteiros e representa de forma única um usuário do Twitter. Para cada pessoa que será analisada neste trabalho, será necessário obter o seu identificador de usuário antes de coletar seus *tweets*.

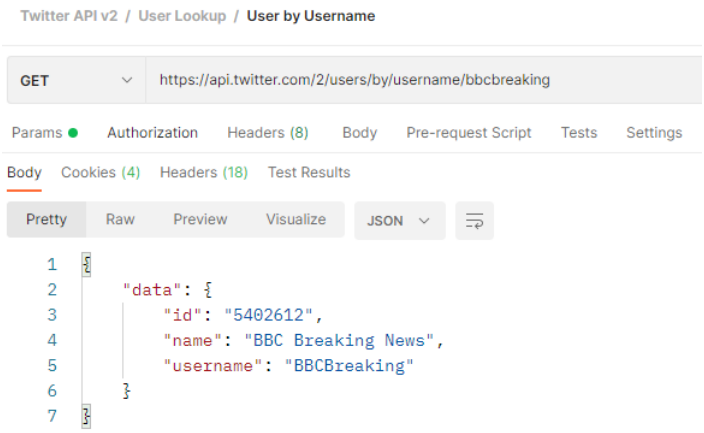
Para obter este identificador único, é necessário utilizar a API do Twitter, através do recurso *User by Username*, conforme a Figura 8. Este recurso permite enviar um nome de usuário, por exemplo *@BBCBreaking*, e obter como retorno, um *user id*. A documentação especifica que somente o nome de usuário, sem o símbolo @, deve ser enviado como parâmetro. A fim de automatizar o processo de transformação de um nome de usuário em identificador de usuário, foi desenvolvido um algoritmo simples em *Python*. O algoritmo executa as seguintes ações: para cada usuário na planilha, busque seu identificador de usuário através de uma chamada na API do Twitter, de forma semelhante à Figura 8 e atribua à uma nova coluna da planilha.

Tabela 1 – Relação de convidados e testemunhas sem uma conta no Twitter

Nome	**Data**	**Cargo ocupado**
Antonio Barra Torres	11/05/2021	Diretor-presidente da Agência Nacional de Vigilância Sanitária (Anvisa)
Carlos Murillo	13/05/2021	Presidente Regional da Pfizer na América Latina e ex-gerente-geral e representante da empresa no Brasil
Eduardo Pazuello	19/05/2021 e 20/05/2021	Ex-ministro da Saúde
Dimas Covas	27/05/2021	Diretor do Instituto Butantan
Luana Araújo	02/06/2021	Ex-secretária Extraordinária de Enfrentamento à Covid
Marcelo Queiroga	08/06/2021	Ministro da Saúde
Alexandre Marques	17/06/2021	Ex-auditor do Tribunal de Contas da União
Paulo Baraúna	18/06/2021	Diretor da White Martins
Francisco Eduardo Cardoso Alves	18/06/2021	Médico
Francisco Emerson Maximiano	23/06/2021	Sócio da Precisa Medicamentos
Luís Ricardo Fernandes Miranda	25/06/2021	Chefe de importação do Departamento de Logística em Saúde do Ministério da Saúde
Luiz Paulo Dominguetti Pereira	01/07/2021	Representante da Davati Medical Supply
Roberto Ferreira Dias	07/07/2021	Ex-diretor de Logística do Ministério da Saúde

Fonte: Adaptado de Senado (2021e)

Figura 8 – Obtendo o identificador de um usuário



3.2.3 Coleta dos *tweets*

De posse de todos os identificadores de usuário, é possível realizar a coleta de *tweets* de um usuário de interesse através do *endpoint User Tweet timeline by ID* da API. De acordo com Twitter (2021), este *endpoint* é reponsável por buscar os *tweets* escritos por um determinado usuário, considerando que seja fornecido como parâmetro o identificador do usuário de interesse. O retorno desta chamada, porém, não retorna todos os *tweets* de

interesse de uma só vez, este *endpoint* faz uso de paginação para dividir o resultado em até 100 *tweets* de cada chamada. Somente os 3200 *tweets* mais recentes de um usuário podem ser adquiridos por este *endpoint*.

Ao chamar a API são retornados dois dados principais: um conjunto de *tweets* e metadados da requisição. Para coletar todos os *tweets* disponíveis de um usuário é necessário realizar primeiramente uma chamada inicial à API para receber os primeiros 100 *tweets*, e os metadados. Nos metadados, existe um código, chamado *next_token*, que, quando anexado à próxima solicitação, permite coletar os próximos 100 tweets, num processo de janela deslizante. A coleta de *tweets* para um usuário, portanto, é baseada em iterações de chamadas ao *endpoint* de linha do tempo, até que todos os *tweets* publicados no período de coleta tenham sido coletado, ou somem 3200 *tweets*, por usuário, no total.

De acordo com Twitter (2021), é possível adicionar informações adicionais ao conteúdo de cada *tweet* retornado pela API. Neste trabalho, para cada *tweet* da base de dados, será incluído o número de curtidas (*likes*), repostagens (*retweets*), respostas (*replies)*, a data de postagem e o número de citações (*quotes*). Nas chamadas desta API é informada a data de início e fim da coleta de dados.

Para agregar os dados coletados de todos os usuários de interesse do trabalho em um único arquivo, todas as requisições realizadas foram adicionadas em uma lista, e esta foi salva em um arquivo no formato *.json* sem nenhum processamento das requisições. O arquivo sem processamento, contendo todas as requisições coletadas no período de análise deste trabalho pode ser encontrada em Libardi (2022b).

3.3 Pré-processamento

Ao realizar a coleta na seção anterior, este trabalho criou duas fontes de dados: uma planilha com informações sobre as pessoas que serão analisadas, e um arquivo contendo todos os *tweets* que foram coletados dessas pessoas. O próximo passo na construção do conjunto de dados é realizar um cruzamento de informações dessas duas fontes e montar um conjunto de dados único, onde as análises serão realizadas.

3.3.1 Organização das pessoas analisadas

Para realizar o cruzamento de dados, cada aba da planilha de informações das pessoas analisadas é lida por um *script* de forma separada, e salva em uma lista, no *Python*. A lista de convidados e testemunhas é filtrada para possuir somente pessoas que possuam um nome de usuário no Twitter. Sendo assim, pessoas que não possuem um perfil no Twitter foram removidas da análise.

Para cada uma das listas construídas, quatro atributos são adicionados à base de

dados principal: identificador do usuário no Twitter, o nome da pessoa, a função da pessoa no contexto do trabalho e o partido político da pessoa. Caso a pessoa pertença ao grupo de convidados e testemunhas, o partido político é definido como "Convidado/Testemunha" de forma fixa.

3.3.2 Relacionamento entre os *tweets* e as pessoas de interesse

Para a construir a base de dados, foi realizado o cruzamento de informações das pessoas com as requisições coletadas anteriormente. Os dados se apresentam da seguinte forma na base de dados:

- Texto do *tweet*;
- Nome da pessoa;
- Função da pessoa (Senador, governador ou convidado/testemunha);
- Data de publicação do *tweet*;
- Partido político da pessoa;
- Número de re-postagens do *tweet* (*retweets*);
- Número de respostas do *tweet* (*replies*);
- Número de curtidas do *tweet* (*likes*);
- Número de citações do *tweet* (*quotes*).

3.4 Categorização das *hashtags*

As etiquetas, ou *tags*, são uma definição de uma categoria, ou palavra-chave, e são utilizadas na *web* para classificação de textos (TONKIN; PFEIFFER; TOURTE, 2012). No Twitter, assim como em outras redes sociais como Facebook e Instagram, as *tags* são comumente chamadas de *hashtags*. De acordo com Cunha (2012), uma *hashtag* é qualquer conteúdo textual precedido do símbolo de cerquilha (#), e pode ser criado a qualquer momento por qualquer usuário, simplesmente fazendo menção a essa *hashtag* em um texto arbitrário publicado no Twitter.

Em seu trabalho, Viana et al. (2019) argumenta que as *hashtags* possuem cada vez mais importância na organização do texto publicado em redes sociais, tendo em vista que muitas campanhas publicitárias são agrupadas em torno de algumas *hashtags* específicas. Neste trabalho, é realizado um agrupamento dos *tweets* coletados em torno das *hashtags* que são referenciadas em cada *tweet*.

3.4.1 Metodologia de categorização

Para realizar a categorização das *hashtags*, foi necessário realizar um agrupamento dos *tweets* considerando o conteúdo textual de cada um. É importante, para o contexto deste trabalho, que as *hashtags* sejam agrupadas desconsiderando as letras maiúsculas de minúsculas. A metodologia de categorização aplicada foi:

1. Para cada *tweet* coletado, transforme-o para ter somente letras minúsculas;
2. Construa uma expressão regular que encontre uma *hashtag* em um *tweet*, considerando acentos e números;
3. Utilize a interface *Counter*, do Python, para agregar os resultados da expressão regular construída no passo anterior;
4. Transforme o resultado do *Counter* em um dicionário do Python.

Foram identificadas no total 4668 *hashtags* nos *tweets* coletados. O ranking das vinte principais *hashtags* encontradas pode ser visualizado na Tabela 2, que relaciona cada *hashtag* com o total de menções.

Tabela 2 – Ranking das 20 *hashtags* mais utilizadas

Hashtag	**Total de menções**
cpidacovid	1289
equipehb	627
covid19	619
equipefb	535
cpidapandemia	458
paraíbavacina	387
usesempremáscara	370
ocuidadocontinua	357
senadofederal	317
senadorconfuciomoura	231
forabolsonaro	223
bahia	220
tbt	208
mdb	208
brasil	203
rondonia	202
wevertonsenador	197
amazonas	186
agenda	181
senadorelmanoferrer	175

Algumas *hashtags* não apresentam significado claro, mas pesquisando cada uma no próprio Twitter foi possível esclarecer que a *hashtag* “equipehb” refere-se ao governador Helder Barbalho, enquanto a *hashtag* “equipefb” está presente nos *tweets* da Governadora Fátima Bezerra. É importante destacar que nas vinte *hashtags* mais utilizadas, cinco

correspondem ao nome dos senadores que as postaram, portanto, 25% das *hashtags* identificadas.

A partir de todas as *hashtags* coletadas, principalmente as que estão presentes na Tabela 2, foram identificadas quatro categorias (assuntos) principais de *hashtag* que estão presentes: CPI da Pandemia, vacinação, utilização de máscara e COVID-19. Optou-se por realizar a categorização de forma manual, pois nem todas as *hashtags* possuem um significado explícito em seu nome.

Todas as 4668 *hashtags* identificadas foram categorizadas de acordo com o assunto que cada uma representa. A partir dessa categorização, uma nova etapa de rotulação classificou todas as *hashtags* que não pertençam às categorias principais como "Não relacionada", com exceção das *hashtags* presentes na Tabela 2. A Tabela 3 relaciona as vinte primeiras *hashtags* com suas respectivas categorias. As *hashtags* que não foram agrupadas em uma das quatro categorias principais, mas que pertençam ao ranking das vinte *hashtags* mais utilizadas serão consideradas para análise no Capítulo 4.

Tabela 3 – Categorização das 20 principais *hashtags*

Hashtag	**Total de menções**	**Categoria**
cpidacovid	1289	CPI da Pandemia
equipehb	627	Equipe HB
covid19	619	Covid
equipefb	535	Equipe FB
cpidapandemia	458	CPI da Pandemia
paraíbavacina	387	Vacinação
usesempremáscara	370	Máscara
ocuidadocontinua	357	Covid
senadofederal	317	CPI da Pandemia
senadorconfuciomoura	231	Confucio Moura
forabolsonaro	223	Fora Bolsonaro
bahia	220	Bahia
tbt	208	Tbt
mdb	208	Mdb
brasil	203	Brasil
rondonia	202	Rondonia
wevertonsenador	197	Weverton Senador
amazonas	186	Amazonas
agenda	181	Agenda
senadorelmanoferrer	175	Senador Elmano Ferrer

3.5 Caracterização do conjunto de dados

A base de dados transformada é composta por 79.280 *tweets*, sendo 22.328 *tweets* postados pelos governadores, representando aproximadamente 28% do total, 43.948 *tweets* foram postados pelos senadores, representando aproximadamente 55% do total, e 13.004 *tweets* foram postados pelos convidados e testemunhas, representando aproximadamente 16% do total de *tweets* coletados no período de 08/04/2021 até 05/11/2021. A base de

dados é composta dos atributos descritos pela Tabela 4 que relaciona o nome de cada atributo com o tipo de dado que representa o atributo na base de dados. Os dados coletados podem ser encontrados em Libardi (2022a).

Tabela 4 – Descrição dos atributos do conjunto de dados coletado

Nome do atributo	**Tipo de dado**
Texto	Textual
Usuário	Textual
Função	Categórico
Data de postagem	Data
Partido	Categórico
Retweets	Numérico
Respostas	Numérico
Curtidas	Numérico
Citações	Numérico

Na Tabela 4 o atributo texto representa o conteúdo textual do *tweet*; usuário identifica quem postou o *tweet* com nome e sobrenome, e é diferente do nome de usuário; função descreve qual foi o papel da pessoa no contexto deste trabalho, e deve assumir um dos três valores: senador, governador ou Convidado/Testemunha. A data de postagem identifica quando o *tweet* foi postado, com data e hora; partido identifica qual partido político da pessoa que postou o *tweet*, convidados e testemunhas foram categorizados com o valor "Convidado/Testemunha", e políticos sem partido assumem o valor "Sem partido". *Retweets* representa o número de vezes que o *tweet* foi compartilhado através de uma repostagem do usuário que *retweetou* em sua própria linha do tempo; o atributo respostas indica quantas respostas um determinado *tweet* obteve; de forma semelhante, curtidas e citações indicam esses respectivos números; e o atributo sentimento indica qual o resultado da análise de sentimento para o texto daquele *tweet*, e deve assumir uma das três categorias: positivo, negativo e neutro.

Tabela 5 – Caracterização dos atributos numéricos do conjunto de dados

	Retweets	**Resposta**	**Curtidas**	**Citações**
Média	178.14	66.61	691.74	11.02
Desvio padrão	854.45	318.72	2793.50	170.39
Coef. Variação	4.79	4.78	4.03	15.46
Mínimo	0.00	0.00	0.00	0.00
1º quartil	1.00	0.00	1.00	0.00
2º quartil	5.00	2.00	20.00	0.00
3º quartil	45.00	16.00	193.00	2.00
Máximo	65948.00	12329.00	124638.00	32497.00

Os atributos numéricos da base de dados podem ser caracterizados estatisticamente de acordo com a Tabela 5, que mostra os valores máximo e mínimo, os quartis, a média, o desvio padrão e o Coeficiente de Variação (CV), que é calculado através da razão entre o desvio padrão e a média de cada atributo. É possível perceber que dentre os *tweets* coletados, os valores máximos são ordens de grandeza maiores do que a média, para cada

atributo. A interpretação é que existem alguns poucos *tweets* que viralizaram e atingiram valores muito altos de *retweets*, curtidas, respostas e citações.

Os *tweets* coletados citam frequentemente termos relacionados ao objeto de estudo desse trabalho: CPI da Pandemia, COVID-19, vacinação e o uso de máscara. Essa relação pode ser visualizada através da nuvem de palavras ilustrada na Figura 9.

Figura 9 – Nuvem de palavras para os 50 termos mais frequentes

3.5.1 Distribuição dos *tweets* no período de coleta

A Figura 10 mostra o total de *tweets* publicados por dia, ao longo da CPI. É possível perceber um comportamento oscilatório de alta frequência. Para entender essa oscilação, a Figura 11 mostra a variação da frequência das postagens ao longo dos dias de uma semana. A oscilação ocorre, portanto, pois existe uma menor frequência de postagem de *tweets* nos fins de semana.

Na Figura 11 existem dois dias que se destacam dos demais, são esses o *outlier* superior da quinta-feira, que ocorreu no dia 30 de setembro de 2021, e o *outlier* inferior da terça-feira, que ocorreu em 13 de abril de 2021. De acordo com Senado (2021e), em 30 de setembro de 2021 esteve presente na CPI da Pandemia Otávio Fakhoury, empresário que segundo G1 (2021b), teve seu sigilo telefônico quebrado pelos parlamentares em agosto de 2021 e conseguiu o direito de permanecer em silêncio durante a sessão. Já no dia 13 de abril de 2021, data anterior à instalação da CPI, segundo G1 (2021c), o presidente do Senado Rodrigo Pacheco oficializou a abertura da CPI da Pandemia.

Figura 10 – Distribuição do número de *tweets* no período da CPI

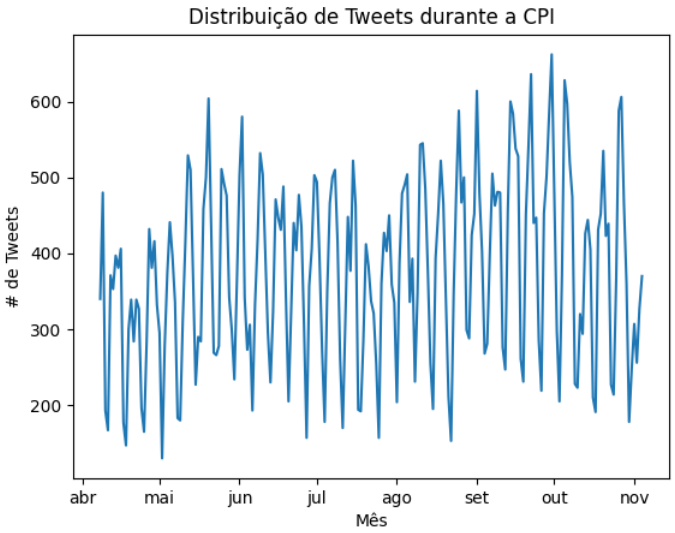
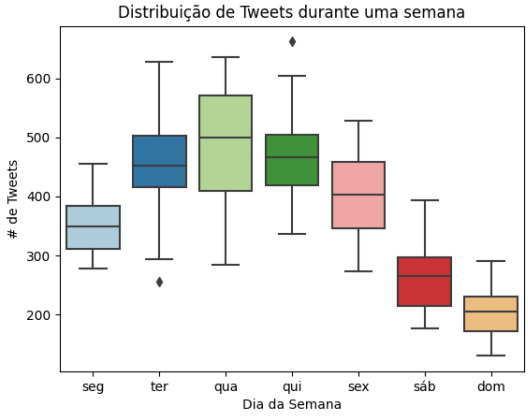


Figura 11 – Distribuição de *tweets* em uma semana



4 Resultados

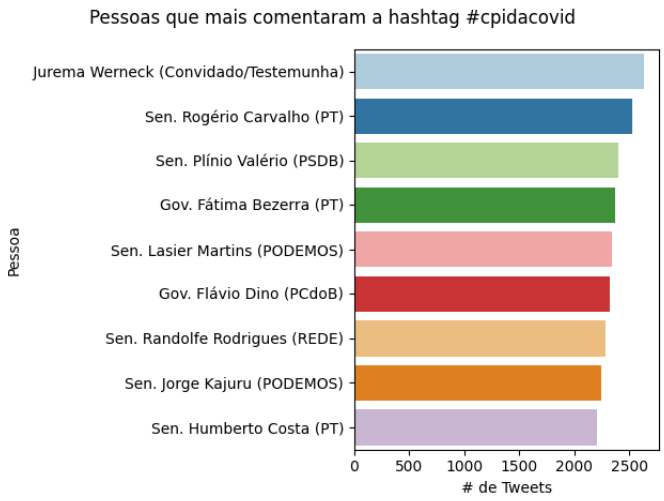
Neste capítulo serão apresentados e discutidos os resultados encontrados. A Seção 4.1 analisa as principais *hashtags* identificadas na coleta de dados e verifica como elas se relacionam entre si. A Seção 4.2 apresenta os tópicos descobertos dentro de cada uma das quatro categorias principais, além de identificar as principais pessoas que publicaram em cada categoria. A Seção 4.3 relaciona os períodos de maior e menor intensidade de postagens com os eventos que ocorreram na CPI da Pandemia.

4.1 Análise das principais *hashtags*

A partir da Tabela 3, é possível perceber que existem quatro categorias que se apresentam de forma mais frequente do que as demais, são elas: CPI da Pandemia, vacinação, covid e máscara. As *hashtags* que compõem as categorias principais possuem 6148 menções em *tweets* no total, $25,2\%$ do total de 24399 menções identificadas em todas as *hashtags*. Essas quatro categorias são as principais categorias de análise deste trabalho.

Ainda na Tabela 3, é possível observar que a categoria CPI da Pandemia obteve o maior número de utilização dentre as demais *hashtags* analisadas. A Figura 12 mostra que Jurema Werneck, convidada a depor na CPI no dia 23 de junho de 2021, foi a pessoa analisada que mais publicou *tweets* utilizando esta *hashtag*. Também é possível visualizar que 60% das pessoas que mais publicaram na *hashtag* #cpidapandemia são Senadores da República.

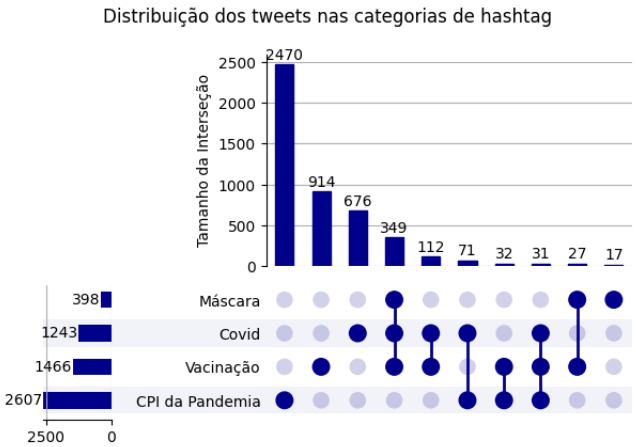
Figura 12 – Total de *tweets* publicados na *hashtag* #cpidacovid



Considerando apenas as principais categorias de *hashtags* (CPI da Pandemia, COVID-19, Máscara e Vacinação), foi feita uma análise para identificar se essas categorias foram utilizadas de forma conjunta pelos usuários em um mesmo *tweet* e com qual frequência. Através da visualização *UpSet Plot*, mostrada na Figura 13, é possível estimar o número de intersecções entre as quatro categorias principais de *hashtag*, ou seja, quantos *tweets* citam, no mesmo texto, *hashtags* que pertencem a categorias diferentes. Nesta visualização, o gráfico de barras superior ilustra o tamanho da intersecção entre as categorias. As categorias e suas intersecções estão identificados pelos círculos preenchidos respectivamente abaixo de cada barra superior. As barras horizontais, na frente de cada categoria, representam o total de vezes que ela foi identificada no conjunto de dados.

Analisando a Figura 13, percebe-se que geralmente as categorias são utilizadas de forma isolada, ou seja, cada *tweet*, de forma geral, utiliza apenas uma das quatro categorias principais. Apesar disso, de forma predominante dentre as demais intersecções de categoria, é interessante notar que as categorias: Máscara, Covid e Vacinação estão bem correlacionadas, apresentando ao todo 349 intersecções entre elas. Em relação à categoria CPI da Pandemia, a categoria Covid é a que mais se correlaciona, aparecendo juntas em 71 *tweets*. Esse resultado se deve ao fato das investigações da CPI terem como objetivo principa investigar as possíveis omissões do Governo Federal durante a Pandemia de COVID-19 (SENADO, 2021e).

Figura 13 – *UpSet Plot* para as 4 categorias principais



Existem *hashtags* que não se pode afirmar que estão diretamente relacionadas com nenhuma das quatro principais categorias, contudo, elas possuem uma alta frequência no conjunto de *tweets* coletados. Sendo assim, foi realizada uma análise de intersecção entre essas categorias consideradas secundárias, utilizando novamente a visualização através de *UpSet Plot*. A Figura 14 mostra as intersecções entre as categorias secundárias, entretanto, a categoria CPI da Pandemia também está presente nessa análise, e desta vez as intersecções

das categorias secundárias com ela terão a cor laranja.

Analisando a Figura 14, é possível perceber que nem todas as categorias secundárias se correlacionam com a categoria CPI da Pandemia de forma significativa. Mas das intersecções que existem, é fundamental perceber que o Senador Elmano Ferrer (PP) utilizou uma das *hashtags* da categoria CPI da Pandemia em seus *tweets* em 171 de 174 *tweets* que ele publicou no período analisado, representando assim 98,2% do total de publicações. De forma semelhante, o Senador Confúcio Moura (MDB) relacionou *hashtags* com o nome de seu partido e o estado onde exerce seu mandato, Rondônia, com a categoria CPI da Pandemia em 164 ocasiões de 231 *tweets* que publicou, representando assim aproximadamente 71% de suas publicações. O Senador Confúcio Moura também realizou 31 publicações associando *hashtags* com seu nome à categoria CPI da Pandemia, representando 13,4% de suas publicações. O Senador ainda utilizou o seu estado, Rondônia, juntamente com seu nome, associados à categoria CPI da Pandemia, com 19 *tweets*, representando assim 8,2% das publicações. Considerando todas as intersecções presentes na Figura 14, Confúcio Moura teve no total 92,6% de suas publicações relacionadas com a CPI da Pandemia.

Ainda na Figura 14 é possível perceber que existem 96 publicações que relacionam a *hashtag* #amazonas com a categoria CPI da Pandemia, e 70 publicações quando considera-se também a intersecção com a *hashtag* #brasil. Esse resultado ocorre, pois um dos objetivos da CPI também era investigar as causas da crise de oxigênio que ocorreu na cidade de Manaus em janeiro de 2021 (SENADO, 2021e).

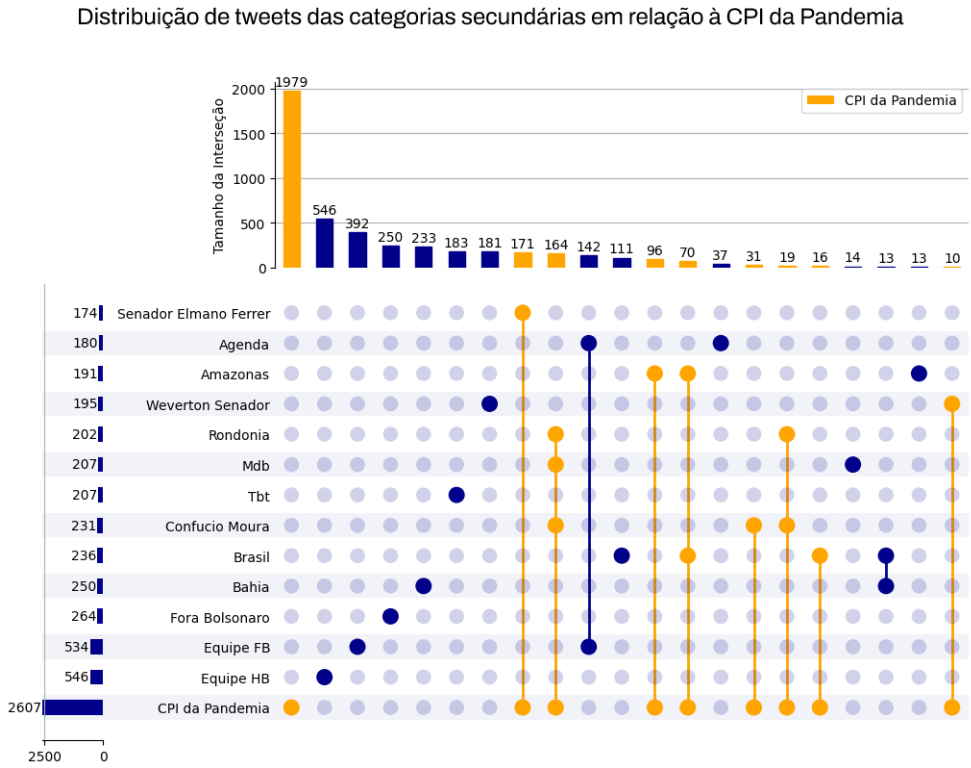
Na Figura 14 ainda é importante notar que categorias como Senador Limano Ferrer, Weverton Senador e Confucio Moura estão fortemente relacionadas à categoria CPI da Pandemia. Considerando também que os senadores realizaram a maioria das postagens no período de coleta, é possível concluir que uma parte dos senadores da República fizeram o uso do Twitter para falar de assuntos referentes à CPI da Pandemia, fazendo menções à *hashtags* que estão nessa categoria juntamente com menções à *hashtags* que citam o seu próprio nome.

4.2 Modelagem de tópicos

Para a modelagem de tópicos foi construído um modelo LDA para cada categoria de *hashtag*. A metodologia para escolher o número ideal de tópicos é realizar o cálculo de coerência para modelos cujo tamanho de tópicos varia entre 2 e 6. Para construir um modelo LDA é realizada uma etapa de pré-processamento cujo objetivo é adequar a base de dados para que o modelo funcione de forma correta.

No pré-processamento dos *tweets* que serão utilizados no LDA, primeiramente são removidos todas as *hashtags*, menções a outros usuários, URLs que possam estar no

Figura 14 – *UpSet Plot* para as categorias secundárias em relação à categoria CPI da Pandemia



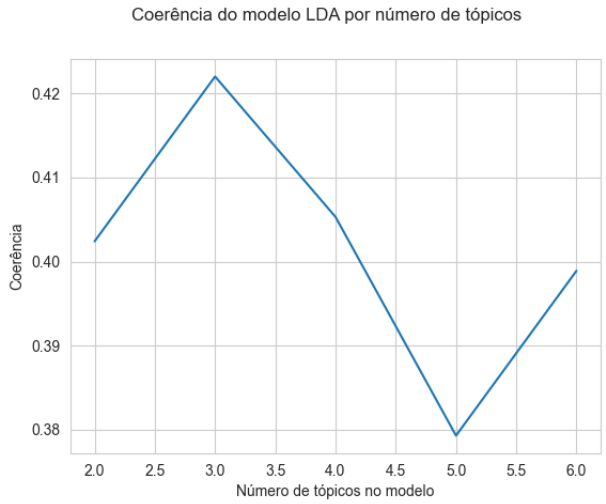
texto do *tweet* e as *stopwords*. Por fim todas as palavras são transformadas para letras minúsculas. A Figura 15 apresenta um gráfico com os valores calculados de coerência no eixo Y para cada modelo LDA construído com a quantidade de tópicos representada no eixo X. O número ideal de tópicos é 3, pois este é o maior valor de coerência calculada para cada modelo.

O modelo LDA possui dois argumentos de entrada, α e β. No modelo construído, utilizando a biblioteca *Gensim*[1], é possível calcular esses parâmetros de forma automática. Esta estimativa de parâmetros automática foi utilizada nos modelos LDA construídos neste trabalho.

Para identificar, dentro de uma categoria de *hashtags*, quais foram os principais assuntos discutidos no Twitter, pelos usuários analisados neste trabalho, foi criado um modelo LDA para representar cada categoria de *hashtag*. Serão criados 3 tópicos por categoria de *hashtag*. Para cada grupo de palavras gerado pelo LDA, é objetivo deste trabalho rotular esse grupo de palavras como um tópico de debate dentro de cada categoria de *hashtag*.

[1] *Gensim* é uma biblioteca de código aberto para modelagem de tópicos não supervisionada e processamento de linguagem natural, utilizando aprendizado de máquina (REHUREK; SOJKA, 2011).

Figura 15 – Valores de coerência para modelos LDA com diferentes número de tópicos



4.2.1 CPI da Pandemia

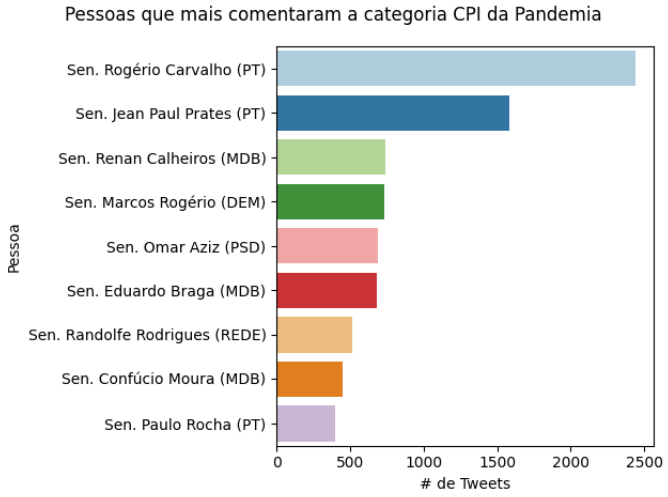
Para a categoria CPI da Pandemia, a modelagem tópicos do LDA resultou na Tabela 6. O rótulo notícias da CPI se refere à *tweets* que foram postados ou *retweetados* pelos usuários analisados, e se referem à noticias sobre a CPI da Pandemia. O rótulo investigações da CPI, refere-se à *tweets* que foram postados com a intenção de debater os acontecimentos da CPI da Pandemia. Por fim, o tópico resultado das investigações agrupa palavras que fazem menção ao governo do presidente Jair Bolsonaro no contexto do relatório final da CPI. De forma geral é importante perceber que no contexto da CPI da Pandemia, o governo do presidente Jair Bolsonaro é citado em todos os três tópicos.

Tabela 6 – LDA de 3 tópicos para a categoria CPI da Pandemia

Rótulo	**Palavras**
Notícias sobre a CPI	cpi, sobre, hoje, saúde, não, ministério, agora, bolsonaro, governo, dia
Investigações da CPI	não, bolsonaro, governo, brasil, mais, mil, pandemia, vacinas, brasileiros, crimes
Resultado das investigações	não, cpi, presidente, bolsonaro, relatório, mais, como, final, contra, governo

A Figura 16 mostra o total de publicações feitas pelas pessoas analisadas que contém ao menos uma menção às *hashtags* que compõem a categoria CPI da Pandemia. É possível analisar que todos as dez pessoas que mais postaram são Senadores da República, e no topo do ranking está o Senador Rogério Carvalho (PT). Também estão presentes na lista o relator da CPI da Pandemia, o Senador Renan Calheiros (MDB), o presidente da Comissão, Senador Omar Azis (PSD), e o vice-presidente da Comissão, Senador Randolfe Rodrigues.

Figura 16 – Pessoas que mais publicaram na categoria CPI da Pandemia



4.2.2 COVID-19

Na categoria COVID-19, o resultado do LDA mostra uma preocupação muito grande com a vacinação das pessoas, apesar do grupo não se tratar especificamente de vacina, de acordo com a Tabela 7. Para tal, no tópico explicando a eficácia das vacinas, percebe-se uma preocupação em convencer o leitor de que as vacinas imunizam as pessoas, e a importância de se tomar a segunda dose. No tópico notícias sobre covid e vacinas, percebe-se o uso recorrente de palavras imediatistas, como hoje e momento, mas abordando também conceitos de vacinação e casos de COVID-19. Por fim, o tópico campanhas de vacinação faz alusão à distribuição de doses de vacina, e utiliza recorrentemente os termos relacionados à vacinação. A forte correlação entre os assuntos abordados entre a categoria COVID-19 e Vacinação pode ser visualizada na Figura 13.

É preciso ressaltar que não é possível associar, tanto pela análise de intersecção entre categorias de *hashtags*, quanto pela análise da modelagem de tópicos, que os *tweets* da categoria Covid fazem referência à eventos da CPI da Pandemia.

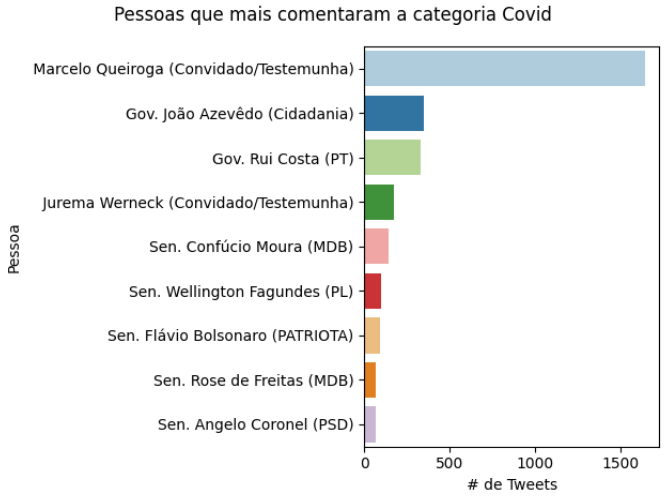
Tabela 7 – LDA de 3 tópicos para a categoria COVID-19

Rótulo	**Palavras**
Explicando a eficácia das vacinas	dose, não, contra, 2ª, vacina, pessoas, saúde, imunização, vacinas, casos
Notícias sobre covid e vacinas	não, brasil, mais, casos, hoje, vacinação, saúde, vacinas, momento, contra
Campanhas de vacinação	doses, milhões, contra, mais, vacinação, vacinas, brasil, vacina, campanha

Na Figura 17 é possível visualizar as pessoas que mais postaram *tweets* que mencionavam ao menos uma das *hashtags* que compõem a categoria Covid. O Ministro da Saúde, Marcelo Queiroga, foi a pessoa que mais publicou nesta categoria, seguido pelo

Governador João Azevêdo (Cidadania), do estado da Paraíba e pelo Governador Rui Costa (PT). É possível verificar uma menor presença dos Senadores da República, quando comparado com a Figura 16, e uma maior presença de convidados e testemunhas, como a diretora-executiva da Anistia Internacional Brasil, Jurema Werneck.

Figura 17 – Pessoas que mais publicaram na categoria Covid



4.2.3 Vacinação

A categoria vacinação, explicada através da Tabela 8, continua abordando o assunto vacinação. No tópico incentivo á vacinação é possível perceber palavras que destacam a importância de se vacinar contra a COVID-19, através de palavras como você, imunização, vidas e vacina. No tópico notícias sobre a vacinação, assim como na categoria CPI da Pandemia é possível perceber palavras imediatistas, como 18h, sobre, vamos e 2021, bem como palavras associadas à vacinação e pandemia. Por fim, o tópico pedidos de compra de vacinas utiliza palavras como: mais, vacina, doses e gente, este tópico foi muito comentado na CPI da Pandemia, e complementando a análise de intersecções entre os conjuntos, justifica-se uma relação entre esses dois assuntos.

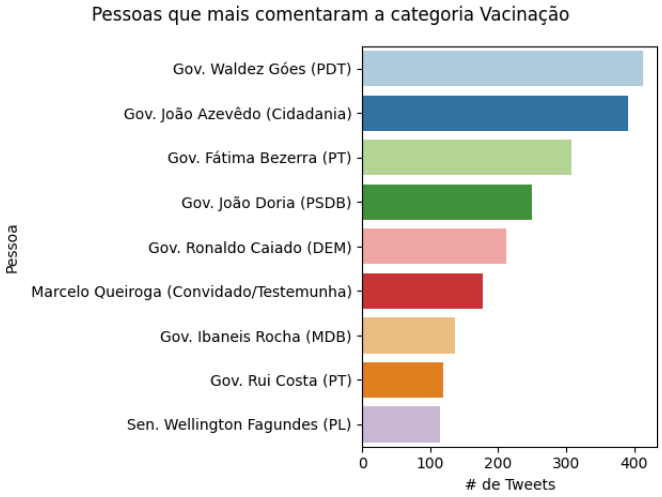
Tabela 8 – LDA de 3 tópicos para a categoria Vacinação

Rótulo	**Palavras**
Incentivo à vacinação	não, vacinas, vacina, vacinação, vidas, pandemia, saúde, brasil, você, imunização
Notícias sobre a vacinação	mais, vacinação, covid-19, brasil, 18h, não, sobre, pandemia, vamos, 2021
Pedidos de compra de vacinas	doses, mais, vacina, contra, vacinas, dose, covid-19, mil, vamos, gente

Na Figura 18 é possível verificar as dez pessoas que mais utilizaram *hashtags* que fazem parte da categoria Vacinação. É possível verificar que a maioria das pessoas que

mais postaram sobre essa categoria são Governadores da República, além do Ministro da Saúde, Marcelo Queiroga. Os Governadores se interessaram mais pelo assunto vacinação, possivelmente porque cabia a eles realizar a compra de vacinas e divulgar informações sobre o andamento da vacinação em seu estado. O Governador que mais publicou foi Waldez Góes (PDT), do estado do Amapá, seguido pelo Governador João Azevêdo (Cidadania) do estado da Paraíba e a Governadora Fátima Bezerra (PT), do estado do Rio Grande do Norte. Esta última além de participar significativamente do debate sobre a vacinação, também foi responsável pela quarta *hashtag* mais utilizada em todo o conjunto de dados coletado.

Figura 18 – Pessoas que mais publicaram na categoria Vacinação



4.2.4 Máscara

Na categoria máscara é possível perceber uma forte influência de *tweets* que referenciam o estado da Paraíba. De acordo com Melo (2022), a Paraíba foi o último estado do Brasil a flexibilizar o uso de máscara, portanto, a estratégia de comunicação está alinhada à uma política de uso de máscara pelo Governo da Paraíba. O primeiro tópico, mensagens de apoio, faz uso de palavras de incentivo, como vamos, vencer e população. O tópico notícias sobre a Paraíba utiliza, novamente, palavras imediatistas, como hoje, assista e acompanhe. Por fim, o tópico informações à população transmite informações em geral para a população, não se tratando somente do uso de máscara. Este resultado corrobora com a análise de intersecções entre categorias, mostrando que não necessariamente esses *tweets* estão relacionados com a CPI da Pandemia.

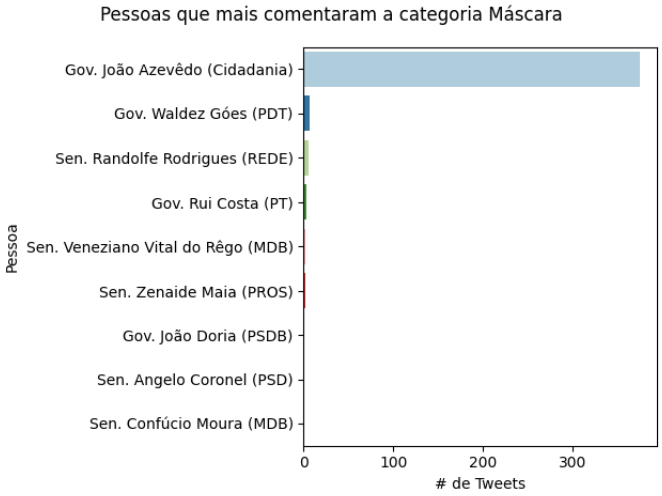
Na Figura 19 é possível verificar as dez pessoas que mais utilizaram *hashtags* que fazem parte da categoria Máscara. É importante destacar que existe uma discrepância muito grande entre o volume de publicações nesta categoria, sendo o principal usuário, o

Tabela 9 – LDA de 3 tópicos para a categoria Máscara

Rótulo	**Palavras**
Mensagens de apoio	mais, profissionais, população, grande, vencer, vamos, nosso, nessa, também, esse
Notícias sobre a Paraíba	mais, paraíba, estado, hoje, doses, paraibanos, vídeo, acompanhe, assista, programa
Informações à população	mais, paraíba, não, vacina, isso, pandemia, saúde, população, quem, muito

Governador da Paraíba, João Azevêdo (Cidadania). O fato de o Governador da Paraíba possuir um alto volume de publicações a respeito desse assunto em seu estado justifica o porquê tantas menções ao estado da Paraíba nos resultados do LDA.

Figura 19 – Pessoas que mais publicaram na categoria Máscara



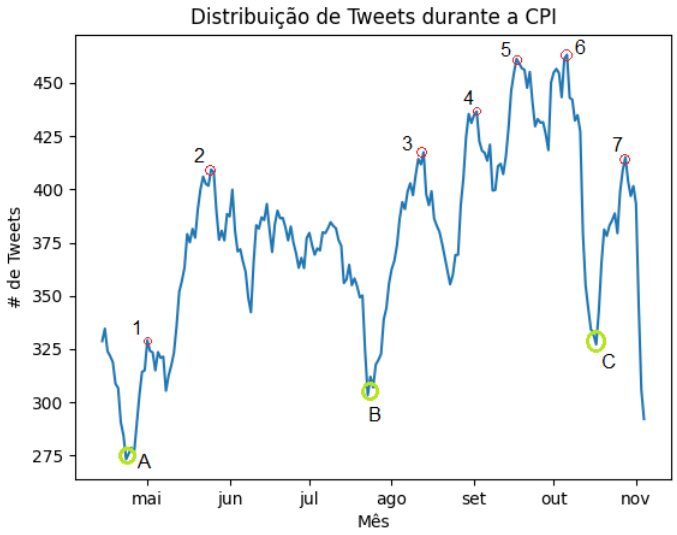
4.3 Identificação dos principais momentos da CPI

Para identificar os períodos de maior e menor movimentação no debate público no Twitter dos senadores, governadores, convidados e testemunhas da CPI, foram demarcados, na Figura 20, pontos de 1 a 7 para representar os topos, e três pontos A, B e C para representar os vales. Os topos foram escolhidos considerando períodos onde ocorre um alto volume de postagens seguido por uma queda, os vales por sua vez, são caracterizados por um declínio no volume de postagens seguido por uma alta repentina. Existem muitos topos e vales na Figura 20, porém será realizada uma análise dos topos e vales mais significativos.

A Figura 20 utiliza uma média móvel de sete dias em relação ao número total de postagens por dia. Para realizar uma análise que considere essa janela de sete dias, para cada ponto demarcado na Figura 20, serão identificados os convidados e testemunhas que foram convocados para depor na CPI na semana da data informada pela Figura 20. A

Tabela 10 relaciona, em ordem cronológica, os pontos demarcados na Figura 20 com os convidados e testemunhas da CPI que compareceram naquela semana, de acordo com Senado (2021e).

Figura 20 – Identificação dos principais momentos do período de análise



Na Tabela 10 os vales A e B foram marcados por presenças pouco marcantes nas sessões da CPI, em contraste com os picos 1, onde a convocação do Ministro da Saúde na época da crise de oxigênio em Manaus, Eduardo Pazuello, provocou grande debate tanto da mídia convencional, quanto nas redes sociais. No pico 2, o não comparecimento de Carlos Wizard fez com que fosse questionado se um precedente de não comparecimento pudesse prejudicar o andamento das investigações, novamente gerando debate nas redes sociais.

A medida com que as investigações da CPI avançam, é possível perceber uma tendência de alta do volume de postagens de *tweets*, nos picos 3, 4 e 5. Esta tendência ocorre pois no ponto 3 começa a se revelar um possível esquema de corrupção envolvendo a compra de vacinas, principalmente com a convocação do líder do governo na Câmara dos Deputados, Ricardo Barros (PP) (SENADO, 2021d).

O comparecimento de Ivanildo Gonçalves da Silva, no ponto 4, corroborando com as investigações e confirmando que houveram pagamentos em espécie em nome da empresa VTCLog, fez com que este assunto fosse ainda mais comentado no Twitter (SENADO, 2021c). No ponto 5 o principal nome convocado foi Pedro Benedito Baptista Júnior, diretor da Prevent Sênior, empresa acusada de realizar testes envolvendo o uso de medicamentos sem eficácia comprovada contra a COVID-19 (SENADO, 2021b).

O auge do volume de postagens ocorre no pico número 6, ele foi marcado pela

Tabela 10 – Principais eventos da CPI

Ponto	**Período**	**Eventos da semana na CPI**
A	10/05/2021 a 14/05/2021	Antônio Barra Torres (Diretor-presidente da Agência Nacional de Vigilância Sanitária (ANVISA)), Fabio Wajngarten (Secretário de comunicação da presidência), Carlos Murillo (Presidente da Pfizer na LATAM)
1	17/05/2021 a 21/05/2021	Eduardo Pazuello (Ex-ministro da Saúde) e Ernesto Araújo (Ex-ministro das relações exteriores)
2	13/06/2021 a 18/06/2021	Não comparecimento de Carlos Wizard (Empresário) à convocação da CPI, Alexandre Marques (Auditor do TCU)
B	11/07/2021 a 17/07/2021	Francisco Maximiliano (Sócio da Precisa Medicamentos), Emanuela Medrades (Diretora técnica da Precisa Medicamentos) e Cristiano Carvalho (Representante da Davati Medical Supply)
3	08/08/2021 a 14/08/2021	Ricardo Barros (Deputado Federal, líder do governo na câmara), Jailton Batista (Diretor da Vitamedic, convidado) e Helcio Bruno de Almeida (Tenente-coronel da reserva)
4	29/08/2021 a 04/09/2021	Ivanildo Gonçalves da Silva (Motoboy que confirmou saques e pagamentos de boletos em nome da VTCLog) e Francisco Araújo Filho (Ex-secretário de Saúde do DF)
5	19/09/2021 a 25/09/2021	Pedro Benedito Baptista Júnior (Diretor da Prevent Sênior), Danilo Berndt Trento (Diretor da Precisa Medicamentos) e Wagner de Campos Rosário (Ministro-chefe da controladoria geral da União)
6	26/09/2021 a 02/10/2021	Luciano Hang (Dono das lojas Havan), Otávio Oscar Fakhoury (Empresário investigado no inquérito das Fake News) e Bruna Morato (Advogada de médicos que trabalharam na Prevent Sênior)
C	10/10/2021 a 16/10/2021	Não houveram sessões na CPI
7	17/10/2021 a 23/10/2021	Apresentação do relatório final da CPI

Fonte: Adaptado de Senado (2021e)

presença do empresário Luciano Hang, acusado de participar do "gabinete paralelo", um grupo de apoiadores do presidente Jair Bolsonaro suspeito de aconselhar o presidente em relação à pandemia de COVID-19, promovendo ideias sem comprovação científica, como o "tratamento precoce" através de medicamentos como hidroxicloroquina e ivermectina (SENADO, 2021a).

O último vale, letra C, foi marcado pela falta de sessões da CPI da Pandemia. Uma semana depois, quando o relatório final da CPI é apresentado, a discussão volta a se fazer presente no debate público através do Twitter, marcado pelo pico número 7.

5 Considerações finais

Em 24 de abril de 2021 foi instalada a CPI da Pandemia no Senado Federal do Brasil, desde então as redes sociais exerceram um papel importante no debate público sobre o comportamento de cada convidado, bem como os fatos que foram investigados e revelados pela comissão. Alguns convidados da CPI fizeram com o que o número de *tweets* postados pelos senadores, governadores e outros convidados disparasse em alguns momentos, foi o caso do Ex-Ministro da Saúde, General Eduardo Pazuello, o empresário Luciano Hang, e o motoboy Ivanildo Gonçalves da Silva.

Este trabalho teve como objetivo principal identificar os principais assuntos discutidos durante a CPI da Pandemia, através da análise das *hashtags* utilizadas pelos usuários analisados no Twitter. Para realizar esta análise, foi estabelecido como objetivo específico coletar os *tweets* dos senadores, governadores, convidados e testemunhas da CPI da Pandemia, identificar as principais *hashtags* utilizadas por esses usuário e relacionar os principais períodos de maior e menos intensidade de postagens com os convidados e testemunhas presentes na CPI nessas datas.

Em posse dos *tweets* coletados foi possível identificar as principais *hashtags* utilizadas por essas pessoas, através do ranking mostrado na Tabela 2. Nessas *hashtags* é possível verificar a presença de termos recorrentes, como CPI da Pandemia, Covid, Vacinação e Máscara. Também foi possível identificar que cinco *hashtags* citavam nominalmente os Senadores que as publicaram, corroborando com o volume de postagens dos Senadores no período analisado, representando 55% dos *tweets* coletados.

Uma vez que foram identificadas as principais *hashtags*, foi possível agrupá-las em quatro assuntos principais: CPI da Pandemia, Covid, Máscara e Vacinação. Através da visualização de intersecção de conjuntos entre as principais categorias de *hashtag*, foi possível verificar uma grande correlação entre as categorias Máscara, Covid e Vacinação, e em relação à CPI da Pandemia, a categoria Covid se mostrou a mais relacionada.

A categoria CPI da Pandemia foi marcada por postagens sobre notícias da CPI e resultado das investigações. Nesta categoria, as dez pessoas que mais participaram do debate são Senadores da República. Na categoria Covid foi identificado um tópico em que foi explicada a eficácia das vacinas e a pessoa que mais publicou nessa categoria foi o ministro da Saúde, Marcelo Queiroga. A categoria Vacinação teve uma participação majoritária dos Governadores da República e do Ministro da Saúde, os assuntos identificados passam por pedidos de compra de vacinas, notícias sobre a vacinação e incentivo à vacinação. A categoria Máscara teve grande participação do Governador da Paraíba, João Azevêdo, que nos tópicos identificados buscou passar informações, notícias e mensagens de apoio à

população através do Twitter.

Por fim, foi possível relacionar os eventos que ocorreram na CPI com os principais pontos de aumento e redução do volume de postagens. Os períodos de baixo volume de postagens passam por convocações na CPI como Antônio Barra Torres, diretor-presidente da Agência Nacional de Vigilância Sanitária (ANVISA), Francisco Maximiliano, sócio da Precisa Medicamentos, entre outros. O primeiro período de maior volume de postagens ocorreu na presença do Ex-ministro da saúde, o general Eduardo Pazuello e do Ex-ministro das relações exteriores, Ernesto Araújo. O não comparecimento de Carlos Wizard, bem como a presença de Alexandre Marques, Auditor do TCU, estão relacionado com o segundo período de maior volume de postagens analisado.

Houve uma tendência de alta no volume de postagens que começou com a presença do líder do governo na câmara dos deputados, Ricardo Barros, passando pelo motoboy Ivanildo Gonçalves da Silva, que confirmou o depósito em espécie à empresa VTCLog, acusada de corrupção pela CPI da Pandemia, até o depoimento de Pedro Benedito Baptista Júnior, diretor da Prevent Sênior, empresa acusada de realizar testes com medicamentos sem eficiência comprovada em pacientes com COVID-19. O convidado que gerou o maior volume de postagens no Twitter foi o Luciano Hang, dono das Lojas Havan.

Para trabalhos futuros, é possível utilizar a mesma base de dados construída para extrair informações sobre o sentimento dos *tweets* coletados, ou utilizar a metodologia de coleta para analisar o comportamento dos atores políticos no Twitter em outros eventos. É importante mencionar que a base de *tweets* coletados neste trabalho será disponibilizada de forma gratuita para a comunidade acadêmica.

Este trabalho apresenta uma caracterização do conteúdo publicado por atores políticos e convidados da CPI da Pandemia, até o momento de publicação não haviam outros trabalhos que analisaram as publicações no Twitter neste contexto. Desta forma, este trabalho é relevante para cientistas políticos e jornalistas que buscam entender melhor os eventos que ocorreram na CPI da Pandemia, no ano de 2021.

Referências

AMARAL, M. S.; PINHO, J. A. G. d. Eleições parlamentares no brasil: O uso do twitter na busca por votos. *Revista de Administração Contemporânea*, SciELO Brasil, v. 22, p. 466–486, 2018. Citado 4 vezes nas páginas 14, 19, 26 e 27.

BARTHOLOMEW, D. J.; KNOTT, M.; MOUSTAKI, I. *Latent variable models and factor analysis: A unified approach*. [S.l.]: John Wiley & Sons, 2011. v. 904. Citado na página 25.

BERNARDES, C. Uso do twitter para engajamento político: análise dos perfis das assembleias legislativas da região sudeste. *Compolítica*, v. 10, p. 5–48, 01 2021. Citado na página 18.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado na página 24.

BRASIL, B. N. *Barroso determina instalação de CPI da Pandemia no Senado*. 2021. Disponível em: <https://www.bbc.com/portuguese/brasil-56684539>. Citado na página 30.

CAESAR, G. *Veja quem são os 27 governadores eleitos nas eleições deste ano*. 2018. Disponível em: <https://g1.globo.com/politica/noticia/2021/04/08/barroso-determina-que-senado-instale-cpi-da-pandemia.ghtml>. Citado na página 30.

CARVALHO, W.; GUIMARÃES, S. Desinformação, negacionismo e automedicação: a relação da população com as drogas "milagrosas" em meio à pandemia da covid-19. *InterAmerican Journal of Medicine and Health*, v. 3, Aug. 2020. Disponível em: <https://www.iajmh.com/iajmh/article/view/147>. Citado na página 14.

CREMONESE, D. Política on-line: a utilização do twitter como ferramenta de capital social nas eleições presidenciais de 2010. *Sociedade e Cultura*, v. 15, n. 1, p. 10–5216, 2012. Citado 3 vezes nas páginas 14, 18 e 19.

CUNHA, E. L. T. P. Etiquetagem de micromensagens no twitter: uma abordagem linguística. Universidade Federal de Minas Gerais, 2012. Citado na página 34.

ESTADO, A. *Anitta bloqueia Bolsonaro no Twitter após interação irônica do presidente*. 2022. Disponível em: <https://www.correiobraziliense.com.br/politica/2022/04/5001149-anitta-bloqueia-bolsonaro-no-twitter-apos-interacao-ironica-do-presidente.html>. Citado 2 vezes nas páginas 18 e 19.

ESTADO DE SÃO PAULO, O. *Criticada por Musk, moderação do Twitter já suspendeu publicações de políticos brasileiros; relembre*. 2022. Disponível em: <https://politica.estadao.com.br/noticias/geral,musk-moderacao-twitter-suspendeu-publicacoes-politicos-brasileiros-nprp,70004048148>. Citado na página 19.

FALCÃO, G. G. M. *Barroso determina que Senado instale CPI da Pandemia*. 2021. Disponível em: <https://g1.globo.com/politica/noticia/2021/04/08/barroso-determina-que-senado-instale-cpi-da-pandemia.ghtml>. Citado na página 30.

FELDMAN, J. S. R. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006. ISBN 0521836573,9780521836579. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=e6f83ec309c84a968a2b85e6adb31c25>. Citado 4 vezes nas páginas 20, 21, 23 e 24.

G1. *Twitter apaga publicações de Jair Bolsonaro por violarem regras da rede*. 2020. Disponível em: <https://g1.globo.com/politica/noticia/2020/03/29/twitter-apaga-publicacoes-de-jair-bolsonaro-por-violarem-regras-da-rede.ghtml>. Citado na página 20.

G1. *A 6 dias do fim, abril se torna o mês mais letal da pandemia no Brasil*. 2021. Disponível em: <https://g1.globo.com/bemestar/coronavirus/noticia/2021/04/24/abril-se-torna-o-mes-mais-letal-da-pandemia-no-brasil.ghtml>. Citado na página 17.

G1. *Quem é Otávio Fakhoury, empresário que senadores dizem ter financiado a divulgação de fake news sobre a pandemia*. 2021. Disponível em: <https://g1.globo.com/politica/cpi-da-covid/noticia/2021/09/30/quem-e-otavio-fakhoury.ghtml>. Citado na página 38.

G1. *Terça-feira, 13 de abril*. 2021. Disponível em: <https://g1.globo.com/resumo-do-dia/noticia/2021/04/13/terca-feira-13-de-abril.ghtml>. Citado na página 38.

GAZEL, A. S.; CRUZ, V. *Crise do Oxigênio no Amazonas completa um ano com impunidade e Incerteza CAUSADA Pela ômicron*. G1, 2022. Disponível em: <https://g1.globo.com/am/amazonas/noticia/2022/01/14/crise-do-oxigenio-no-amazonas-completa-um-ano-com-impunidade-e-incerteza-causada-pela-omicron.ghtml>. Citado na página 14.

IBGE. *Pesquisa mostra que 82,7% dos domicílios brasileiros têm acesso à internet*. 2021. Disponível em: <https://www.gov.br/mcom/pt-br/noticias/2021/abril/pesquisa-mostra-que-82-7-dos-domicilios-brasileiros-tem-acesso-a-internet>. Citado na página 19.

INSTITUTE, L. *Covid performance index*. 2021. Disponível em: <https://interactives.lowyinstitute.org/features/covid-performance/>. Citado na página 17.

JELODAR, H. et al. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, Springer, v. 78, n. 11, p. 15169–15211, 2019. Citado 2 vezes nas páginas 24 e 25.

JO, T. *Text mining: concepts, implementation, and big data challenge*. Springer, 2019. (Studies in big data 45). ISBN 978-3-319-91814-3,3319918141,978-3-319-91815-0. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=669c2c62a9a83a24c0736558e33a6439>. Citado 5 vezes nas páginas 20, 21, 22, 23 e 24.

JONES, M.; HARDT, D. *The oauth 2.0 authorization framework: Bearer token usage*. [S.l.], 2012. Citado na página 28.

JURiDICO, C. *Por decisão do STF, Twitter e Facebook apagam contas de aliados de Bolsonaro*. 2020. Disponível em: <https://www.conjur.com.br/2020-jul-24/decisao-alexandre-twitter-apaga-contas-aliados-bolsonaro>. Citado na página 20.

LAVOR, A. d. et al. Amazônia sem respirar: falta de oxigênio causa mortes e revela colapso em manaus. ENSP/Fiocruz, 2021. Citado na página 14.

LEX, A. et al. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis)*, v. 20, n. 12, p. 1983–1992, 2014. Citado na página 29.

LIBARDI, G. *Base de dados coletada no Twitter, durante a CPI da pandemia*. 2022. Disponível em: <https://github.com/GuilhermeLibardi/tweets-cpi-pandemia/blob/9625888a91a68b030ea8168f7ac55ffea1b71dd4/tweets-sentiment-analysis.xlsx>. Citado na página 37.

LIBARDI, G. *Requisições dos tweets da CPI da pandemia*. 2022. Disponível em: <https://github.com/GuilhermeLibardi/tweets-cpi-pandemia/blob/f92b82c6063b9809888381e84905564609672f2f/full-requests.rar>. Citado na página 33.

LIN, J. On the dirichlet distribution. *Department of Mathematics and Statistics, Queens University*, 2016. Citado na página 25.

LUBANOVIC, B. *Introducing Python: Modern Computing in Simple Packages*. [S.l.]: "O'Reilly Media, Inc.", 2014. Citado na página 29.

MACHADO, R. *Saiba quais são os principais alvos da CPI da Covid e o que pesa contra eles*. 2021. Disponível em: <https://www1.folha.uol.com.br/poder/2021/04/entenda-quem-sao-os-principais-alvos-da-cpi-da-covid-e-o-que-pesa-contra-eles.shtml>. Citado na página 17.

MALAGOLI, L. G. et al. A look into covid-19 vaccination debate on twitter. In: *13th ACM Web Science Conference 2021*. New York, NY, USA: Association for Computing Machinery, 2021. (WebSci '21), p. 225–233. ISBN 9781450383301. Disponível em: <https://doi.org/10.1145/3447535.3462498>. Citado 2 vezes nas páginas 26 e 27.

MARTINI, P. *Mudanças no Twitter preocupam pesquisadores brasileiros sobre regulação das redes sociais*. 2022. Disponível em: <https://www.cnnbrasil.com.br/tecnologia/mudancas-no-twitter-preocupam-pesquisadores-brasileiros-sobre-regulacao-das-redes-sociais>. Citado na página 19.

MATOS, E. O.; DOURADO, T. M.; MESQUITA, P. @ dilmabr no impeachment: Uma análise das estratégias de comunicação política de dilma rousseff no twitter. *Comunicação & Sociedade*, v. 39, n. 3, p. 61–77, 2017. Citado 2 vezes nas páginas 25 e 27.

MCKINNEY, W. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. [S.l.]: "O'Reilly Media, Inc.", 2012. Citado na página 29.

MELO, K. *Paraíba é último estado a flexibilizar uso de máscara em espaço aberto*. 2022. Disponível em: <https://agenciabrasil.ebc.com.br/saude/noticia/2022-04/paraiba-e-ultimo-estado-flexibilizar-uso-de-mascara-em-espaco-aberto>. Citado na página 47.

MINER, G. A. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press, 2012. ISBN 978-0-12-386979-1. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=e39f01d054442fc8577e37397ed3be9b>. Citado 3 vezes nas páginas 20, 21 e 24.

NACIONAL, J. *CPI Da Covid: Ministério da saúde queria convencer médicos E doentes a USAREM remédios sem eficácia comprovada contra covid*. G1, 2021. Disponível em: <https://g1.globo.com/jornal-nacional/noticia/2021/07/21/cpi-da-covid-ministerio-da-saude-queria-convencer-medicos-e-doentes-a-usarem-remedios-sem-eficacia ghtml>. Citado na página 14.

OLIVEIRA, L. S. et al. When politicians talk about politics: Identifying political tweets of brazilian congressmen. In: *Twelfth International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2018. Citado 2 vezes nas páginas 26 e 27.

ORELLANA, J. D. Y. et al. Excesso de mortes durante a pandemia de covid-19: subnotificação e desigualdades regionais no brasil. *Cadernos de Saúde Pública*, SciELO Public Health, v. 37, p. e00259120, 2021. Citado na página 17.

OSMAN, M. *Estatísticas e Fatos do Twitter Sobre a Nossa Rede Favorita*. 2021. Disponível em: <https://kinsta.com/pt/blog/estatisticas-e-fatos-do-twitter>. Citado 2 vezes nas páginas 18 e 19.

REHUREK, R.; SOJKA, P. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, v. 3, n. 2, 2011. Citado na página 43.

REUTERS. *Médicos sem Fronteiras chamam pandemia no Brasil de catástrofe humanitária*. 2021. Disponível em: <https://exame.com/brasil/medicos-sem-fronteiras-chamam-pandemia-no-brasil-de-catastrofe-humanitaria/>. Citado na página 17.

SAÚDE, O. P.-A. da. *Entenda a infodemia e a desinformação na luta contra a COVID-19*. [S.l.]: Opas, 2020. Citado na página 14.

SAÚDE, O. P.-A. da. *Histórico da Pandemia de Covid-19*. 2020. Website. Disponível em: <https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19>. Acesso em: 01 mai 2022. Citado na página 14.

SENADO, A. *CPI da Pandemia ouve nesta quarta depoimento do empresário Luciano Hang*. 2021. Disponível em: <https://www12.senado.leg.br/noticias/materias/2021/09/24/cpi-da-pandemia-marca-para-quarta-depoimento-do-empresario-luciano-hang>. Citado na página 50.

SENADO, A. *CPI: diretor da Prevent Senior é acusado de mentir e passa à condição de investigado*. 2021. Disponível em: <https://www12.senado.leg.br/noticias/materias/2021/09/22/cpi-diretor-da-prevent-senior-e-acusado-de-mentir-e-passa-a-condicao-de-investigado>. Citado na página 49.

SENADO, A. *CPI elogia Ivanildo por depor e critica Tolentino e Marconny por atestados médicos*. 2021. Disponível em: <https://www12.senado.leg.br/noticias/materias/2021/09/01/tolentino-alega-mal-estar-para-nao-ir-a-cpi-motoboy-ivanildo-comparece-e-confirma-saques>. Citado na página 49.

SENADO, A. *CPI ouve deputado Ricardo Barros sobre irregularidades na compra da Covaxin*. 2021. Disponível em: <https://www12.senado.leg.br/noticias/materias/2021/08/06/cpi-ouve-deputado-ricardo-barros-sobre-irregularidades-na-compra-da-covaxin>. Citado na página 49.

SENADO, F. *CPI da Pandemia*. 2021. Disponível em: <https://legis.senado.leg.br/comissoes/comissao?codcol=2441>. Citado 9 vezes nas páginas 14, 17, 30, 32, 38, 41, 42, 49 e 50.

SENADO, F. *Senadores em Exercício*. 2022. Disponível em: <https://www25.senado.leg.br/web/senadores/em-exercicio>. Citado na página 30.

SETIAWAN, R. E. B. et al. Analysis of the governor's communication model on twitter. In: ATLANTIS PRESS. *International Conference on Public Organization (ICONPO 2021)*. [S.l.], 2022. p. 215–219. Citado na página 18.

SILVA, T. V. O uso do twitter pelos deputados federais brasileiros: estudo sobre atuação e tendências de comportamento. *Biblioteca Digital da Câmara dos Deputados do Brasil*, 2012. Citado na página 18.

SMALL, T. A. What the hashtag? *Information, Communication & Society*, Routledge, v. 14, n. 6, p. 872–895, 2011. Disponível em: <https://doi.org/10.1080/1369118X.2011.554572>. Citado na página 15.

SRIVASTAVA, M. S. A. *Text Mining: Classification, Clustering, and Applications*. 1. ed. Chapman Hall, 2009. (Chapman Hall/CRC Data Mining and Knowledge Discovery Series). ISBN 1420059408,9781420059403. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=acee2645cd1a772c992e9706c6b1f195>. Citado 2 vezes nas páginas 24 e 25.

TONKIN, E.; PFEIFFER, H. D.; TOURTE, G. Twitter, information sharing and the london riots? *Bulletin of the American Society for Information Science and Technology*, Wiley Online Library, v. 38, n. 2, p. 49–57, 2012. Citado na página 34.

TWITTER. *Timelines*. [S.l.], 2021. Disponível em: <https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-tweets>. Acesso em: 04/06/2022. Citado 2 vezes nas páginas 32 e 33.

VEJA. *Bolsonaro foi 2º governante a ter post apagado pelo Twitter; 1º foi Maduro*. 2020. Disponível em: <https://veja.abril.com.br/mundo/antes-de-bolsonaro-twitter-apagou-post-de-maduro-com-antidoto-caseiro/>. Citado na página 20.

VERGEER, M.; HERMANS, L.; SAMS, S. Online social networks and micro-blogging in political campaigning: The exploration of a new campaign tool and a new campaign style. *Party politics*, Sage Publications Sage UK: London, England, v. 19, n. 3, p. 477–501, 2013. Citado 2 vezes nas páginas 14 e 19.

VIANA, J. Q. et al. A recuperação da informação em redes sociais: o uso e aplicação das hashtags. Universidade Federal de Minas Gerais, 2019. Citado na página 34.

VIJAYARANI, S. et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, v. 5, n. 1, p. 7–16, 2015. Citado 2 vezes nas páginas 22 e 23.

VIJAYARANI, S.; JANANI, R. et al. Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, v. 3, n. 1, p. 37–47, 2016. Citado na página 22.

WANG, H. et al. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: *Proceedings of the ACL 2012 system demonstrations*. [S.l.: s.n.], 2012. p. 115–120. Citado na página 15.

ZACHARIAS, B. *CPI da Pandemia gera 11,8 milhões de publicações no Twitter*. 2021. Disponível em: <https://www.cnnbrasil.com.br/politica/cpi-da-pandemia-gera-11-8-milhoes-de-publicacoes-no-twitter>. Citado na página 14.