UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA

DANIELA CARVALHO FERRAZ NOLASCO NEVES

EVOLUÇÃO DAS *FAKE NEWS* NO BRASIL NO DECORRER DA PANDEMIA

São Luís – MA
2022

EVOLUÇÃO DAS *FAKE NEWS* NO BRASIL NO DECORRER DA PANDEMIA

Trabalho de Conclusão de Curso II apresentado como requisito para obtenção do título de Bacharel em Engenharia da Computação pela Universidade Federal do Maranhão

Grande Área: Ciências Exatas e da Terra

Área: Ciência da Computação

Subárea: Metodologia e Técnicas da Computação

Orientador: Prof. Dr. Bruno Feres de Souza

DANIELA CARVALHO FERRAZ NOLASCO NEVES

EVOLUÇÃO DAS *FAKE NEWS* NO BRASIL NO DECORRER DA PANDEMIA

Trabalho de Conclusão de Curso II apresentado como requisito para obtenção do título de Bacharel em Engenharia da Computação pela Universidade Federal do Maranhão

BANCA EXAMINADORA:

Prof. Dr. Bruno Feres de Souza
Orientador

Prof. Ms. Cláudio Manoel Pereira Aroucha
Universidade Federal do Maranhão

Profa. Ms. Rayanne Maria Cunha Silveira
Universidade Federal do Maranhão

SÃO LUÍS – MA
2022

AGRADECIMENTOS

Gostaria de agradecer a todos que contribuíram para minha jornada de 5 anos na Universidade Federal do Maranhão. Primeiramente, à minha mãe, por ter me dado todo apoio necessário para que eu pudesse mudar de cidade e garantir meus estudos por tantos anos de minha vida. Por toda a minha família que me apoiou na mudança, nas conquistas, e contribuíram financeiramente para que eu conseguisse ter a estrutura e recursos necessários aqui em São Luís.

Quero agradecer imensamente ao meu orientador, Prof. Dr. Bruno Feres de Souza, que sempre me apoiou desde 2017 e me ajudou e orientou por tantas vezes durante todos estes anos. Ambos os professores Dr. Bruno Feres de Souza e Dra. Maira Silva Ferreira foram os grandes responsáveis para que eu ingressasse na Engenharia da Computação. Sou eternamente grata pela paciência, apoio e influência de ambos. Dedico outro agradecimento especial para o professor Dr. Bruno Feres de Souza por também ter sido uma inspiração na área da computação, e ter sido um verdadeiro educador na minha vida. Todo o apoio, todas as conversas e trocas de opiniões, foram inestimáveis para minha construção pessoal, acadêmica e profissional.

Quero agradecer a todo o corpo docente da Engenharia da Computação, que sempre se esforçaram para darem as melhores aulas, criarem projetos com os alunos, realizar apresentações na Feira das Profissões, e fizeram inestimáveis esforços para trazer conhecimento e sabedoria na sala de aula. Todos vocês foram essenciais para minha formação e lembro de todos vocês com frequência no meu dia a dia.

Quero deixar um agradecimento especial também para os professores Ms. Cláudio Manoel Pereira Aroucha, Dr. Davi Viana dos Santos e Dr. Rafael Fernandes Lopes por me ajudarem a conseguir um laboratório de estudos no Centro de Empreendedorismo da UFMA, sem o qual não teria recursos necessários para desempenhar meus estudos com bons equipamentos. Também agradeço por todo o apoio durante essa jornada. Tive o prazer de estudar com os três e todos foram muito solícitos e compreensíveis comigo sempre que precisei.

Gostaria de agradecer também ao corpo docente do Bacharelado Interdisciplinar em Ciência e Tecnologia, o qual teve formação inicial no meu processo de busca por um diploma. Tive o prazer de ter professores incríveis na minha formação inicial. Sou grata, em especial, ao prof. Dr. José Renato de Oliveira Lima, que foi um coordenador incrível, além de ter me ajudado a conseguir minha primeira bolsa de estudos no Núcleo de Tecnologia da Informação.

Quero também agradecer ao Prof. Dr. Areolino de Almeida Neto que me acompanhou por várias vezes nessa jornada universitária, especialmente na disciplina de Cálculo Numérico e no meu projeto de pesquisa.

Sou grata a todos os amigos que fiz nessa jornada incrível. Por tantas noites estudando juntos, buscando ajuda em conjunto, pelos grupos de estudo e por todas as incríveis experiências que compartilhamos juntos. Fiz amizades com os mais diversos grupos de pessoas, nos mais diversos cursos, e sou eternamente grata a todos. Levarei-os sempre comigo!

Agradeço a todos que contribuíram, direta ou indiretamente, com minha formação acadêmica. Vivi anos incríveis graças a todos vocês e, por isso, serei eternamente grata.

Por fim, queria agradecer à Daniela de 2017, que iniciou este curso em uma cidade nova, enfrentando os mais diversos desafios, e que teve incrível disciplina para alcançar os seus objetivos. Hoje, consegui chegar muito além do que eu imaginava, e devo tudo aos sacrifícios e esforços que fiz no passado.

"É possível não cometer erros e ainda assim perder. Isso não é fraqueza, isso é a vida."
(Capitão Jean-Luc Picard, Star Trek: The Next Generation)

RESUMO

No começo do pandemia do coronavírus, foi observado outro fenômeno: a disseminação de informações falsas (popularmente conhecidas como *fake news*, termo em inglês) em larga escala nos mais diversos períodos, reconhecida como uma ”infodemia”pelo diretor da Organização Mundial da Saúde. Para o estudo das informações falsas disseminadas, foi utilizada uma base de dados relacionada ao tópico COVID-19 com informações de 2020 a 2022 coletadas no Boatos.org. A partir dos dados coletados, foram realizados um tratamento e análise de dados obtidos. Além do mais, foi necessária a aplicação de técnicas de mineração de texto, a fim de realizar pré-processamento dos títulos de reportagens falsas obtidos - remoção de palavras vazias, normalização morfológica, filtragem e ponderação com técnica TF-IDF, e análise de componente principal (PCA). Por fim, utilizou-se o *K-Means* para agrupamento de informações coletadas, definindo o número de *clusters* necessários via método *Elbow*. A partir dos grupos definidos, para visualização dos termos mais frequentes, foram exibidas nuvens de palavras de cada *cluster*. Diante dos termos mais frequentes de cada grupo em cada semestre dos anos 2020 a 2022, foram classificadas as *fake news* em um ou mais 9 possíveis grupos de categorias de *fake news* relacionadas à COVID-19. A partir da categorização das *fake news* obtidas, realizou-se um levantamento dos grupos de categorias mais frequentes em cada semestre e em cada ano, a fim de verificar o andamento das falsas narrativas à medida que novas informações eram dadas em cada semestre e em cada ano da pandemia no Brasil.

Palavras-chave: COVID-19; *Fake News*; Mineração de Texto; Agrupamento; Categorias.

ABSTRACT

At the beginning of the coronavirus pandemic, another phenomenon was observed: the dissemination of false information (popularly known as fake news) on a large scale in the most diverse periods, recognized as an ”infodemic”by the director of the Organization World Health. To study the spread of false information, a database related to the topic of COVID-19 was used, with information from 2020 to 2022 collected at Boatos.org. From the collected data, a treatment and analysis of the obtained data were carried out. Furthermore, it was necessary to apply text mining techniques in order to perform pre-processing of the false news headlines obtained - removal of empty words, morphological normalization, filtering and weighting with TF-IDF technique, and principal componente analysis (PCA). Finally, K-Means was used to group the collected information, defining the number of clusters needed via the Elbow method. From the defined groups, to view the most frequent terms, word clouds of each cluster were displayed. Given the most frequent terms in each group in each semester from 2020 to 2022, fake news were classified into one or more 9 possible groups of fake news categories related to COVID-19. From the categorization of the fake news obtained, a survey was carried out of the most frequent groups of categories in each semester and in each year, in order to verify the progress of the false narratives as new information was given in each semester and in each year of the pandemic in Brazil.

Keywords: COVID-19; Fake News; Text Mining; Clustering; Categories.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

CSV	*Comma Separated Values* – valores separados por vírgula, em português
DTM	*Document-Term Matrix* – matriz de termos do documento, em português
OMS	Organização Mundial da Saúde
PCA	*Principal Component Analysis* – análise de componente principal, em português
SSE	*Sum of Squared Error* – soma do erro quadrado, em português
TF-IDF	*Term Frequency–Inverse Document Frequency* – frequência do termo–inverso da frequência nos documentos, em português

SUMÁRIO

1 INTRODUÇÃO

Em um cenário global, em que notícias falsas aumentam paralelamente às diferenças sociopolíticas, é imprescindível a necessidade de análise à natureza de tal tipo de informação disseminada, os atores por trás de sua produção, quais as partes interessadas, os possíveis objetivos que a divulgação das mesmas e os mecanismos utilizados para chamar a atenção do seu público-alvo.

A ocorrência de disseminação de notícias falsas não é novidade, no entanto, com o advento da pandemia do COVID-19, foi trazido à tona o lado negativo das redes sociais com os mais diversos tipos de teorias e ações referentes ao coronavírus. O mundo vivenciou confusão, medo e desconfiança em relação ao vírus, à vacina, às tomadas de decisões políticas e afins.

No cenário brasileiro, foram exibidos em jornais, mídias sociais e meios televisivos, os mais diversos tipos de declarações: negacionistas, falsas, alarmistas, xenofóbicas e até mesmo supostas curas para o COVID-19. O cenário político-social nos anos de pandemia estava intrinsicamente correlacionado com a divulgação de tais tipos de informações. Enquanto diversos países seguiam protocolos e diretrizes da Organização Mundial da Saúde, no Brasil, as práticas governamentais adotadas pelo governo executiva foram controversas, principalmente devido às falas e posicionamentos do presidente da república, Jair Bolsonaro, fazendo do Brasil um alvo de críticas da comunidade científica, meios de comunicação mundais, gestores de saúde e afins (HUR; CAMESELLE; ALZATE, 2021).

A Organização Mundial da Saúde, atráves do seu diretor, informou que a pandemia do coronavírus não é apenas uma ameaça global que o mundo estava enfrentando, mas também uma ”infodemia”que precisa ser controlada (NIELSEN et al., 2020), uma vez que as informações fabricadas estavam se espalhando mais rápidos que o vírus (SOCIETY, 2020). Em todo o mundo, o assunto COVID-19 tem sido um ícone para desinformação, uma vez que muitos artigos falsos ou tendeciosos foram publicados a respeito do mesmo (PALADE; BALABAN, 2020).

Neste trabalho, será realizado um estudo comparativo de informações falsas relacionadas ao COVID-19 nos anos de 2020 a 2022 no Brasil, assim como o agrupamento dos tipos de notícias falsas espalhadas. O estudo utiliza como base o Boatos.org, um site independente que faz levantamento de notícias falsas sobre os mais diversos tópicos.

1.1 Justificativa

A desinformação existe bem antes da pandemia da COVID-19. As inverdades estão presentes nos mais diversos assuntos, desde informações com cunho político até econômico. As inverdades que foram expostas durante a pandemia utilizaram as mesmas

fontes de disseminação tradicionalmente usadas para espalhar notícias em larga escala.

A desinformação referente ao tópico da COVID-19 trouxe, dentro do cenário de enfrentamento de uma pandemia, a perda de confiança nas instituições de pesquisa e ciência, disseminando medo e aumentando as chances de propagação da doença (GALHARDI et al., 2020). Por isso, trouxe prejuízos tangíveis para a sociedade.

Neste contexto, este trabalho busca, através de uma base de dados de notícias falsas, categorizá-las em uma linha de tempo com início no primeiro ano de pandemia no Brasil a fim de verificar a mudança de narrativa a partir de novas informações referentes ao coronavírus e tratamento da doença COVID-19 que foram surgindo no decorrer de 2020 a 2022.

1.2 Objetivos

O objetivo principal deste trabalho é, através do algoritmo de agrupamento *K-Means*, categorizar títulos de reportagens falsas referentes à COVID-19 no Brasil.

Como objetivos específicos, tem-se:

- Análise e tratamento de base de dados de notícias falsas referentes à pandemia do coronavírus;
- Extrair padrões e tendências em títulos de reportagens falsas com técnicas de mineração de texto: remoção de palavras vazias, normalização morfológica, filtragem e ponderação e análise de componente principal;
- Definir quantidade de possíveis grupos de informações;
- Dada a definição de quantidade de grupos, agrupar informações coletadas;
- Exibir termos frequentes de cada grupo de informações e associá-los com uma das possíveis categorias de notícias falsas relacionadas à COVID-19.

1.3 Estrutura do trabalho

Este trabalho está organizado da seguinte forma: no capítulo 2, traz-se trabalhos relacionados ao tema aqui estudado e as principais diferenças propostas em relação ao que existe na literatura; no capítulo 3, descreve-se a literatura atual a respeito do tema proposto; no capítulo 4, mostra-se a metodologia que será utilizada para a realização dos experimentos; no capítulo 5, apresentam-se os resultados deste trabalho; e, no capítulo 6, a conclusão.

2 TRABALHOS RELACIONADOS

O trabalho de (ENDO et al., 2022) foi explorado o uso de técnicas de aprendizado de máquina e aprendizado profundo para identificar notícias falsas em comunicações online no idioma português brasileiro relacionadas à pandemia da COVID-19. O trabalho resultou em uma análise exploratória que sugere grande número de notícias falsas sobre a vacina prevaleceram no Brasil, sendo muitas relacionadas a comunicações governamentais.

O trabalho de (ENDO et al., 2022) difere-se deste pois seu principal foco é na detecção de uma notícia falsa, para então analisar o método de classificação de notícia falsa com melhor resultado. Este trabalho já utilizada dados catalogados como falsos, divindo-os e classificando cada vertence de notícia falsa obtida.

O artigo de (JÚNIOR et al., 2020) analisou a busca do termo ”coronavírus”em notícias falsas, apontando resultados para um crescente interesse da população por informações sobre o termo. Foi realizado também um levantamento das medidas tomadas pelo Ministério da Saúde do Brasil e por veículos da mídia tradicional, a fim de combater a disseminação desse tipo de informação. O estudo fez também um comparativo com o aumento de buscas dado alguma nova notícia em relação ao vírus.

O artigo de (JÚNIOR et al., 2020) difere-se deste trabalho porque analisa a busca por informações relacionadas ao COVID-19, não necessariamente a busca por informações falsas, e correlaciona-as com o período em que haviam notícias divulgadas pela mídia. Este trabalho não analisa a busca por termos relacionados ao COVID-19.

O trabalho de (GALHARDI et al., 2020) realizou uma análise sobre as notícias falsas a respeito do coronavíruas mais disseminadas nas redes sociais. Os resultados obtidos mostraram que o WhastApp é o principal canal de compartilhamento de notícias falsas, seguido do Instagram e Facebook, correlacionando a disseminação de conteúdos falsos com o descrédito para com a ciência e instituições globais de saúde.

O estudo de (GALHARDI et al., 2020), apesar de também realizar estudo de notícias falsas relacionadas ao COVID-19, teve como foco a disseminação das mesmas pelas principais redes sociais utilizadas. Este trabalho não faz correlação das notícias falsas com redes sociais.

3 REFERENCIAL TEÓRICO

Neste capítulo, serão apresentadas as revisões da literatura acerca dos temas basilares para a construção desta pesquisa, a saber: o novo coronavírus, a disseminação de notícias falsas e a categorização de notícias falsas.

3.1 Coronavírus

Conforme (TEIXEIRA et al., 2021), coronavírus é um vírus zoonótico, um RNA vírus da ordem *Nidovirales*, da família *Coronaviridae*. Esta é uma família de vírus que causam infecções respiratórias, os quais foram isolados pela primeira vez em 1937 e descritos como tal em 1965, em decorrência do seu perfil na microscopia parecendo uma coroa. Os tipos de coronavírus conhecidos até o momento são: alfa coronavírus HCoV-229E e alfa coronavírus HCoV-NL63, beta coronavírus HCoV-OC43 e beta coronavírus HCoV-HKU1, SARS-CoV (causador da síndrome respiratória aguda grave ou SARS), MERS-CoV (causador da síndrome respiratória do Oriente Médio ou MERS) e SARS-CoV-2, um novo coronavírus descrito no final de 2019 após casos registrados na China. Este provoca a doença chamada de COVID-19.

Segundo os estudos de (HOEK et al., 2004), foi identificada a presença de coronavírus em "[...] camundongos, ratos, galinhas, perus, suínos, cães, gatos, coelhos, cavalos, gado e humanos [...]" e alertaram que essa família de vírus poderia causar "[...] uma variedade de doenças graves, incluindo gastroenterites e doenças do trato respiratório".

Essa família de vírus foi classificada como comum e afirmado que a maioria da população mundial já se infectou ou irá se infectar com algum tipo de coronavírus ao longo da vida, uma vez que é o tipo mais comum de espécies desse vírus é causador de resfriados comuns, enquanto existem tipos mais severos que causam pneumonias com risco de vida (HOEK et al., 2004).

Até 2019, sabia-se que dentro da família *Coronaviridae* existem quatro gêneros – alfacoronavírus, betacoronavírus, gamacoronavírus e deltacoronavírus – e havia seis espécies de coronavírus causadores de doenças humanas – 229E, OC43, NL63 e HKU1, que causam sintomas de resfriado comum, e SARS-CoV e MERS-CoV, que são cepas de origem zoonótica associadas a doenças com síndromes respiratórias por vezes fatais (CHAVES; BELLEI, 2020).

No entanto, em dezembro de 2019, na cidade de Wuhan, na China, foi descoberto o novo agente do coronavírus com capacidade de infectar humanos. O vírus foi descoberto a partir de uma amostra de um grupo de pessoas com pneumonia com causas desconhecidas (JÚNIOR et al., 2020).

O novo agente do coronavírus, o SARS-COV-2, causador da doença COVID-19,

de probabilidade de contágio superior aos anteriores, fez com que, dois meses depois de sua descoberta, o contágio tomasse uma proporção global a ponto de a Organização Mundial de Saúde decretar estado de pandemia (JÚNIOR et al., 2020).

Em 31 de janeiro de 2020, o Ministério da Saúde do Brasil instaurou o Grupo de Trabalho Interministerial de Emergência em Saúde Pública de Importância Nacional e Internacional para acompanhamento da situação e definição de protocolos de ação, para a vigilância do SARS–CoV-2 no país (JÚNIOR et al., 2020).

O acompanhamento do avanço exponencial dos casos da doença COVID-19 fez com que, no dia 3 de fevereiro de 2020, o governo brasileiro decretasse Emergência de Saúde Pública de Importância Nacional, e, no dia 6 de fevereiro deste ano, foi sancionada a Lei da Quarentena para o enfrentamento da pandemia (JÚNIOR et al., 2020).

3.2 *Fake News*

A expressão *fake news* - em português, notícias falsas - denomina a propagação de notícias falsas com objetivo de distorcer fatos intencionalmente, de modo a atrair audiência, enganar, desinformar, induzir a erros, manipular a opinião pública, desprestigiar ou exaltar uma instituição ou uma pessoa, diante de um assunto específico, para obter vantagens econômicas e políticas (GALHARDI et al., 2020).

A expressão *fake news* popularizou-se mundialmente durante a cobertura jornalística da eleição presidencial de 2016, nos Estados Unidos. O termo foi usado na mídia pelo candidato a presidente dos Estados Unidos contra seus adversários, visando a desqualificar informações que favorecessem a candidatura deles (GALHARDI et al., 2020).

No cenário de enfrentamento de pandemia, o uso e disseminação excessiva de notícias falsas relevou uma inquietante perda de confiança nas instituições conhecidas por apresentarem fatos e verdades comprovada, como as próprias instituições de pesquisa e ciência. As informações que contrariaram o conhecimento e comprovação científica não apenas disseminaram medo, mas aumentaram as chances de infecção e propagação da doença (GALHARDI et al., 2020).

Ademais, a febre de desinformação pode fazer com que as pessoas se sintam ansiosas, deprimidas, sobrecarregadas, emocionalmente exaustas e incapazes de atender a demandas importantes. Também pode afetar os processos de tomada de decisões, quando se esperam respostas imediatas e não se dedica tempo suficiente para analisar com cuidado as evidências, afinal, não há controle de qualidade do que é publicado (SOUZA, 2021).

3.3 *Classificação das Fake News*

Conforme o trabalho da UNESCO realizado por (POSETTI; BONTCHEVA, 2020), foram identificados nove temas essenciais presentes em conteúdos associados à pandemia da COVID-19, sendo eles:

1. Origens e propagação do coronavírus/da doença COVID-19;
2. Estatísticas falsas e equivocadas;
3. Impactos econômicos;
4. Desacreditar jornalistas e veículos de notícias fidedignos;
5. Ciência médica: sintomas, diagnóstico e tratamento;
6. Impactos na sociedade e no meio ambiente;
7. Politização;
8. Conteúdo impulsionado para ganho financeiro fraudulento;
9. Desinformação cujo foco são as celebridades.

Os temas identificados no trabalho de (POSETTI; BONTCHEVA, 2020) variam desde informações falsas sobre a origem do vírus, a incidência, os sintomas e os remédios para o tratamento, até ataques políticos contra jornalistas. Os formatos mais utilizados incluem: construções de narrativas e memes com alto teor emotivo; imagens e vídeos fabricados, alterados de forma fraudulenta ou descontextualizados; infiltrações e campanhas orquestradas de desinformação; e fontes, páginas de internet e bases de dados falsas.

4 METODOLOGIA DA PESQUISA

O estudo conduzido foi dividido nas seguintes etapas: tratamento e análise de dados coletados com informações falsas a respeito da COVID-19, mineração de texto e agrupamento de informações coletadas.

4.1 Tratamento e análise de dados

Neste trabalho, foi utilizada a base de dados pública do trabalho de (ENDO et al., 2022), com informações obtidas do Boatos.org e atualizadas para o período de 2020 a 2022 relacionadas ao tópico COVID-19. Todas as informações coletadas são classificadas como falsas pelo Boatos.org, um site independente que faz levantamento de notícias falsas sobre os mais diversos tópicos. Mais informações a respeito da coleta de dados estão disponíveis no trabalho de (ENDO et al., 2022).

Para realizar a definição de próximas etapas deste projeto, foi feita uma análise da base de dados. A análise consiste na visualização de um dos principais pontos mais relevantes para esta pesquisa, os termos mais frequentes observados no *dataset*, para então, empiricamente, observar os principais temas abordados na criação de notícias falsas referentes ao COVID-19.

4.2 Mineração de texto

A mineração de texto consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural para objetivos específicos (ARANHA; PASSOS, 2006). Isto é, uma série de procedimentos realizados para a transformação de conhecimento implícito para conhecimento explícito.

A mineração de texto tem suas bases descritas primordialmente no trabalho de (FELDMAN; DAGAN, 1995), em que os autores descrevem uma forma de extração de conhecimento a partir de uma coleção de dados de texto.

Conforme (FELDMAN; SANGER et al., 2007), sistemas de mineração de texto recebem como entrada documentos em formato bruto e podem gerar vários tipos de saídas, tais como padrões, predições, etc. Os autores ainda fornecem uma arquitetura geral de sistemas de mineração de texto, que, de forma mais específica, divide-se em: coleta de documentos, pré-processamento, extração de conhecimento e, por fim, avaliação e interpretação dos resultados (GONÇALVES, 2002).

A coleta de documentos é a etapa inicial na qual os documentos são extraídos da(s) fonte(s) de domínio do problema. Os documentos coletados são, então, estruturados em um formato legível pelo computador e manipulável pelo(a) analista, tais como CSV

(do inglês, *Comma Separated Values*) e JSON (do inglês, *JavaScript Object Notation*), a fim de facilitar sua leitura e processamento.

A etapa de pré-processamento consiste em um conjunto de operações aplicadas sobre o conjunto de dados textuais de modo a se construir a matriz de termos do documento (do inglês, *Document-Term Matrix*, ou simplesmente DTM).

A extração de conhecimento compreende o aprendizado de máquina e utilização de algoritmos que irão extrair conhecimento da DTM na forma de um classificador de texto.

Na última etap,a de avaliação e interpretação dos resultados, as técnicas utilizadas são avaliadas por algum método de desempenho e/ou comparação, de modo a fornecer uma avaliação objetiva para a tomada de decisão.

4.2.1 Coleta de documentos

Neste trabalho, a coleta de documentos foi realizada pelo trabalho de (ENDO et al., 2022). Os dados gerados em CSV são utilizados nas etapas seguintes deste trabalho.

4.2.2 Pré-processamento

Esta etapa tenta identificar similaridades em função da morfologia ou do significado dos termos nos textos (MORAIS; AMBRÓSIO, 2007).

4.2.2.1 Remoção de palavras vazias

Esta etapa consiste na remoção de palavras que não trazem significado algum para o processo de análise de texto (WIVES, 2002). Geralmente, palavras vazias (ou *stopwords*, em inglês) são classes gramaticais auxiliares como preposições, pronomes, artigos, advérbios e afins.

Neste trabalho, as palavras vazias aplicadas foram todas em português brasileiro, visto que este trabalho propõe-se a analisar somente notícias falsas no Brasil.

4.2.2.2 Normalização morfológica

A normalização morfológica consiste em transformar os termos do *corpus* analisado em sua forma raiz. Esta subetapa é importante para agrupar termos morfologicamente diferentes quando foram redigidos e que, no entanto, possuem a mesma configuração morfológica raiz. As técnicas mais comuns de normalização morfológica são *stemming* e *lemmatization*. A técnica adotada neste trabalho foi a *stemming*.

Stemming objetiva a identificação dos radicais da palavra e consequente eliminação dos prefixos e/ou sufixos, adicionando os radicais aos índices dos termos (MORAIS; AMBRÓSIO, 2007).

4.2.2.3 Filtragem e ponderação

Visto que nem todos os termos presentes na DTM contribuem positivamente para a etapa de análise de texto, nem mesmo após a remoção de *stopwords*, a abordagem utilizada neste trabalho e em outros trabalhos encontrados na literatura, consiste em atribuir pesos aos termos do *corpus*.

Para realizar a ponderação, foi utilizada a técnica TF-IDF (em inglês, *Term Frequency-Inverse Document Frequency*, ou frequência do termo–inverso da frequência nos documentos).

A TF-IDF é uma estatística numérica que reflete a importância de uma palavra para um documento da coleção ou corpus (SALTON; BUCKLEY, 1988). Este método é frequentemente usado como um fator de ponderação na recuperação de informações e mineração de texto. O TF-IDF é usado principalmente para interromper a filtragem de palavras na aplicação de categorização e resumo de texto. Por convenção, o valor de TF-IDF aumenta proporcionalmente ao número de vezes que uma palavra aparece em um documento, mas é compensada pela frequência da palavra no *corpus*, o que ajuda a controlar o fato de que algumas palavras são mais comuns do que outras. O termo de frequência significa a frequência bruta de um termo em um documento. Além disso, o termo referente à frequência inversa do documento é uma medida de se o termo é comum ou raro em todos os documentos em que pode ser obtido dividindo o número total de documentos pelo número de documentos que contêm o termo (MUNOT; GOVILKAR, 2014).

Essencialmente, o TF-IDF funciona determinando a frequência relativa de palavras em um documento específico em comparação com a proporção inversa dessa palavra em todo o *corpus* do documento. Intuitivamente, esse cálculo determina a relevância de uma determinada palavra em um determinado documento. Palavras comuns em um único ou em um pequeno grupo de documentos tendem a ter números TF-IDF mais altos do que palavras comuns, como artigos e preposições (RAMOS et al., 2003).

Dada uma coleção de documentos D, uma palavra w e um documento individual $d \in D$, calcula-se:

$$w_d = f_{w,d} * log(|D|/f_{w,D}) \tag{1}$$

em que $f_{w,d}$ é igual ao número de vezes que w aparece em d, $|D|$ é o tamanho do *corpus*, e $f_{w,D}$ é igual ao número de documentos em que w aparece em D (SALTON; BUCKLEY, 1988). Há algumas situações diferentes que podem ocorrer para cada palavra, dependendo dos valores de $f_{w,d}$, $|D|$ e $f_{w,D}$ (RAMOS et al., 2003).

Assumindo que $|D|$ $f_{w,D}$, isto é, o tamanho do *corpus* é aproximidamente igual à frequência de w sobre D. Se $1 < log(|D|/f_{w,D}) < c$ para alguma constante muito pequena c, então w_d será menor que $f_{w,d}$ mas ainda positivo. Isso implica que w é relativamente comum por todo o *corpus* mas ainda possui alguma importância por todo D. Um exemplo seria, por exemplo, o uso de pronomes e preposições ao longo de textos, os quais não

possuem relevância de significa ao longo do texto (a menos que esteja sendo estudado explicitamente o uso de tais palavras). Tais palavras comuns possui pontuação de TF-IDF muito baixa, tornando-as essenciamente insignificantes na busca.

Por fim, supondo que $f_{w,d}$ é grande e $f_{w,D}$ é pequeno. Então, $log(|D|/f_{w,D})$ será bastante grande, e então w_d será igualmente grande. Este é o caso de estudo do TF-IDF, uma vez que palavras com alto w_d implicam que w é uma palavra importante em d mas não é comum em D. O termo w é dito como o qual possui maior poder discriminatório. Portanto, quando uma *query* contém o mesmo w, retornar um documento d em que w_d é algo, irá muito provavelmente satisfazer o usuário.

4.2.2.4 *Principal Component Analysis - PCA*

A análise de componentes principais (PCA, em inglês) é uma técnica multivariada de modelagem da estrutura de covariância, que transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, em um conjunto substancialmente menor de variáveis não-correlacionadas e que contêm a maior parte da informação do conjunto original (HONGYU; SANDANIELO; JUNIOR, 2016).

O principal objetivo do PCA é o de explicar a estrutura da variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, por meio de combinações lineares das variáveis originais. Essas combinações lineares são denominadas componentes principais e são não-correlacionadas entre si (HONGYU; SANDANIELO; JUNIOR, 2016).

Neste trabalho, após aplicar a técnica de TF-IDF para ponderação dos termos do *corpus*, foi aplicada a técnica de *Principal Component Analysis* a fim de reduzir a dimensionalidade dos dados.

4.3 Agrupamento de informações coletadas

Agrupamento, ou *clustering* (termo em inglês), é a divisão de dados em grupos de objetos similares. Cada grupo, chamado de *cluster*, consiste em objetos que são similares entre os mesmos e diferentes quando comparados com objetos de outros grupos. A representação de dados por menos *clusters* necessariamente perde certos detalhes finos, mas alcança a simplificação. Ela representa muitos objetos de dados por poucos *clusters* e, portanto, modela os dados por seus *clusters* (HAN; KAMBER; MINING, 2006).

Por sua vez, a análise por agrupamento é a organização da coleção de padrões em *clusters* baseado na similaridade. Padrões de objeto dentro de um mesmo *cluster* são mais similares uns aos outros do que a padrões que pertencem a *clusters* diferentes. No entanto, é importante entender a diferença entre *clustering* (classificação não-supervisionada) e análise discriminatória (classificação supervisionada). Conforme (ABBAS, 2008), em classificação supervisionada, temos uma coleção de padrões rotulados (pré-classificados); o problema é rotular um padrão recém-encontrado, mas não rotulado. Normalmente, os padrões rotulados

(treinamento) fornecidos são usados para rotular um novo padrão. No caso de *clustering*, o problema é agrupar uma determinada coleção de padrões não rotulados em *clusters* significativos. De certa forma, os rótulos também estão associados a *clusters*, mas esses rótulos de categoria são orientados por dados; ou seja, são obtidos apenas a partir dos dados.

Para este estudo, foi utilizado o algoritmo *K-Means* para agrupamento dos dados

4.3.1 *K-Means*

O algoritmo *k-means* é uma abordagem incremental para o agrupamento que adiciona dinamicamente um centro de *cluster* por vez por meio de um procedimento de pesquisa global determinístico que consiste em execuções do algoritmo *k-means* na mesma quantidade que o tamanho do conjunto de dados (LIKAS; VLASSIS; VERBEEK, 2003).

O algoritmo *k-means* depende do vlaor de k, o qual sempre precisa ser especificado para executar qualquer análise de agrupamento. O agrupamento com diferentes valores de k acabará por produzir diferentes resultados.

A associação ao *cluster* é determinada calculando o centróide para cada grupo (a versão multidimensional da média) e atribuindo cada objeto ao grupo com o centróide mais próximo. Essa abordagem minimiza a dispersão geral dentro do *cluster* pela realocação iterativa dos membros do *cluster* (ABBAS, 2008).

Consoante à definição de (ABBAS, 2008), um algoritmo de k-particionamentos tem como entrada um conjunto S de objetos e um inteiro k, gerando uma partição S de subconjuntos $S_1, ...S_2, S_k$. Ele usa a soma dos quadrados como critério de otimização. Sendo x_r^i o r-ésimo elemento de S_i, $|S_i|$ é o número de elementos em S_i, e $d(x_r^i, x_s^i)$ é a distância entre x_r^i e x_s^i. O critério da soma dos quadrados é definido pelo função de custo:

$$c(S_i) = \sum_{r=1}^{|S_i|} \sum_{x=1}^{|S_i|} (d(x_r^i, x_s^i))^2 \tag{2}$$

Em particular, *k-means* funciona através do cálculo da centróide de cada *cluster* S_i, denotado por x^{-i}, e pela otimização da função de custo:

$$c(S_i) = \sum_{r=1}^{|S_i|} (d(x^{-i}, x_r^i))^2 \tag{3}$$

O objetivo do algoritmo é o de minimizar o custo total:

$$c(S_i) + ... + c(S_k) \tag{4}$$

O pseudo-código a seguir exemplifica o funcionamento do algoritmo *k-means*.

A popularidade do *k-means* como escolha para algoritmo de agrupamento deve-se aos seguintes fatores, conforme (ABBAS, 2008):

Algorithm 1 K-Means

Definir K como o número de *clusters*
Inicializar os vetores de cada *cluster* K
Para cada novo vetor:

- Computar a distância entre o novo vetor com todos os outros vetores do *cluster*
- Recomputar a distância mais próxima dos vetores com o novo vetor, usando uma taxa de aprendizagem que diminui a cada iteração

- Sua complexidade de tempo é de $O(nkl)$, em que n é o número de padrões, k é o número de *clusters* e l é o número de iterações obtidas pelo algoritmo até convergir;
- Sua complexidade de espaço é de $O(k + n)$. O mesmo requer espaço adicional para armazenar a matriz de dados;
- É independente de ordem; para um determinado conjunto de sementes inicial de centros de *clusters*, ele gera a mesma partição dos dados, independentemente da ordem em que os padrões são apresentados ao algoritmo

4.3.1.1 *Definição de K*

O método *Elbow* (cotovelo, em inglês) é um método utilizado para produzir informações na determinação do melhor número de *clusters*, observando a porcentagem da comparação entre o número de *clusters* que formarão um cotovelo em um ponto (NAINGGOLAN et al., 2019).

Este método fornece sugestões de seleções de valores de *cluster* e, em seguida, adicionando o valor do cluster a ser usado como modelo de dados na determinação do melhor cluster. Além disso, a porcentagem do cálculo resultante é uma comparação entre o número de clusters adicionados (NAINGGOLAN et al., 2019).

Diferentes resultados percentuais de cada valor de *cluster* podem ser mostrados usando o gráfico como fonte de informação. Se o valor do primeiro *cluster* com o valor do segundo *cluster* der o ângulo no gráfico ou o valor tem a maior diminuição, então o valor do *cluster* é o melhor. Para obter uma comparação, é necessário calcular a SSE (soma do erro quadrado, em português) de cada valor de *cluster*. Isso porque, quanto maior o número de cluster K, menor será o valor de SSE (NAINGGOLAN et al., 2019).

As etapas de definição de K pelo método de *Elbow* são:

1. Inicializar K com um primeiro valor;
2. Aumentar o valor de K;
3. Calcular o resultado da soma do erro quadrado para cada valor de K;
4. Analisar a soma dos resultados do erro quadrático do valor de K que diminuiu drasticamente;
5. Localizar e definir o valor de K em forma de cotovelo.

A SSE é um dos métodos estatísticos usados para medir a diferença total do valor real para o valor alcançado (NAINGGOLAN et al., 2019), calculado por:

$$SSE = \sum_{i=1}^{n} (d)^2 \tag{5}$$

em que d é a distância entre os dados e o centro do *cluster*.

A SSE é uma fórmula usada para medir a diferença entre os dados obtidos pelo modelo de previsão que foi feito anteriormente. Também é frequentemente usado como referência de pesquisa na determinação de *clusters* ideais (NAINGGOLAN et al., 2019).

4.4 Fluxograma

As etapas da metodologia adotada neste trabalho podem ser melhor visualizadas na figura 1 a seguir.

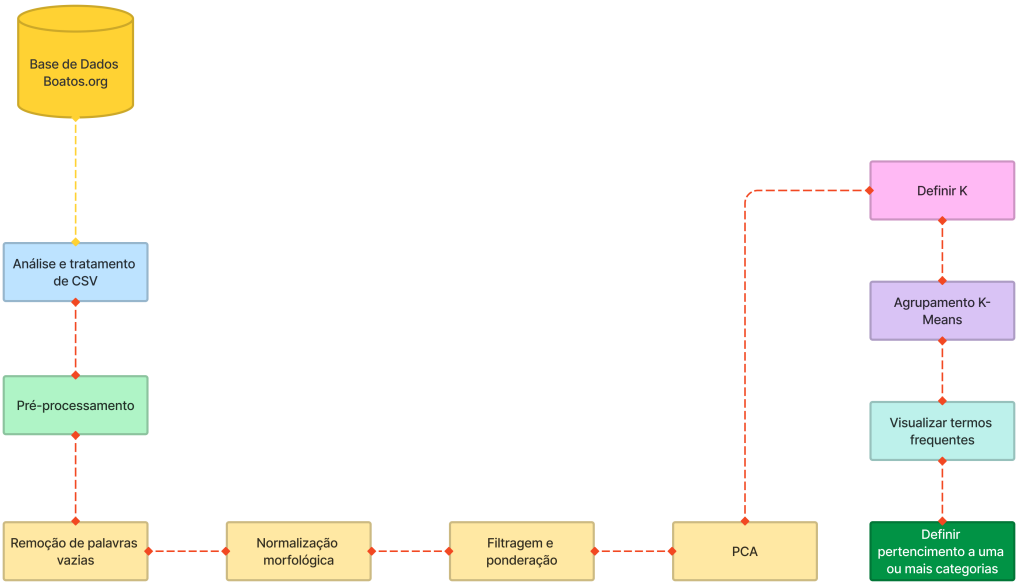


Figura 1 – Fluxograma das etapas realizadas neste trabalho. Fonte: Autoria Própria (2022).

5 RESULTADOS

Este capítulo apresenta os resultados obtidos seguindo as etapas descritas nos métodos da pesquisa, estando divido em quatro subseções. Na primeira subseção, trata-se da definição do problema e do intervalo de datas adotado para tratamento de dados. Na segunda subseção, mostra-se os resultados obtidos pelo método Elbow e como este foi utilizado para definir quantidade de agrupamentos a tratar. Na terceira subseção, exibe-se os resultados do agrupamento de dados com a quantidade previamente definida e trazendo a visualização dos termos mais frequentes de cada grupo e a associação dos termos com categorias de *fake news* associadas ao COVID-19. Na quarta e última seção, . Na quarta seção, serão exibidas as frequências de cada categoria por período e por ano.

5.1 Tratamento da base de dados

O *dataset* obtido contém dados de 26 de janeiro de 2020 a 30 de novembro de 2022, totalizando 893 notícias falsas reportadas no site Boatos.org. Para melhor verificar as notícias obtidas por período, foi realizada a seguinte filtragem de dados: o *dataset* foi subdivido por período de 6 meses para cada ano.

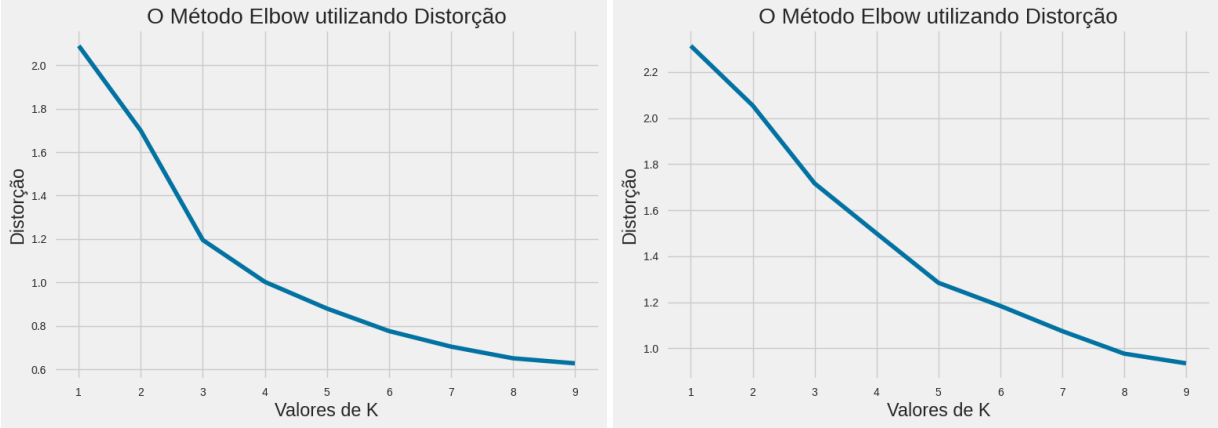
- Janeiro a junho de 2020;
- Julho a dezembro de 2020;
- Janeiro a junho de 2021;
- Julho a dezembro de 2021;
- Janeiro a junho de 2022;
- Julho a dezembro de 2022.

Isto foi definido de forma empírica, a fim de buscar dados de cada período que assemelhavam-se às notícias obtidas e realidade de cada semestre: surgimento do vírus, número de infectados, óbtidos, tratamentos realizados, período em que saíram notícias relacionadas às vacinas, etc.

5.2 Método *Elbow* e definição de *K*

O método *Elbow* foi implementado para ajudar os cientistas de dados a selecionar o número ideal de *clusters*, ajustando o modelo com um intervalo de valores para K. Se o gráfico de linhas assemelhar-se a um braço, então o "cotovelo"(o ponto de inflexão na curva) é uma boa indicação de que o modelo subjacente ajusta-se melhor naquele ponto. No visualizador destes resultados, o "cotovelo"indicado está sendo exibido com uma linha tracejada.

O método *Elbow* utilizado teve a métrica do parâmetro de pontuação definida por distorção, a qual calcula a soma das distâncias quadradas de cada ponto até seu centro



(a) Janeiro a junho com $K = 4$ (b) Julho a dezembro com $K = 5$

Figura 2 – Método *Elbow* por distorção com dados do primeiro (a) e segundo (b) semestre de 2020. Fonte: Autoria Própria (2022).

atribuído. O método utilizado também exibe a quantidade de tempo para treinar o modelo de agrupamento através de uma linha verde tracejada.

Os resultados podem ser visualizados nas figuras 2 a 4.

Para definição de K, foi escolhido um intervalo de 1 a 10 possíveis *clusters* a partir dos dados de PCA do primeiro e segundo semestre de 2020 – cada um feito separadamente. A partir dos dados de PCAs com ambos os semestres, foi calculada a média das distâncias ao quadrado dos centros dos *clusters* dos respectivos *clusters* utilizando a métrica de distância euclidiana.

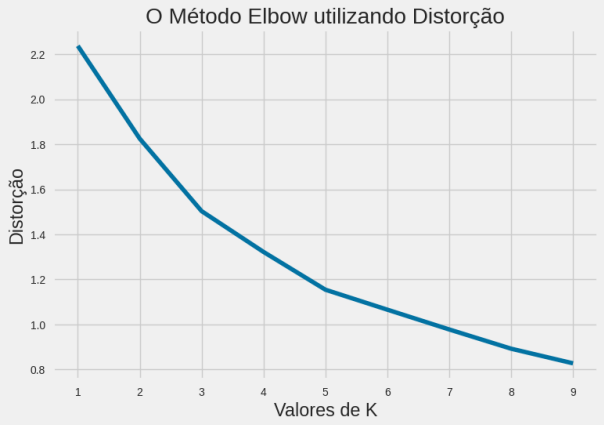
Os resultados são mapeados para cada possível K – como dito previamente, em um intervalo de 1 a 10.

Para determinar o número ideal de *clusters*, é preciso selecionar o valor de K no ”cotovelo”, ou seja, no ponto a partir do qual a distorção começa a diminuir de forma linear.

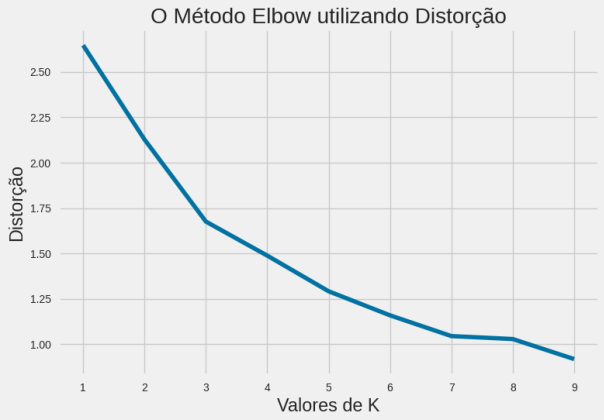
Na figura 2.a, o valor de K começa a diminuir de forma linear quando $K = 4$. Com raciocínio semelhante, o valor de K escolhido na figura 2.b foi $K = 5$.

Com a mesma definição de K em um intervalo de 1 a 10, e mesma métrica do parâmetro de pontuação definida por distorção. Na figura 3.a, o valor de K começa a diminuir de forma linear quando $K = 5$, enquanto que na figura 3.b, o valor de K é definido para 3.

Igualmente, K definido para valores entre 1 a 10, fazendo uso da métrica do parâmetro de pontuação por distorção e com raciocínio similar aos métodos descritos anteriormente nesta seção, na figura 4.a, de janeiro a junho de 2022, o valor de K começa a diminuir de forma linear quando $K = 4$, ao passo que na figura 4.b, o seu valor adotado é $K = 3$.

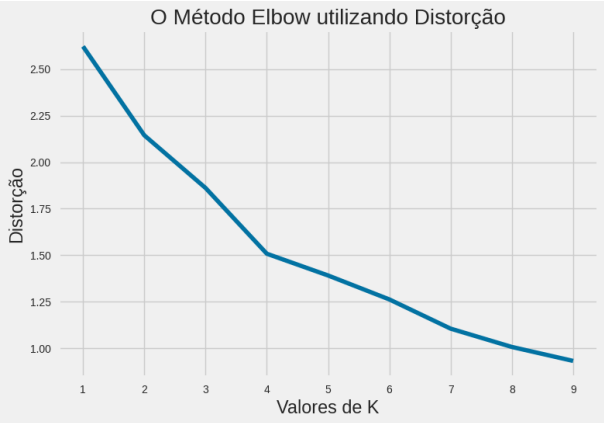


(a) Janeiro a junho com $K = 5$

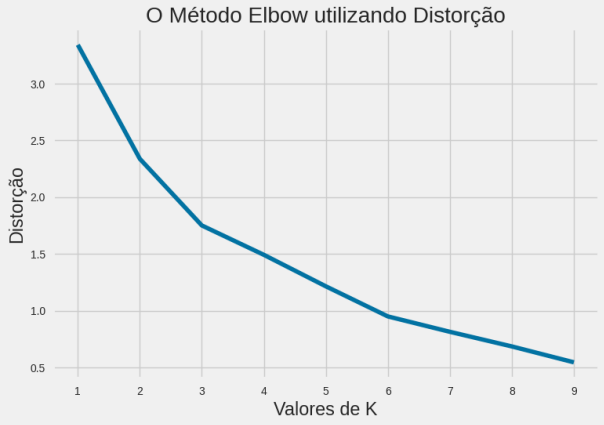


(b) Julho a dezembro com $K = 3$

Figura 3 – Método *Elbow* por distorção com dados do primeiro (a) e segundo (b) semestre de 2021. Fonte: Autoria Própria (2022).



(a) Janeiro a junho com $K = 4$



(b) Julho a dezembro com $K = 3$

Figura 4 – Método *Elbow* por distorção com dados do primeiro (a) e segundo (b) semestre de 2022. Fonte: Autoria Própria (2022).

5.3 Agrupamento de dados para cada *K*

Para a coluna de PCAs gerados em cada período estudado, foi utilizado o algoritmo de agrupamento *K-Means* para o respectivo K obtido pelo método *Elbow*. Os resultados obtidos por período são exibidos a seguir.

5.3.1 Janeiro a junho de 2020

Dada a coluna de PCAs gerados para cada título de notícia no período de janeiro a junho de 2020, e com número de *clusters* definido para $K = 4$, o resultado pode ser visualizado na figura 5.

Dado o resultado obtido na figura 5, é possível observar 3 *clusters* com resultados muito próximos e semelhantes entre si, podendo indicar pertencimento a uma ou mais categoria de *fake news* em comum, ao passo que o quatro *cluster* pode ser associado com

uma categoria exclusiva.

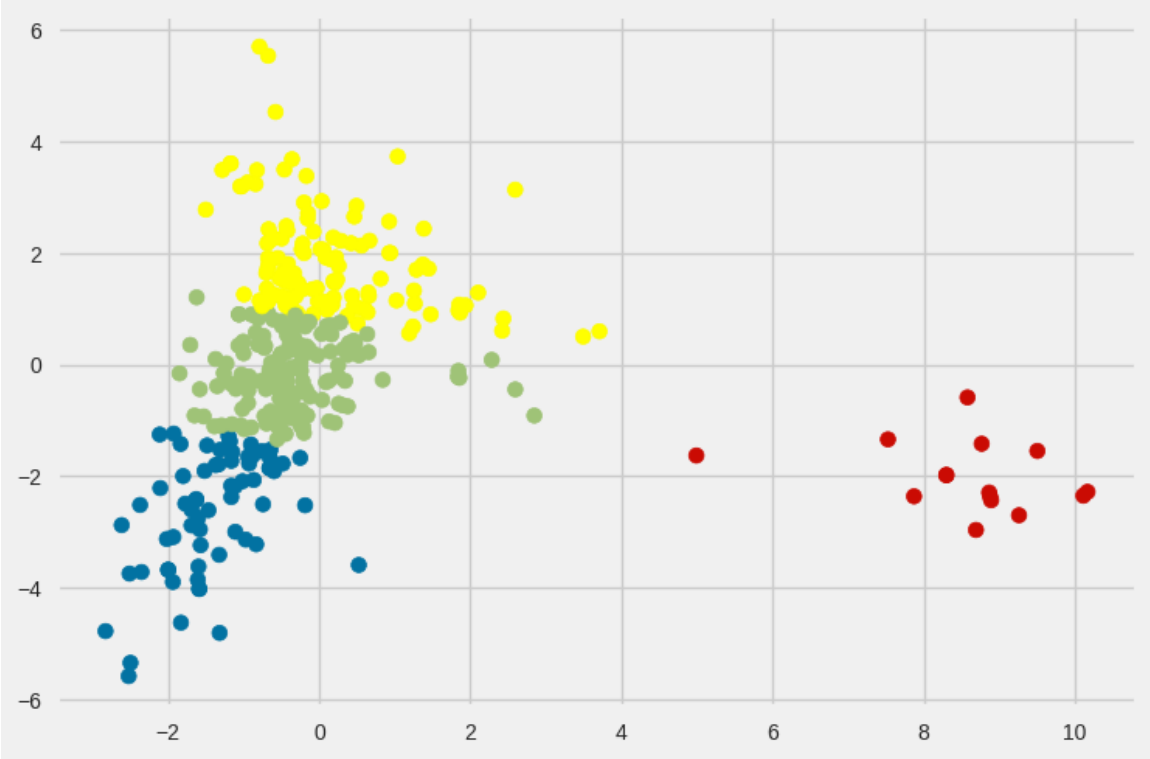


Figura 5 – Método de agrupamento *K-Means* aplicado para $K = 4$ com dados de *Principal Component Analysis* no período de janeiro a junho de 2020. Fonte: Autoria Própria (2022).

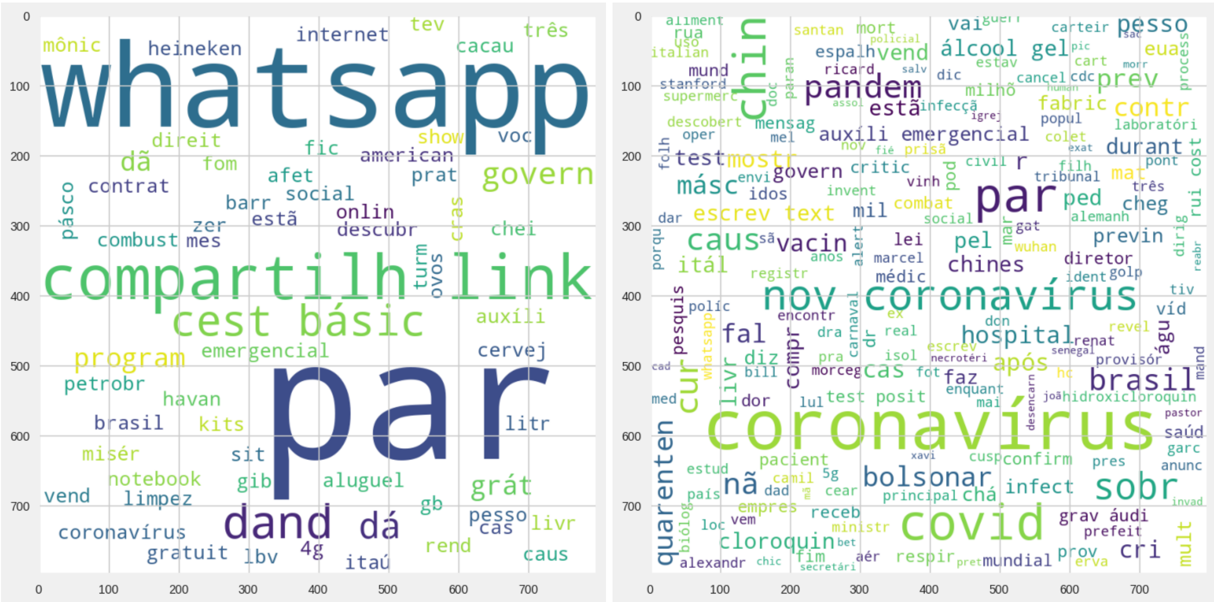
Para melhor avaliar o que pode ser interpretado do resultado obtido no *K-Means*, é interessante analisar quais os termos mais frequentes associado a cada *cluster*. Ao realizar a correlação do PCA associado com os dados de seus respectivos *clusters*, é possível obter os termos mais evidentes das reportagens falsas.

Alguns dos termos mais frequentes para o cenário de *cluster* associado com $n = 0$, são: ”par”, ”compartilh”, ”link”, ”whatsapp”, ”dand”, ”cest”, ”básic”, ”dá”, ”govern”. O resultado pode ser enquadrado nas categorias: 2) Estatísticas falsas e equivocadas; 4) Desacreditar jornalistas e veículos de notícias fidedignos; 8) Conteúdo impulsionado para ganho financeiro fraudulento. Vale ressaltar que o *cluster* $n = 0$, exibido na figura 6.a, é o cluster mais distante exibido na figura 5.

Por sua vez, nos *clusters* mais próximos entre si, são exibidos muitos termos em comum em 6.b, 6.c e 6.d. Os termos mais comuns são: ”coronavírus”, ”covid”, ”chin”, ”par”.

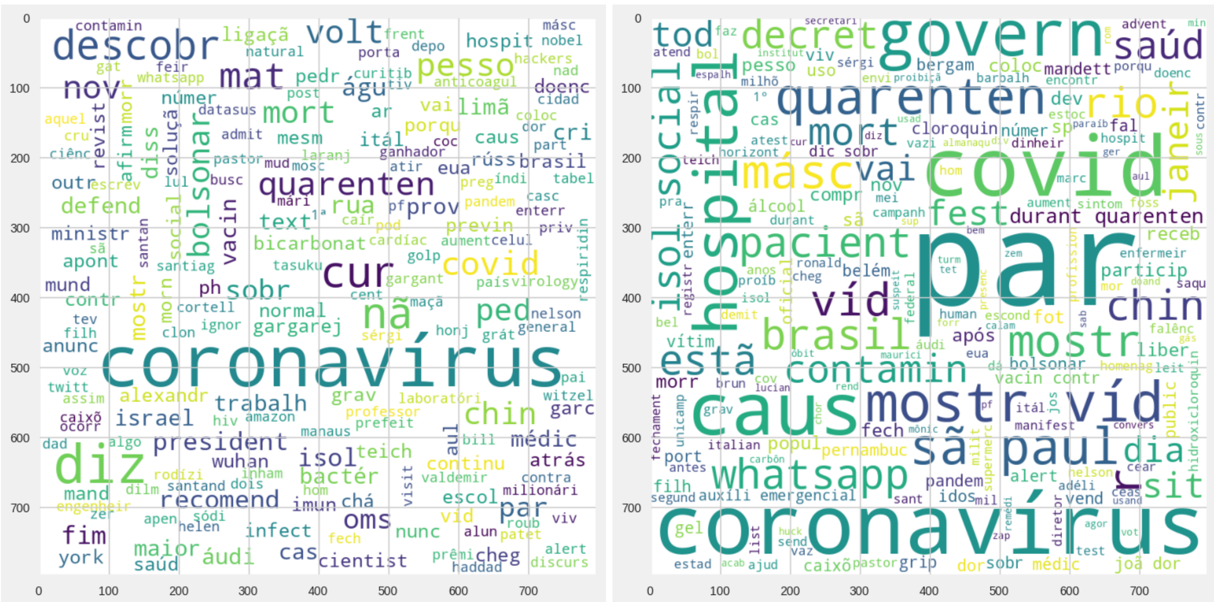
No *cluster* referente à figura 6.b, os termos mais frequentes são: ”coronavírus”, ”covid”, ”par”, ”nov”, ”chin”, ”sobr”, ”pandem”, ”caus”, ”cur”, ”brasil”. O resultado pode ser enquadrado nas categorias: 1) Origens e propagação do coronavırus/da doença COVID-19; 5) Ciência médica: sintomas, diagnóstico e tratamento.

No *cluster* referente à figura 6.c, os termos mais frequentes são: ”coronavírus”, ”diz”, ”cur”, ”nã”, ”descobr”, ”mat”, ”chin”, ”quarenten”, ”volt”, ”nov”, ”covid”. O resultado pode ser



(a) Termos mais frequentes do *cluster* associado com *label* n = 0

(b) Termos mais frequentes do *cluster* associado com *label* n = 1



(c) Termos mais frequentes do *cluster* associado com *label* n = 2

(d) Termos mais frequentes do *cluster* associado com *label* n = 3

Figura 6 – Nuvem de palavras mais frequentes formadas pelos títulos das reportagens falsas para os quatro *clusters* formados no período de janeiro a junho de 2020. Fonte: Autoria Própria (2022).

associado às categorias: 1) Origens e propagação do coronavırus/da doença COVID-19; 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos.

No *cluster* referente à figura 6.d, os termos mais frequentes são: ”par”, ”coronavírus”, ”covid”, ”caus”, ”mostr”, ”víd”, ”govern”, ”hospital”, ”quarenten”. O resultado poder ser enquadrado nas categorias: 5) Ciência médica: sintomas, diagnóstico e tratamento; 7) Politização; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos.

5.3.2 Julho a dezembro de 2020

Semelhante ao que foi realizado na subseção anterior, dada a coluna de PCAs gerados para cada título de notícia no período de julho a dezembro de 2020, e com número de *clusters* definido para $K = 5$, o resultado pode ser visualizado na figura 7.

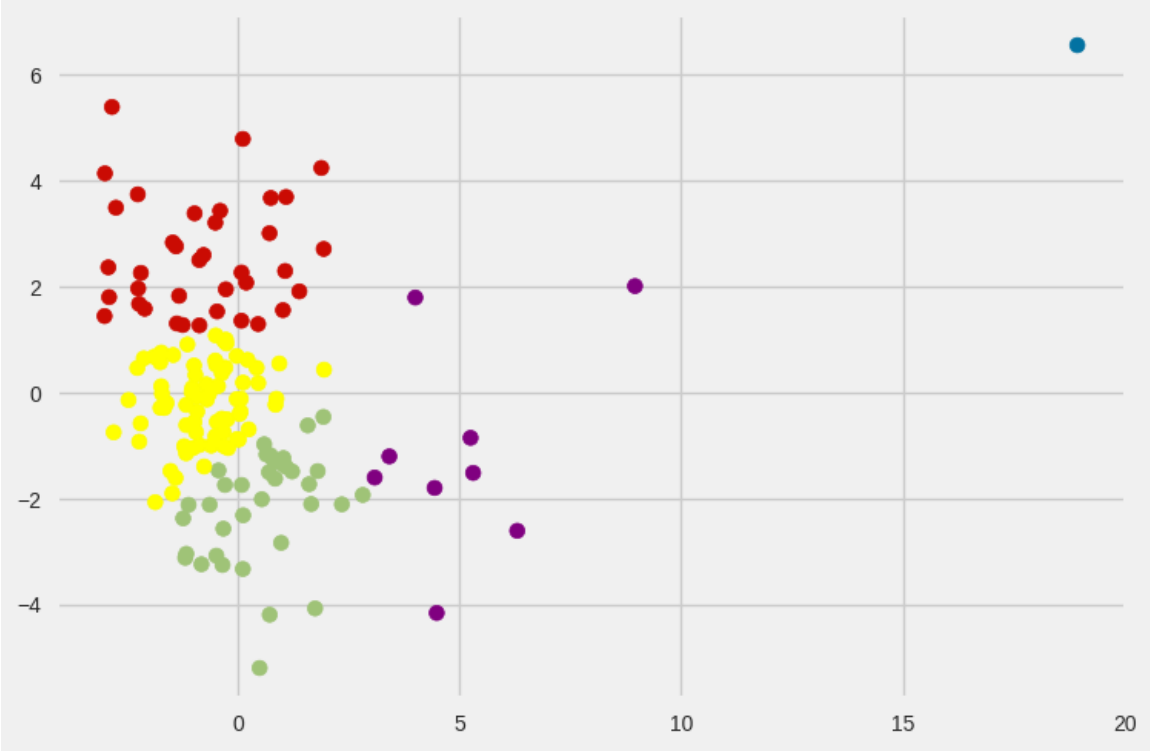


Figura 7 – Método de agrupamento *K-Means* aplicado para $K = 5$ com dados de *Principal Component Analysis* no período de julho a dezembro de 2020. Fonte: Autoria Própria (2022).

Dado o resultado exibido na fig. 7, é possível observar quatro *clusters* com resultados muito próximos e semelhantes entre si, indicando pertencimento a uma ou mais categorias de *fake news* em comum, ao passo que há um quinto *cluster* que diferencia-se dos demais.

Para melhor avaliar o que pode ser interpretado do resultado obtido no *K-Means* da figura 7, é interessante analisar quais os termos mais frequentes associado a cada *cluster*.

De forma similar ao que foi escrito na subseção anterior, os termos mais evidentes das reportagens falsas são exibidos na figura 8.

No *cluster* referente à figura 8.a, alguns dos termos mais frequentes são: ”covid”, ”pandem”, ”mort”, ”caus”, ”durant”, ”cloroquin”, ”mil”, ”port”, ”hidroxicloroquin”, ”tev”, ”acab”. O resultado pode evidenciar pertencimento às seguintes categorias de notícias falsas: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos.

No *cluster* referente à figura 8.b, alguns dos termos mais frequentes são: ”vacin”, ”covid”, ”contr”, ”tom”, ”par”, ”chines”, ”chin”, ”nã”, ”pesso”, ”dna”. O resultado pode indicar pertencimento às seguintes categorias de notícias falsas: 1) Origens e propagação do coronavırus/da doença COVID-19; 6) Impactos na sociedade e no meio ambiente; 5) Ciência médica: sintomas, diagnóstico e tratamento.

No *cluster* da figura 8.c, alguns dos termos mais frequentes evidenciados são: ”vacin”, ”fal”, ”verdad”, ”covid”, ”fars”, ”médic”, ”cert”, ”nã”, ”dna”, ”diz”, ”alert”. O resultado pode indicar pertencimento às seguintes categorias de notícias falsas: 1) Origens e propagação do coronavırus/da doença COVID-19; 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos.

No *cluster* da figura 8.d, os termos mais frequentes evidenciados são: ”covid”, ”par”, ”vacin”, ”contr”, ”másc”, ”caus”, ”nã”, ”test”, ”após”, ”mostr”, ”morr”, ”víd”. O resultado pode ser associado às categorias de: 5) Ciência médica: sintomas, diagnóstico e tratamento; 6) Impactos na sociedade e no meio ambiente; 2) Estatısticas falsas e equivocadas.

Por fim, no *cluster* da figura 8.e, os termos mais frequentes são: ”verdad”, ”médic”, ”la”, ”fal, ”relat”, ”pandem”, ”covid”, ”fars”, ”vacin”, ”grip”, ”5g”. O resultado pode indicar pertencimento às seguintes categorias de notícias falsas: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos.

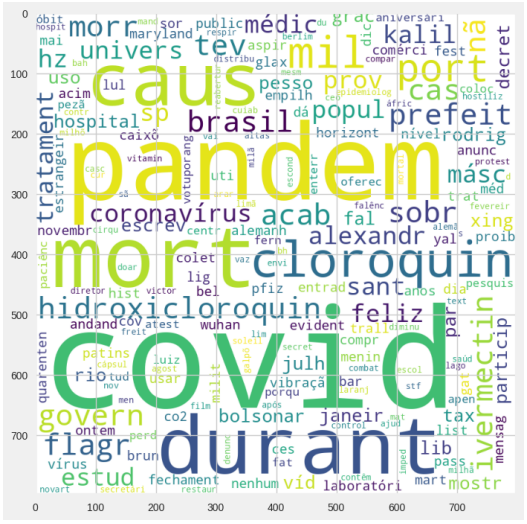
5.3.3 Janeiro a junho de 2021

Semelhante ao que foi realizado nas subseções anteriores, dada a coluna de PCAs gerados para cada título de notícia no período de janeiro a junho de 2021, e com número de *clusters* definido para $K = 5$, o resultado pode ser visualizado na figura 9.

Dado o resultado exibido na fig. 9, é possível avaliar que todos os *clusters* possuem similaridades entre si, sem haver grandes diferenças de proximidade entre si.

De forma similar ao que foi trabalhado nas subseções anteriores, para melhor avaliar o que pode ser interpretado do resultado obtido no *K-Means* da figura 9, é interessante analisar quais os termos mais frequentes associado a cada *cluster*. Os termos evidenciados são exibidos na figura 10.

No *cluster* referente à figura 10.a, os termos frequentes em evidência são: ”covid”, ”caus”, ”tratament”, ”precoc”, ”morr”, ”mort”, ”contr”, ”vacin”, ”zer”, ”hidroxicloroquin”, ”ivermectin”. O resultado sugere que esse grupo pode pertencer às seguintes categorias



(a) Termos mais frequentes do *cluster* associado com *label* n = 0

(b) Termos mais frequentes do *cluster* associado com *label* n = 1

(c) Termos mais frequentes do *cluster* associado com *label* n = 2

(d) Termos mais frequentes do *cluster* associado com *label* n = 3

(e) Termos mais frequentes do *cluster* associado com *label* n = 4

Figura 8 – Nuvem de palavras mais frequentes formadas pelos títulos das reportagens falsas para os quatro *clusters* formados no período de julho a dezembro de 2020. Fonte: Autoria Própria (2022).

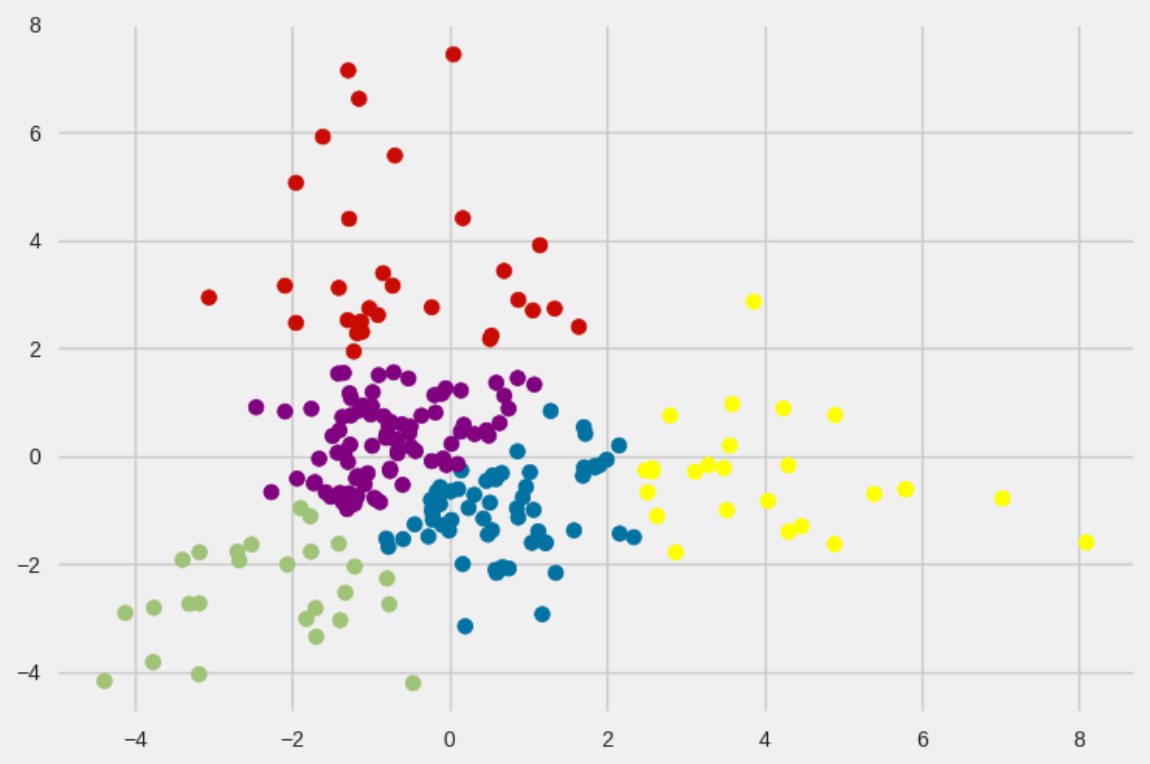


Figura 9 – Método de agrupamento *K-Means* aplicado para $K = 5$ com dados de *Principal Component Analysis* no período de janeiro a junho de 2021. Fonte: Autoria Própria (2022).

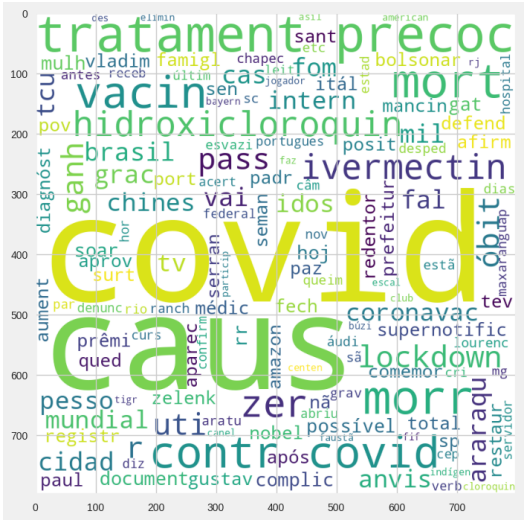
de notícias falsas: 5) Ciência médica: sintomas, diagnóstico e tratamento; 2) Estatısticas falsas e equivocadas; 1) Origens e propagação do coronavırus/da doença COVID-19.

No *cluster* referente à figura 10.b, os termos mais frequentes são: ”vacin”, ”nã”, ”covid”, ”diz”, ”tom”, ”par”, ”prov”, ”contr”, ”pfiz”, ”médic”. O resultado sugere pertencimento às seguintes categorias: 5) Ciência médica: sintomas, diagnóstico e tratamento; 6) Impactos na sociedade e no meio ambiente; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos.

No *cluster* referente à figura 10.c, os termos em evidência são: ”vacin”, ”par”, ”mostr”, ”víd”, ”covid”, ”nã”, ”contr”, ”pandem”, ”bolsonar”, ”govern”. O resultado sugere pertencimento às seguintes categorias: 5) Ciência médica: sintomas, diagnóstico e tratamento; 9) Desinformação cujo foco são as celebridades; 7) Politização; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos.

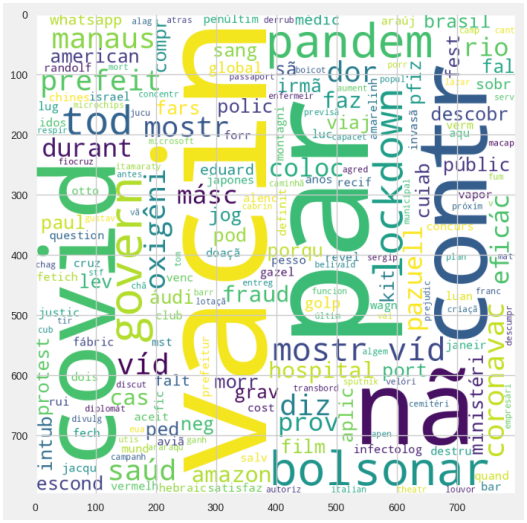
No *cluster* apresentando na figura 10.d, os termos em evidência são: ”sã”, ”paul”, ”lockdown”, ”par”, ”vacin”, ”anos”, ”dor”, ”decret”, ”marc”, ”divulg”. O resultado sugere pertencimento às seguintes categorias: 1) Origens e propagação do coronavırus/da doença COVID-19; 6) Impactos na sociedade e no meio ambiente; 5) Ciência médica: sintomas, diagnóstico e tratamento.

No *cluster* apresentado na figura 10.e, os termos em evidência são: ”covid”, ”vacin”, ”contr”, ”par”, ”cur”, ”ivermectin”, ”tom”, ”morr”, ”hidroxicloroquin”, ”nã”, ”comprov”, ”másc”.

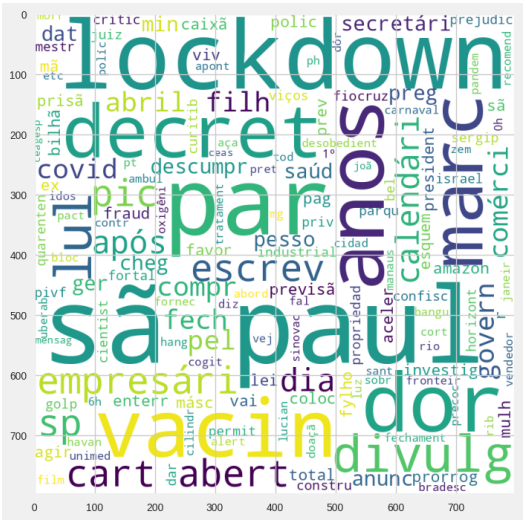


(a) Termos mais frequentes do *cluster* associado com *label* n = 0

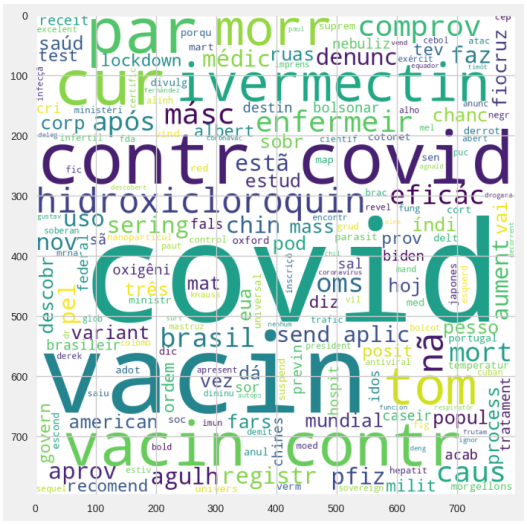
(b) Termos mais frequentes do *cluster* associado com *label* n = 1



(c) Termos mais frequentes do *cluster* associado com *label* n = 2



(d) Termos mais frequentes do *cluster* associado com *label* n = 3



(e) Termos mais frequentes do *cluster* associado com *label* n = 4

Figura 10 – Nuvem de palavras mais frequentes formadas pelos títulos das reportagens falsas para os quatro *clusters* formados no período de janeiro a junho de 2021. Fonte: Autoria Própria (2022).

O resultado sugere pertencimento às seguintes categorias: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veículos de notícias fidedignos.

5.3.4 Julho a dezembro de 2021

Semelhante ao que foi realizado nas subseções anteriores, dada a coluna de PCAs gerados para cada título de notícia no período de julho a dezembro de 2021, e com número de *clusters* definido para $K = 3$, o resultado pode ser visualizado na figura 11.

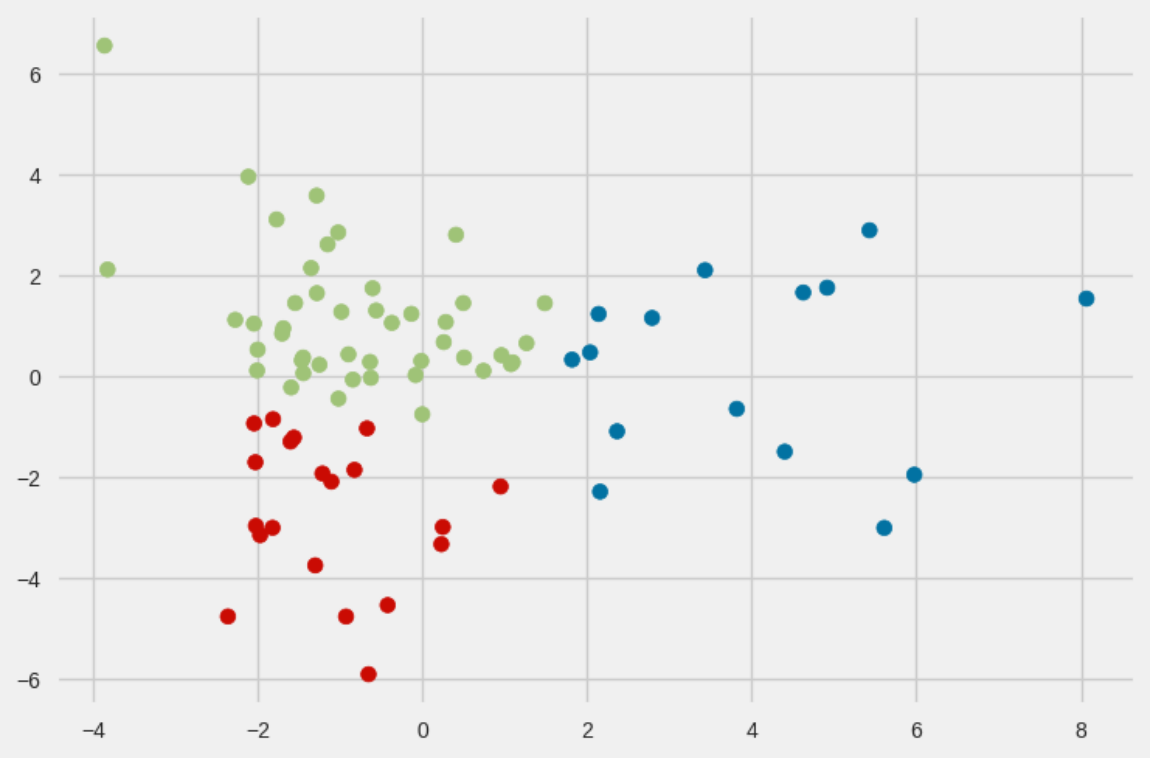


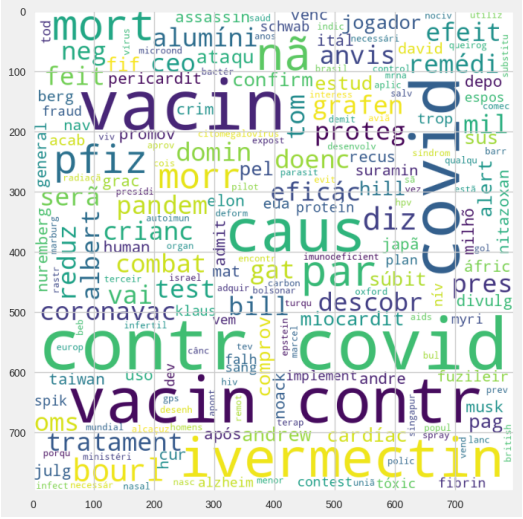
Figura 11 – Método de agrupamento *K-Means* aplicado para $K = 3$ com dados de *Principal Component Analysis* no período de julho a dezembro de 2021. Fonte: Autoria Própria (2022).

Dado o resultado exibido na fig. 11, é possível avaliar que todos os *clusters* possuem algumas similaridades entre si, mas também são evidenciados distanciamentos suficientes para distingui-los.

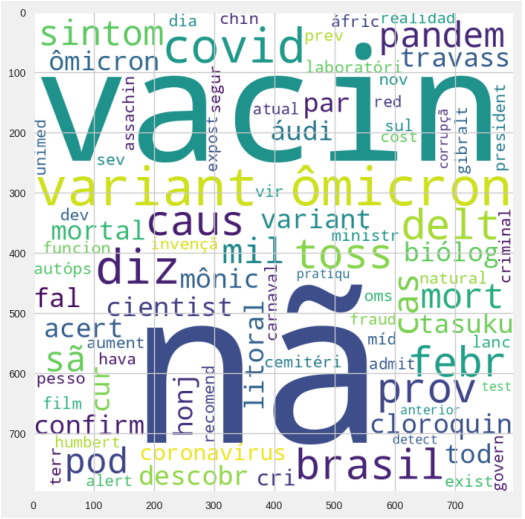
De forma similar ao que foi trabalhado nas subseções anteriores, para melhor avaliar o que pode ser interpretado do resultado obtido no *K-Means* da figura 11, é interessante analisar quais os termos mais frequentes associado a cada *cluster*. Os termos evidenciados são exibidos na figura 12.

No *cluster* da figura 12.a, os termos mais frequentes são: ”vacin”, ”par”, ”sobr”, ”covid”, ”crianc”, ”vai”, ”cert”, ”anos”, ”austrál”, ”cri”, ”forc”, ”telefon”. Os termos sugerem integração às seguintes categorias: 5) Ciência médica: sintomas, diagnóstico e tratamento; 6) Impactos na sociedade e no meio ambiente; 2) Estatísticas falsas e equivocadas.

(a) Termos mais frequentes do *cluster* associado com *label* n = 0



(b) Termos mais frequentes do *cluster* associado com *label* n = 1



(c) Termos mais frequentes do *cluster* associado com *label* n = 2

Figura 12 – Nuvem de palavras mais frequentes formadas pelos títulos das reportagens falsas para os quatro *clusters* formados no período de julho a dezembro de 2021. Fonte: Autoria Própria (2022).

No *cluster* da figura 12.b, os termos mais evidentes são: ”covid”, ”vacin”, ”contr”, ”caus”, ”ivermectin”, ”nã”, ”par”, ”mort”, ”pfiz”, ”morr”. O resultado sugere pertencimento às seguintes categorias de notícias falsas: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; 6) Impactos na sociedade e no meio ambiente.

Por fim, no *cluster* da figura 12.c, os termos mais frequentes são: ”nã”, ”variant”, ”ômicron”, ”vacin”, ”diz”, ”covid”, ”toss”, ”febr”, ”delt”, ”brasil”. O resultado apresentado sugere pertencimento às seguintes categorias: 5) Ciência médica: sintomas, diagnóstico e tratamento; 1) Origens e propagação do coronavırus/da doença COVID-19.

5.3.5 Janeiro a junho de 2022

Semelhante ao que foi realizado nas subseções anteriores, dada a coluna de PCAs gerados para cada título de notícia no período de janeiro a junho de 2022, e com número de *clusters* definido para $K = 4$, o resultado pode ser visualizado na figura 13.

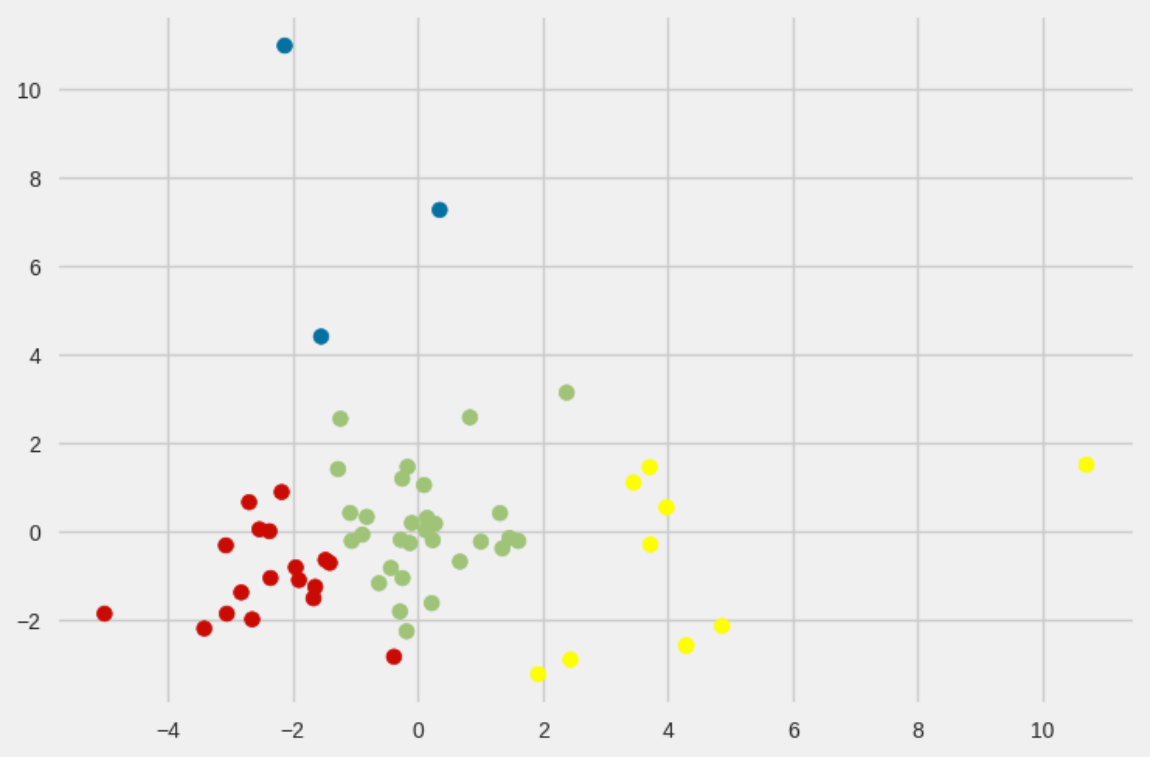
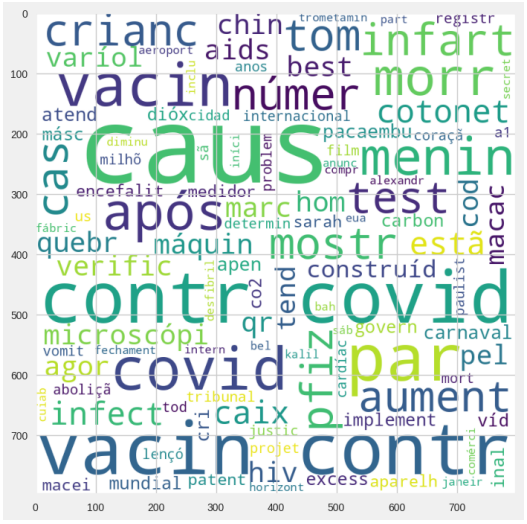


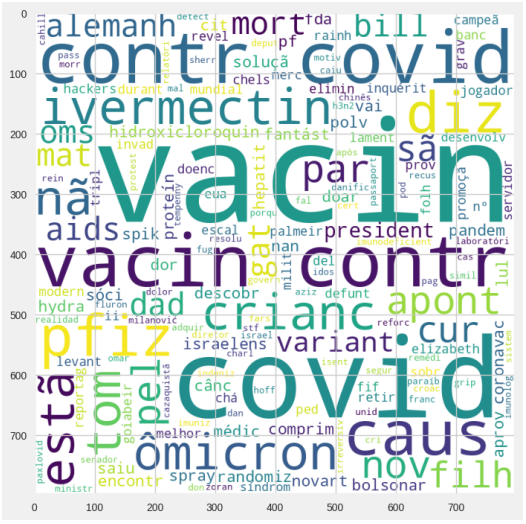
Figura 13 – Método de agrupamento *K-Means* aplicado para $K = 4$ com dados de *Principal Component Analysis* no período de janeiro a junho de 2022. Fonte: Autoria Própria (2022).

Dado o resultado exibido na fig. 13, é possível avaliar que dois *clusters* possuem mais similaridades entre si em relação aos demais, enquanto os outros dois estão mais distantes em relação a todos os demais.

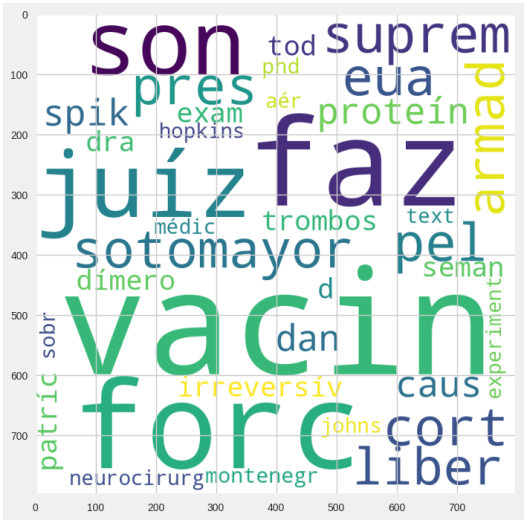
De forma similar ao que foi trabalhado nas subseções anteriores, para melhor avaliar o que pode ser interpretado do resultado obtido no *K-Means* da figura 13, é



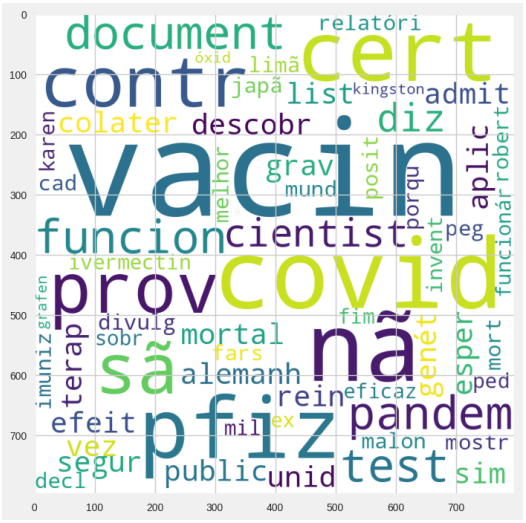
(a) Termos mais frequentes do *cluster* associado com *label* n = 0



(b) Termos mais frequentes do *cluster* associado com *label* n = 1



(c) Termos mais frequentes do *cluster* associado com *label* n = 2



(d) Termos mais frequentes do *cluster* associado com *label* n = 3

Figura 14 – Nuvem de palavras mais frequentes formadas pelos títulos das reportagens falsas para os quatro *clusters* formados no período de janeiro a junho de 2022. Fonte: Autoria Própria (2022).

interessante analisar quais os termos mais frequentes associado a cada *cluster*. Os termos evidenciados são exibidos na figura 14.

No *cluster* da figura 14.a, os termos mais evidenciados são: ”vacin”, ”covid”, ”caus”, ”contr”, ”par”, ”após”, ”tom”, ”vcas”, ”menin”, ”morr”, ”crianc”, ”test”. O resultado sugere pertencimento às seguintes categorias de notícias falsas relacionadas à pandemia: 5) Ciência médica: sintomas, diagnóstico e tratamento; 6) Impactos na sociedade e no meio ambiente; 2) Estatısticas falsas e equivocadas.

No *cluster* da figura 14.b, os termos mais evidencias são: ”vacin”, ”covid”, ”contr”, ”caus”, ”pfiz”, ”diz”, ”ivermectin”, ”apont”, ”par”, ”pel”, ”nov”, ”crianc”. O resultado sugere pertencimento às seguintes categorias: 5) Ciência médica: sintomas, diagnóstico e trata-

mento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; 2) Estatısticas falsas e equivocadas.

No *cluster* da figura 14.c, os termos mais evidentes são: ”vacin”, ”forc”, ”faz”, ”juiz”, ”son”, ”sotomayor”, ”suprem”, ”cort”, ”eua”, ”lib”, ”spik”, ”caus”, ”dan”. O resultado sugere pertencimento às seguintes categorias: 6) Impactos na sociedade e no meio ambiente; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; 5) Ciência médica: sintomas, diagnóstico e tratamento.

Por fim, no *cluster* da figura 14.d, os termos mais evidentes são: ”vacin”, ”covid”, ”nã”, ”pfiz”, ”cert”, ”contr”, ”prov”, ”sã”, ”document”, ”funcion”, ”test”, ”pandem”. Os resultados sugerem inclusão nas seguintes categorias: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; 2) Estatısticas falsas e equivocadas.

5.3.6 Julho a dezembro de 2022

Semelhante ao que foi realizado nas subseções anteriores, dada a coluna de PCAs gerados para cada título de notícia no período de julho a dezembro de 2022, e com número de *clusters* definido para $K = 3$, o resultado pode ser visualizado na figura 15.

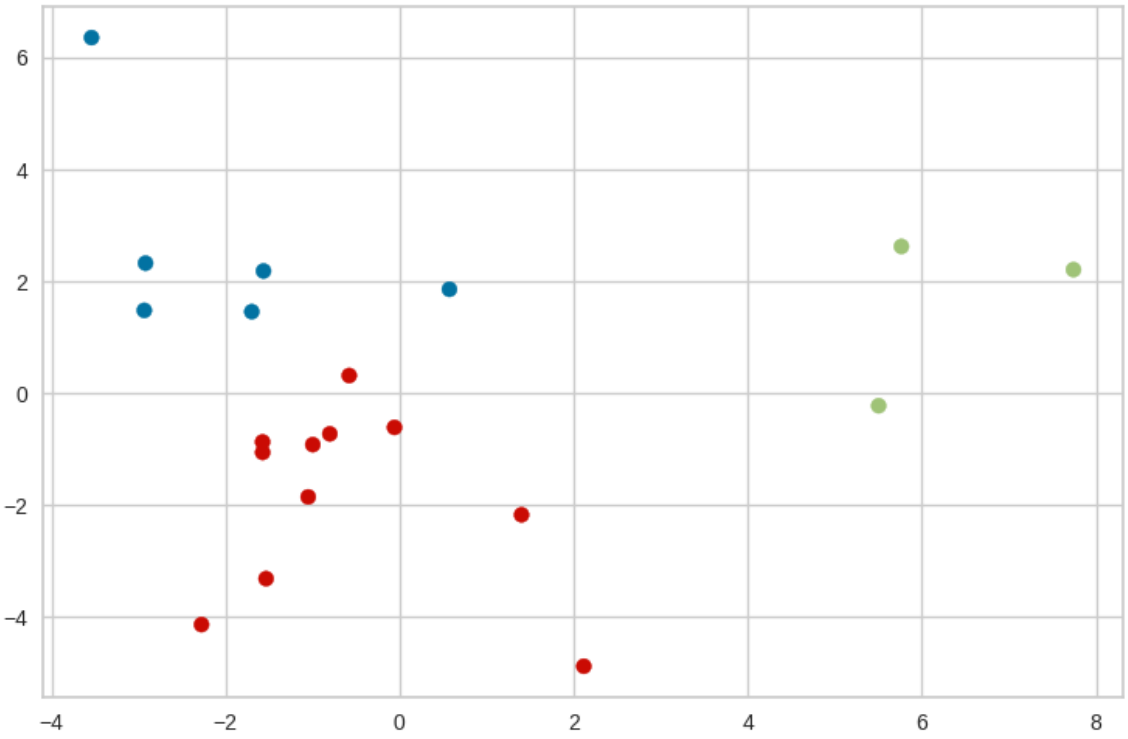
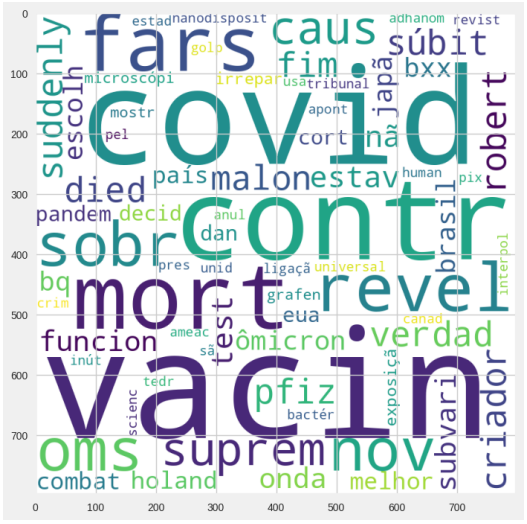
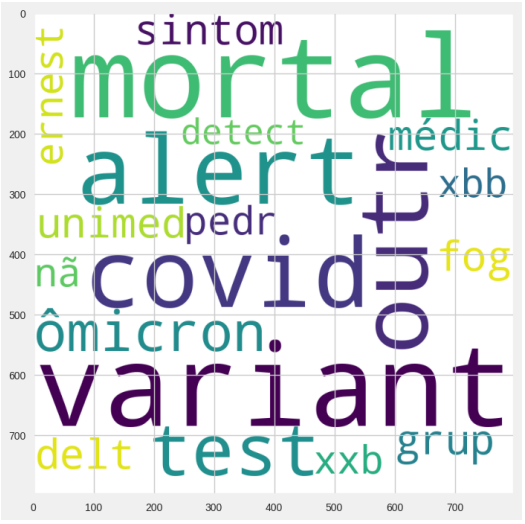


Figura 15 – Método de agrupamento *K-Means* aplicado para $K = 3$ com dados de *Principal Component Analysis* no período de julho a dezembro de 2022. Fonte: Autoria Própria (2022).

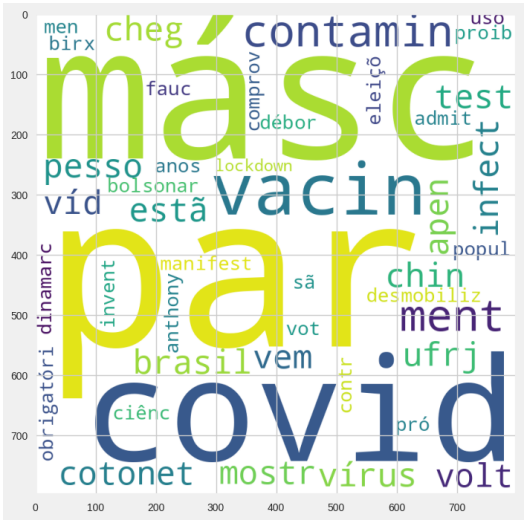
Dado o resultado exibido na fig. 15, é possível observar três *clusters* com significativas diferenças entre si, havendo um ainda mais distante dos outros dois.



(a) Termos mais frequentes do *cluster* associado com *label* n = 0



(b) Termos mais frequentes do *cluster* associado com *label* n = 1



(c) Termos mais frequentes do *cluster* associado com *label* n = 2

Figura 16 – Nuvem de palavras mais frequentes formadas pelos títulos das reportagens falsas para os quatro *clusters* formados no período de julho a dezembro de 2022. Fonte: Autoria Própria (2022).

De forma similar ao que foi trabalhado nas subseções anteriores, para melhor avaliar o que pode ser interpretado do resultado obtido no *K-Means* da figura 15, é interessante analisar quais os termos mais frequentes associado a cada *cluster*. Os termos evidenciados são exibidos na figura 16.

No *cluster* da figura 16.a, os termos mais evidentes são: ”vacin”, ”covid”, ”contr”, ”mort”, ”fars”, ”revel”, ”sobr”, ”nov”, ”oms”, ”suprem”, ”caus”. O resultado sugere pertencimento às seguintes categorias: 4) Desacreditar jornalistas e veículos de notícias fidedignos; 5) Ciência médica: sintomas, diagnóstico e tratamento; 7) Politização.

No *cluster* da figura 16.b, os termos mais evidentes são: ”variant”, ”mortal”, ”alert”, ”covid”, ”outr”, ”test”, ”ômicron”, ”grup”, ”unimed”, ”delt”, ”sintom”. O resultado sugere

pertencimento às categorias: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veículos de notícias fidedignos; 8) Conteúdo impulsionado para ganho financeiro fraudulento.

Por fim, no *cluster* da figura 16.c, os termos mais evidentes são: ”par”, ”másc”, ”covid”, ”vacin”, ”contamin”, ”ment”, ”está”, ”cheg”, ”chin”, ”pesso”, ”brasil”. O resultado sugere pertencimento às seguintes categorias: 1) Origens e propagação do coronavírus/da doença COVID-19; 5) Ciência médica: sintomas, diagnóstico e tratamento; 6) Impactos na sociedade e no meio ambiente.

5.4 Análise de categorização de *fake news*

Nesta seção, serão exibidas as frequências de cada categoria por período e por ano.

5.4.1 2020

No ano de 2020, a frequência das categorias de *fake news* presentes no período de janeiro a junho são exibidos na figura 17.

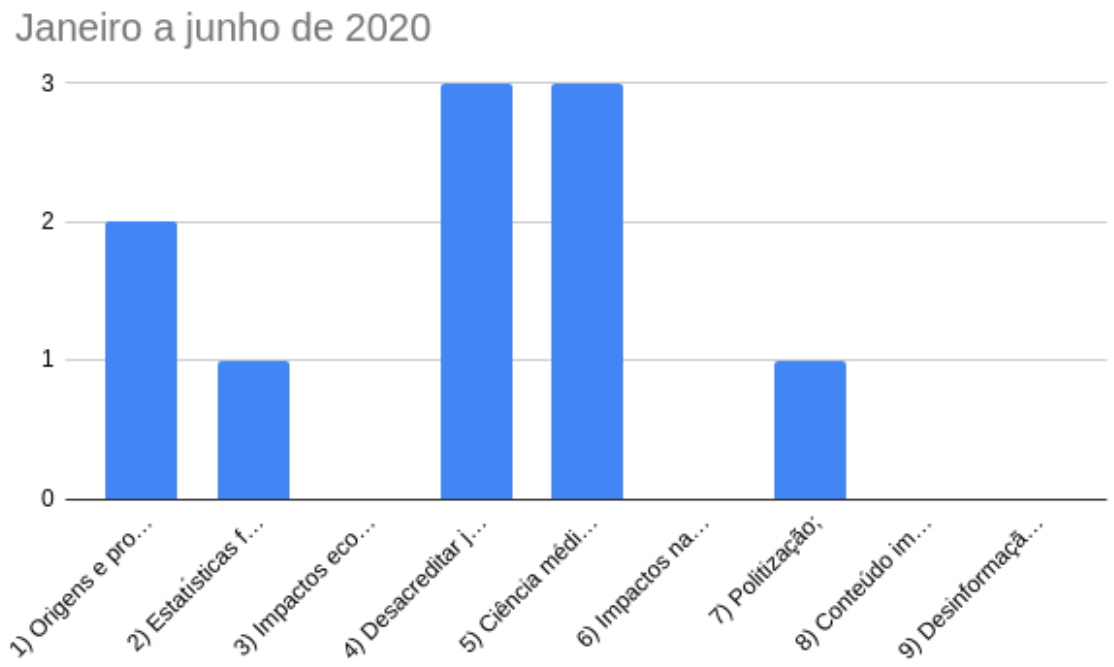


Figura 17 – Frequência de categorias de *fake news* no período de janeiro a junho de 2020. Fonte: Autoria Própria (2022).

Na figura 17, evidencia-se que as categorias de *fake news* mais presentes no primeiro semestre do primeiro ano da pandemia no Brasil foram: 4) Desacreditar jornalistas e veículos de notícias fidedignos; 5) Ciência médica: sintomas, diagnóstico e tratamento; e 1) Origens e propagação do coronavírus/da doença COVID-19.

Na figura 18, a frequência das categorias no período de julho a dezembro de 2020 é exibida.

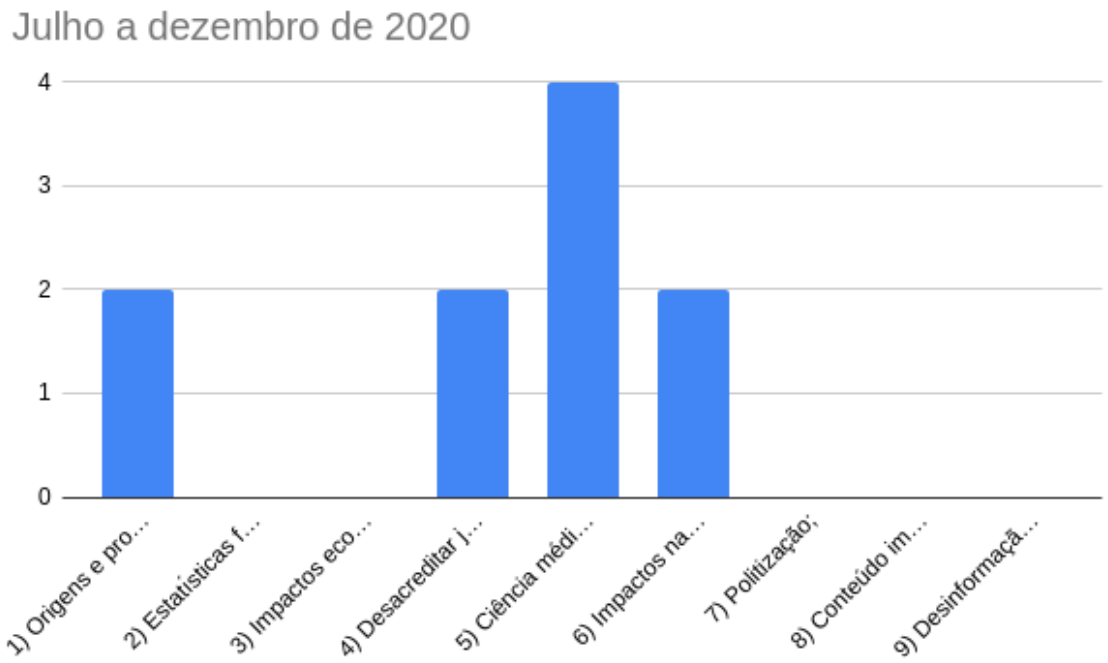


Figura 18 – Frequência de categorias de *fake news* no período de julho a dezembro de 2020. Fonte: Autoria Própria (2022).

Através do gráfico exibido na figura 18, evidencia-se que no segundo semestre do primeiro ano da pandemia no Brasil, a categoria de *fakw news* mais presente foi a 5) Ciência médica: sintomas, diagnóstico e tratamento.

Para uma análise completa do ano de 2020, o gráfico das categorias mais frequentes no decorrer de todo o ano, é exibido na figura 19.

Portanto, confere-se que, no decorrer de todo o primeiro ano da pandemia da COVID-19 no Brasil, as três categorias mais frequentes de *fake news* geradas foram: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; e 1) Origens e propagação do coronavırus/da doença COVID-19.

5.4.2 2021

No ano de 2021, durante o segundo ano de pandemia no Brasil, a frequência das categorias de *fake news* presentes no período de janeiro a junho são exibidos na figura 20.

Na figura 20, evidencia-se que as categorias de *fake news* mais presentes no primeiro semestre do segundo ano de pandemia no Brasil foram: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; 1) Origens e propagação do coronavırus/da doença COVID-19; e 6) Impactos na sociedade e no meio ambiente.

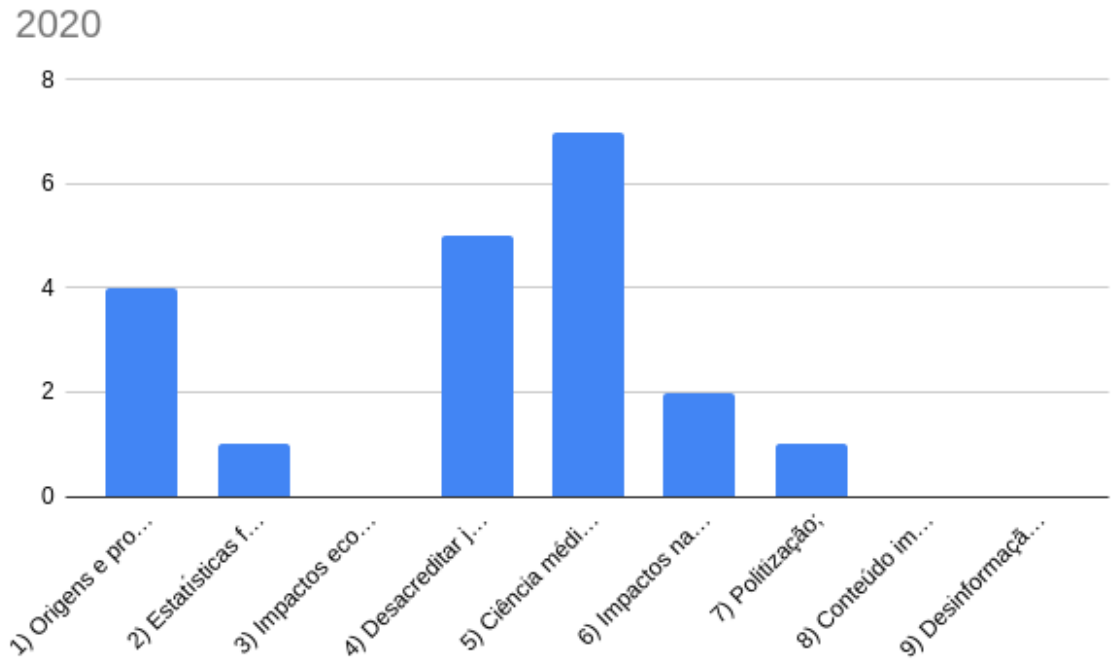


Figura 19 – Frequência de categorias de *fake news* no ano de 2020. Fonte: Autoria Própria (2022).

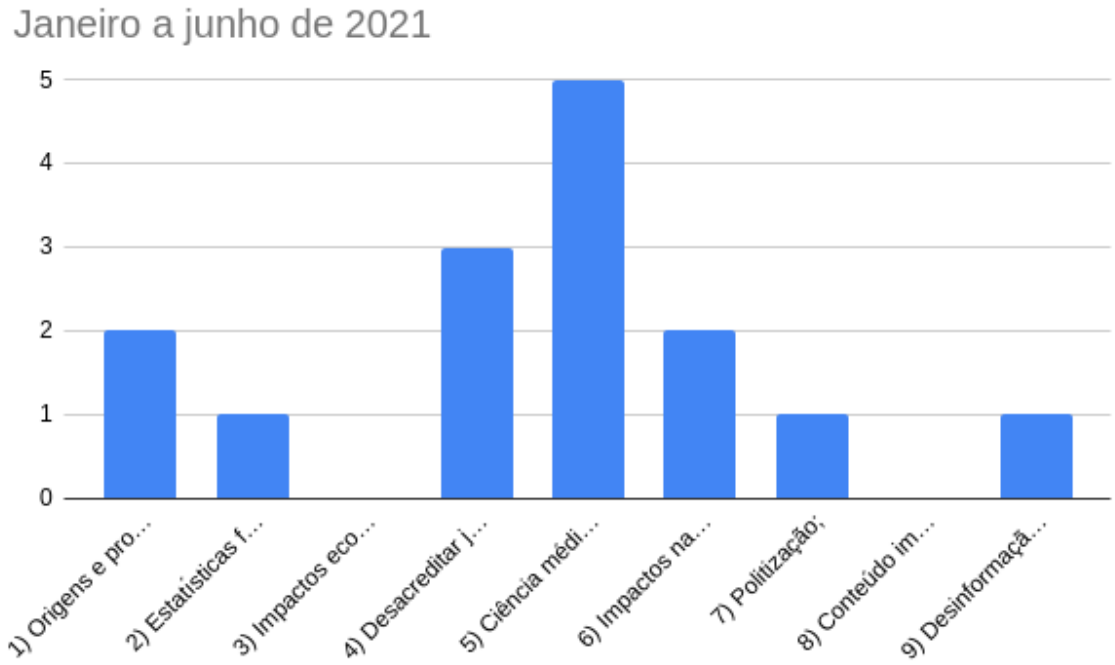


Figura 20 – Frequência de categorias de *fake news* no período de janeiro a junho de 2021. Fonte: Autoria Própria (2022).

Na figura 21, a frequência das categorias no período de julho a dezembro de 2021 é apresentada.

Através do gráfico exibido na figura 21, evidencia-se que, no segundo semestre do segundo ano de pandemia no Brasil, as categorias de *fake news* mais presentes foram: 5)

Figura 21 – Frequência de categorias de *fake news* no período de julho a dezembro de 2021. Fonte: Autoria Própria (2022).

Ciência médica: sintomas, diagnóstico e tratamento; e 6) Impactos na sociedade e no meio ambiente.

Semelhante à subseção anterior, para uma análise completa de 2021, o segundo ano de pandemia no Brasil, o gráfico das categorias mais frequentes no decorrer de todo o ano, é exibido na figura 22.

Portanto, confere-se que, no decorrer de todo o segundo ano da pandemia da COVID-19 no Brasil, as três categorias mais frequentes de *fake news* geradas foram: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; e 6) Impactos na sociedade e no meio ambiente.

5.4.3 2022

No ano de 2022, durante o terceiro ano de pandemia no Brasil, a frequência das categorias de *fake news* presentes no período de janeiro a junho são exibidos na figura 23.

Na figura 23, evidencia-se que as categorias de *fake news* mais presentes no primeiro semestre do terceiro ano de pandemia no Brasil foram: 5) Ciência médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; e 1) Origens e propagação do coronavırus/da doença COVID-19.

Na figura 24, as categorias mais frequentes no período de julho a dezembro de 2022 são apresentadas.

Através do gráfico exibido na figura 24, mostra-se que, no segundo semestre do terceiro ano de pandemia no Brasil, as categorias de *fake news* mais presentes foram: 5)

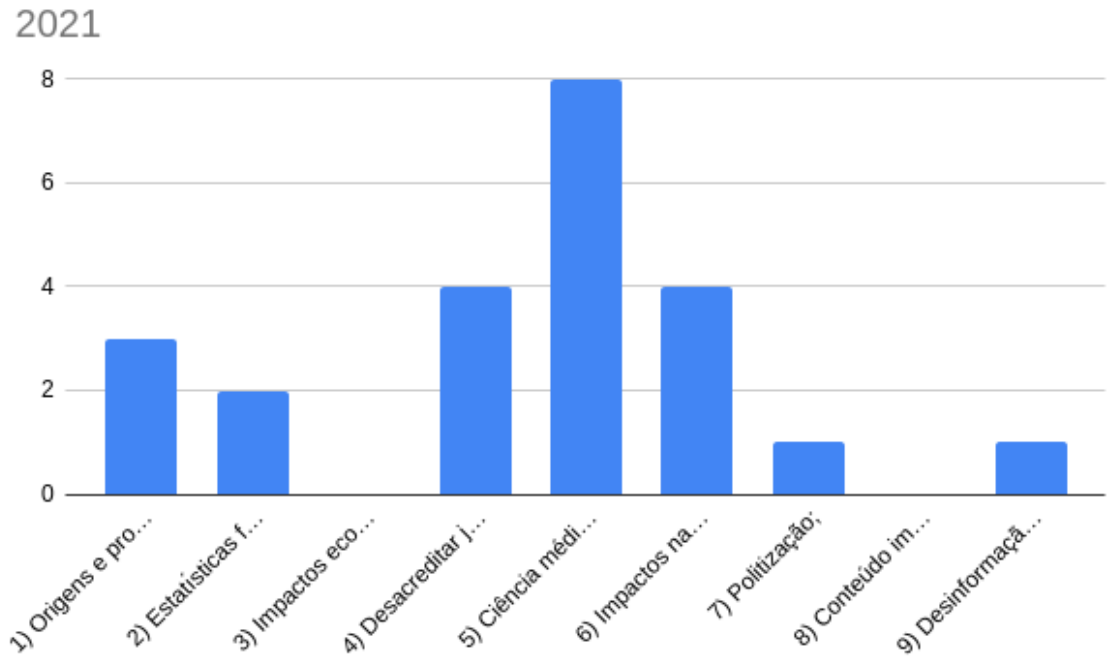


Figura 22 – Frequência de categorias de *fake news* no ano de 2021. Fonte: Autoria Própria (2022).

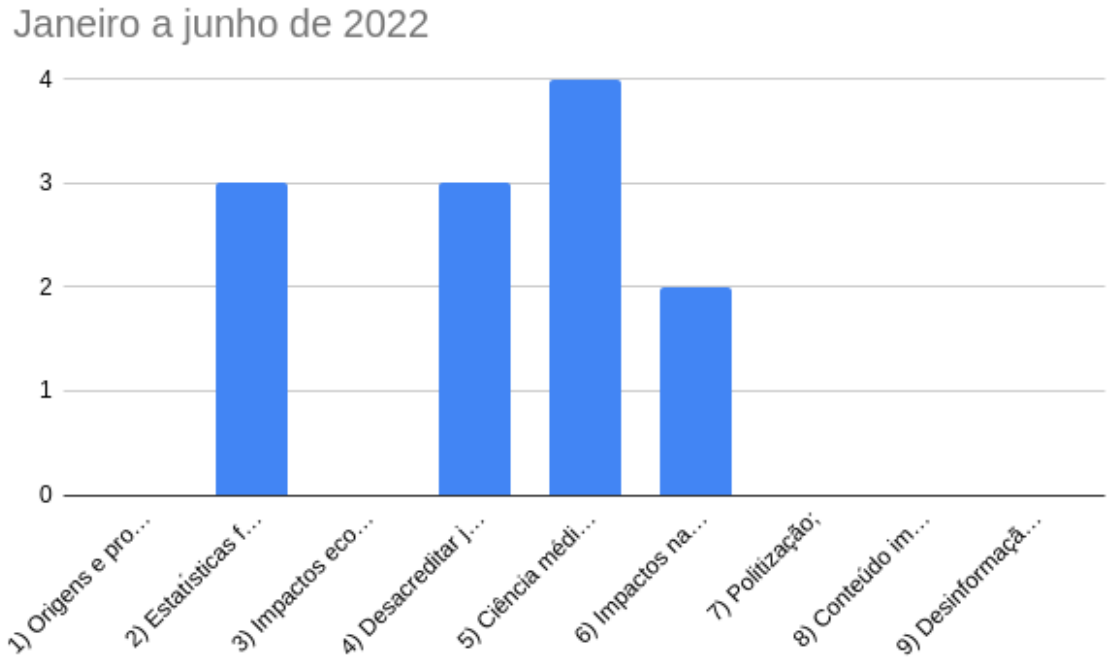


Figura 23 – Frequência de categorias de *fake news* no período de janeiro a junho de 2022. Fonte: Autoria Própria (2022).

Ciência médica: sintomas, diagnóstico e tratamento; e 4) Desacreditar jornalistas e veículos de notícias fidedignos.

Semelhante às subseções anteriores, para uma análise completa do ano de 2022, o terceiro ano de pandemia no Brasil, o gráfico da frequência das categorias no decorrer de

Figura 24 – Frequência de categorias de *fake news* no período de julho a dezembro de 2022. Fonte: Autoria Própria (2022).

todo o ano é exibido na figura 25.

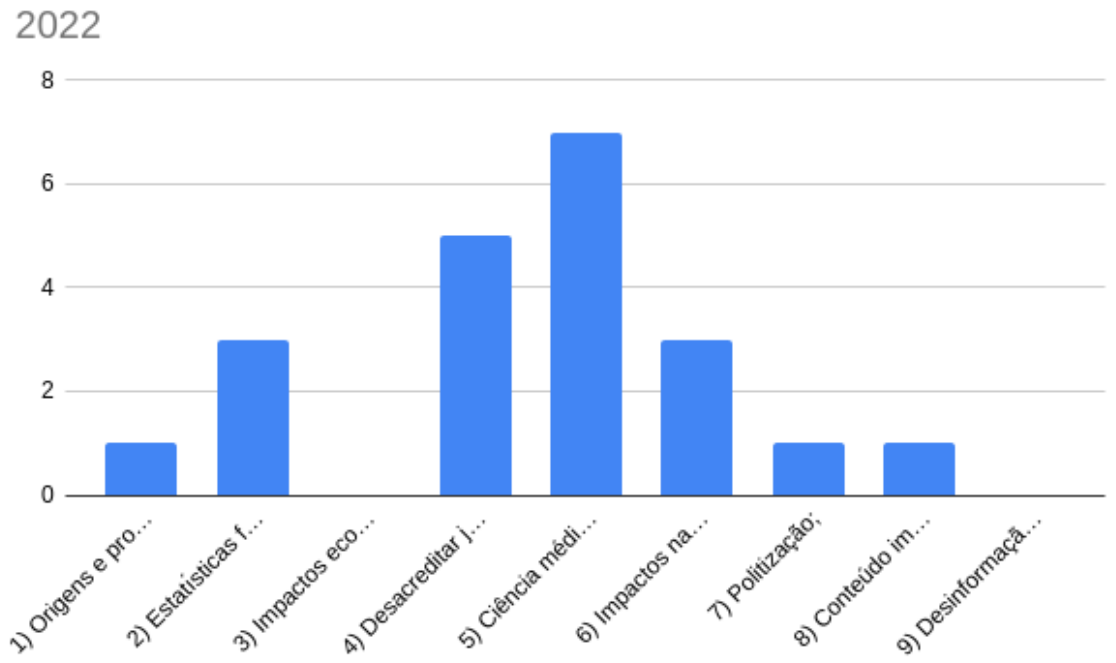


Figura 25 – Frequência de categorias de *fake news* no ano de 2022. Fonte: Autoria Própria (2022).

Portanto, percebe-se que, no decorrer de todo o terceiro ano de pandemia da COVID-19 no Brasil, as categorias mais frequentes de *fake news* geradas foram: 5) Ciência

médica: sintomas, diagnóstico e tratamento; 4) Desacreditar jornalistas e veıculos de notıcias fidedignos; 2) Estatısticas falsas e equivocadas; e 6) Impactos na sociedade e no meio ambiente.

6 CONCLUSÃO

Este trabalho propôs-se a usar um *dataset* com informações falsas sobre a COVID-19 no Brasil e associá-las ao período nos quais foram publicadas, a fim de buscar padrões para categorizações das notícias falsas espalhadas.

Para fazer tais associações, foi necessária uma etapa de tratamento de dados textuais, a fim de buscar palavras e expressões parecidas e categorizá-las como um mesmo assunto a ser abordado na *clusterização*. Além da coleta de documentos, foi feito um pré-processamento que removeu palavras vazias, realizou normalização morfológica, filtragem e ponderação dos termos que apareciam com frequência e análise de componente principal.

Após tal etapa, com os resultados de PCAs obtidos, foi feito o agrupamento das informações coletadas com a técnica de *K-Means*. Para a definição de K, foi utilizado método *Elbow* para cada período.

Por fim, foi realizado um último levantamento dos termos mais frequentes de cada *cluster*. Para cada *cluster*, foram associados os termos frequentes com categorias de *fake news* relacionadas ao COVID-19.

Através dos resultados obtidos neste trabalho, foi possível ver a frequência de cada categoria baseada na realidade de cada semestre e de cada ano de pandemia da COVID-19 no Brasil, a perceber que a categoria de *fake news* mais frequente em todos os anos de pandemia foi a categoria 5) Ciência médica: sintomas, diagnóstico e tratamento.

É importante ressaltar algumas limitações presentes neste trabalho: foi utilizada somente uma base de dados e não há nenhuma métrica definida para escolha de categorização de *fake news*, sendo esta última etapa realizada de forma empírica.

Para trabalhos futuros, sugere-se uma rede de treinamento de categorização de notícias falsas relacionadas à COVID-19, a fim de associar as categorias com os *clusters* de forma objetiva e melhorar processo de identificação e associação de termos de títulos com cada categoria.

Referências

ABBAS, O. A. Comparisons between data clustering algorithms. **International Arab Journal of Information Technology (IAJIT)**, v. 5, n. 3, 2008. Citado 2 vezes nas páginas 10 e 11.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, v. 5, n. 2, 2006. Citado na página 7.

CHAVES, T. d. S. S.; BELLEI, N. C. J. Sars-cov-2, o novo coronavírus: uma reflexão sobre a saúde única (one health) e a importância da medicina de viagem na emergência de novos patógenos. **Revista de Medicina**, v. 99, n. 1, p. i–iv, 2020. Citado na página 4.

ENDO, P. T. et al. Illusion of truth: Analysing and classifying covid-19 fake news in brazilian portuguese language. **Big Data and Cognitive Computing**, MDPI, v. 6, n. 2, p. 36, 2022. Citado 3 vezes nas páginas 3, 7 e 8.

FELDMAN, R.; DAGAN, I. Knowledge discovery in textual databases (kdt). In: **KDD**. [S.l.: s.n.], 1995. v. 95, p. 112–117. Citado na página 7.

FELDMAN, R.; SANGER, J. et al. **The text mining handbook: advanced approaches in analyzing unstructured data**. [S.l.]: Cambridge university press, 2007. Citado na página 7.

GALHARDI, C. P. et al. Fato ou fake? uma análise da desinformação frente à pandemia da covid-19 no brasil. **Ciência & Saúde Coletiva**, SciELO Brasil, v. 25, p. 4201–4210, 2020. Citado 3 vezes nas páginas 2, 3 e 5.

GONÇALVES, L. S. M. **Categorizaçao em text mining**. Tese (Doutorado) — Universidade de São Paulo, 2002. Citado na página 7.

HAN, J.; KAMBER, M.; MINING, D. Concepts and techniques. **Morgan Kaufmann**, v. 340, p. 94104–3205, 2006. Citado na página 10.

HOEK, L. V. D. et al. Identification of a new human coronavirus. **Nature medicine**, Nature Publishing Group, v. 10, n. 4, p. 368–373, 2004. Citado na página 4.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. de O. Análise de componentes principais: resumo teórico, aplicação e interpretação. **ES Engineering and Science**, v. 5, n. 1, p. 83–90, 2016. Citado na página 10.

HUR, D. U.; CAMESELLE, J. M. S.; ALZATE, M. Bolsonaro e covid-19: negacionismo, militarismo e neoliberalismo. **Revista Psicologia Política**, Associação Brasileira de Psicologia Política, v. 21, n. 51, p. 550–569, 2021. Citado na página 1.

JÚNIOR, J. H. de S. et al. Da desinformação ao caos: uma análise das fake news frente à pandemia do coronavírus (covid-19) no brasil. **Cadernos de Prospecção**, v. 13, n. 2 COVID-19, p. 331–331, 2020. Citado 3 vezes nas páginas 3, 4 e 5.

LIKAS, A.; VLASSIS, N.; VERBEEK, J. J. The global k-means clustering algorithm. **Pattern recognition**, Elsevier, v. 36, n. 2, p. 451–461, 2003. Citado na página 11.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007. Citado na página 8.

MUNOT, N.; GOVILKAR, S. S. Comparative study of text summarization methods. **International Journal of Computer Applications**, Foundation of Computer Science, v. 102, n. 12, 2014. Citado na página 9.

NAINGGOLAN, R. et al. Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2019. v. 1361, n. 1, p. 012015. Citado 2 vezes nas páginas 12 e 13.

NIELSEN, R. et al. **Navigating the 'infodemic': How people in six countries access and rate news and information about coronavirus**. [S.l.]: Reuters Institute for the Study of Journalism, 2020. Citado na página 1.

PALADE, I.; BALABAN, D. C. An analysis of covid-19-related fake news from romania. a pilot qualitative study. **Journal of Media Research**, v. 13, n. 2, 2020. Citado na página 1.

POSETTI, J.; BONTCHEVA, K. Desinfodemia: descifrando la desinformación sobre el covid-19. Organización de las Naciones Unidas para la Educación, Ciencia y Cultura-UNESCO, 2020. Citado 2 vezes nas páginas 5 e 6.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: NEW JERSEY, USA. **Proceedings of the first instructional conference on machine learning**. [S.l.], 2003. v. 242, n. 1, p. 29–48. Citado na página 9.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information processing & management**, Elsevier, v. 24, n. 5, p. 513–523, 1988. Citado na página 9.

SOCIETY. **Disinformation: how to recognise and tackle Covid-19 myths**. 2020. Último acesso em: 15 de novembro de 2022. Disponível em: <https://www.europarl.europa.eu/news/en/headlines/society/20200326STO75917/disinformation-how-to-recognise-and-tackle-covid-19-myths>. Citado na página 1.

SOUZA, P. F. e A. Pandemia de desinformação: as fake news no contexto da covid-19 no brasil. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, v. 15, n. 1, 2021. ISSN 1981-6278. Disponível em: <https://homologacao-reciis.icict.fiocruz.br/index.php/reciis/article/view/2219>. Citado na página 5.

TEIXEIRA, L. d. A. C. et al. Saúde mental dos estudantes de medicina do brasil durante a pandemia da coronavirus disease 2019. **Jornal Brasileiro de Psiquiatria**, SciELO Brasil, v. 70, p. 21–29, 2021. Citado na página 4.

WIVES, L. K. Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva. **Exame de Qualificação EQ-069, PPGC-UFRGS**, v. 18, 2002. Citado na página 8.