UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE COMPUTAÇÃO

CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LEONARDO EMERSON ANDRÉ ALVES

CARACTERIZAÇÃO, EVOLUÇÃO E IDENTIFICAÇÃO DE PADRÕES EM NOTÍCIAS FALSAS: UMA ABORDAGEM VOLTADA À MODELAGEM DE TÓPICOS

RIO DE JANEIRO

2023

LEONARDO EMERSON ANDRÉ ALVES

CARACTERIZAÇÃO, EVOLUÇÃO E IDENTIFICAÇÃO DE PADRÕES EM NOTÍCIAS FALSAS: UMA ABORDAGEM VOLTADA À MODELAGEM DE TÓPICOS

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Jonice de Oliveira Sampaio

Co-orientador: Sirius Thadeu Ferreira da Silva

RIO DE JANEIRO

2023

CIP - Catalogação na Publicação

A474c

Alves, Leonardo Emerson André
Caracterização, evolução e identificação de padrões em notícias falsas: uma abordagem voltada à modelagem de tópicos / Leonardo Emerson André Alves. -- Rio de Janeiro, 2023.
202 f.

Orientadora: Jonice de Oliveira Sampaio.
Coorientador: Sirius Thadeu Ferreira da Silva.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Computação, Bacharel em Ciência da Computação, 2023.

1. Processamento de linguagem natural. 2. Modelagem de tópicos. 3. Análise textual. I. Sampaio, Jonice de Oliveira , orient. II. Silva, Sirius Thadeu Ferreira da , coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

LEONARDO EMERSON ANDRÉ ALVES

CARACTERIZAÇÃO, EVOLUÇÃO E IDENTIFICAÇÃO DE PADRÕES EM NOTÍCIAS FALSAS: UMA ABORDAGEM VOLTADA À MODELAGEM DE TÓPICOS

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 28 de junho de 2023.

BANCA EXAMINADORA:

Jonice de Oliveira Sampaio
D.Sc. (IC/UFRJ)

Sirius Thadeu Ferreira da Silva
M.Sc. (PPGI/UFRJ)

Diogo Nolasco Ferreira Sousa
M.Sc. (PPGI/UFRJ)

Silas Pereira Lima Filho
D.Sc. (IC/UFRJ)

“Portanto, não vos escrevo porque vos falta o conhecimento da verdade, mas justamente porque a conheceis e, porquanto, nenhuma mentira tem origem na verdade.”

1 João 2:21

RESUMO

As notícias falsas constituem um problema central na sociedade atual. O avanço das tecnologias e mídias digitais tem alavancado esse problema, visto que se caracterizam como meios extremamente rápidos para disseminação de informação. Dessa forma, a disseminação de desinformações pode implicar em diversos problemas para a sociedade, tais como: influenciar processos democráticos, dificultar o contingenciamento de pandemias, ocasionar crises sociais, que podem trazer graves consequências para a população, entre outros. Este estudo tem como intuito a criação de um processo voltado para a caraterização, descrição da evolução e identificação de padrões em notícias com foco no estudo de notícias falsas escritas em português. Nesse sentido, o foco deste trabalho consiste na caracterização das notícias falsas estudadas por meio da análise textual das mesmas a partir da utilização de uma base de dados de notícias coletadas entre 2013 e 2021, com o uso de técnicas de processamento de linguagem natural e modelagem de tópicos. Portanto, este estudo realizou o tratamento e aperfeiçoamento de um *corpus* com uso de técnicas tanto de limpeza de dados, quanto de *Web Scraping*, e posteriormente uma análise das notícias falsas desse *corpus*, com o uso da linguagem de programação Python, e também com o uso de bibliotecas conhecidas para processamento de linguagem natural e modelagem de tópicos, como NLTK, *gensim* e spaCy; e fazendo uso de algoritmos tradicionais para modelagem de tópicos como *Latent Dirichlet Allocation* (LDA) e *Latent Semantic Analysis* (LSA); em conjunto com as bibliotecas para indexação, visualização e análise de dados Pandas, Matplotlib, Seaborn, Numpy; foi possível dessa forma compreender o avanço dos assuntos e padrões de escrita de notícias falsas, criando um dicionário que caracteriza tais notícias.

Palavras-chave: notícias falsas; análise textual; processamento de linguagem natural; *web scraping*; modelagem de tópicos.

ABSTRACT

Fake news is a central problem in today's society. The advancement of technologies and digital media has leveraged this problem, as they are characterized as extremely fast means of disseminating information. In this way, the dissemination of disinformation can lead to several problems for society, such as: influencing democratic processes, making it difficult to contain pandemics, causing social crises, which can have serious consequences for the population, among others. This study aims to create a process aimed at the characterization, description of the evolution and identification of patterns in news with a focus on the study of fake news written in Portuguese. In this sense, the focus of this work is to characterize the fake news studied through textual analysis of the same from the use of a database of news collected between 2013 and 2021, with the use of natural language processing techniques and modeling of topics. Therefore, this study carried out the treatment and improvement of a corpus using both data cleaning and Web Scraping techniques, and subsequently an analysis of the fake news in this corpus, using the Python programming language, and also with the use of popular libraries for natural language processing and topical modeling such as NLTK, gensim, and spaCy; and making use of traditional algorithms for modeling topics such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA); together with libraries for indexing, visualization and data analysis Pandas, Matplotlib, Seaborn, Numpy; In this way, it was possible to understand the advancement of subjects and patterns of writing fake news, creating a dictionary that characterizes such news.

Keywords: fake news; textual analysis; natural language processing; web scraping; topic modeling.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, o aumento da quantidade de usuários nas mídias sociais transformou a interação humana na Internet. Segundo o estudo *Digital 2022: Global Overview Report*[1], publicado pelo site *Datareportal*, o número de usuários ativos se aproximou da marca de 5 bilhões de pessoas em janeiro de 2022. Ainda segundo o estudo, os dados compilados pelo relatório apontam que um usuário típico de internet gasta em média atualmente 7 horas por dia online, sendo o Brasil um dos países onde as pessoas mais passam tempo na internet, com uma média de 10 horas e 19 minutos por dia. Este aumento registrado no número de usuários, assim como o grande intervalo de tempo que os usuários gastam na internet acaba permitindo a rápida disseminação de informações e atingindo um público amplo e diversificado. Nesse contexto, as informações falsas são frequentemente – e facilmente – propagadas (VOSOUGHI et al., 2018; GUO et al., 2019a; SU et al., 2020), representando um risco não apenas à integridade dos meios de informação, mas também atingindo todos os setores da sociedade, prejudicando a tomada de decisões, e causando prejuízos e danos à sociedade.

As *fake news* de certa forma tanto podem moldar a opinião de cidadãos que não verificam com precisão a fonte de notícias que consomem, quanto podem ser consumidas por indivíduos que consideram informações que condizem com suas crenças já estabelecidas a respeito de determinado assunto, e que podem eventualmente não ser de fato crenças com informações de credibilidade (GELFERT, 2021), e isso ocasiona mudanças no fluxo de acontecimentos sociais, devido ao fato de que esse fenômeno pode trazer mudanças comportamentais nos indivíduos que são afetados por estas informações falsas (BASTICK, 2021). Com isso, a investigação e o combate à desinformação são assuntos muito importantes no atual contexto do fluxo informacional global, tendo em vista que podem alterar o curso de diversos processos sociais, tais como: eleições, administração de crises de saúde, consultas públicas de opinião, entre outros. Os impactos dessas desinformações podem ser vistos tanto em processos democráticos mais corriqueiros como os já citados, quanto em violações de direitos humanos ao redor do mundo (COLOMINA; MARGALEF; YOUNGS, 2021). Porém, até o momento existem poucos trabalhos que destacam

[1] Disponível em: https://datareportal.com/reports/digital-2022-global-overview-report

estudos de *corpus* não-balanceados de notícias falsas com textos em língua portuguesa, investigando o avanço temporal dessas notícias, bem como identificando padrões de palavras e tópicos mais frequentes ao decorrer do texto, e suas influências na construção de uma *fake news*, visando a construção de um dicionário de tópicos referentes as notícias, de forma a caracterizar as notícias de acordo com o seu período temporal de propagação.

1.1 CENÁRIO

Nesta seção será definido o cenário de atuação do presente trabalho, em outros termos, define-se o *corpus* utilizado para o estudo, bem como a estruturação dos dados presentes nesse *corpus*, e também o contexto de procedência das notícias presentes no *corpus* utilizado.

No presente trabalho foi utilizado o *corpus Fakepedia*[2] na versão de 2021, que contém notícias exclusivamente falsas que foram extraídas do site *Boatos*[3]. Esse site, por sua vez, é um site de *fact-checking* feito por jornalistas, criado para compilar notícias falsas.

A estrutura da versão do *corpus* Fakepedia utilizado no presente trabalho é composta por dez colunas que representam as características presentes no *corpus*, ou seja, para cada notícia presente no *corpus*, temos que o *corpus* armazena as características resumidas no Quadro 1.

De maneira geral, as notícias contidas no *corpus* são todas do site Boatos, consequentemente essas notícias advêm tanto de *websites*, quanto de redes sociais, o que exemplifica o contexto das notícias utilizadas no estudo. Portanto, durante o presente estudo foi possível identificar tanto notícias falsas, quanto notícias que visavam desmentir notícias falsas presentes no *corpus*.

[2] Disponível em: https://github.com/andersoncordeiro/Fakepedia-Corpus/blob/main/dataset/fakepedia-corpus-v1.csv
[3] Disponível em: https://www.boatos.org

Quadro 1 – Características presentes no Corpus

Características textuais	Descrição
title	Título original da notícia
title_norm	Título normalizado da notícia
message	Conteúdo da notícia original
message_norm	Conteúdo da notícia normalizado
tokens	Palavras-chave extraídas da notícia
features	Classes gramaticais presentes nas notícias
entities	Pessoas e entidades mencionadas nas notícias
class	Indica se a notícia é falsa ou verdadeira
source	Indica o autor de checagem
url_review	URL da notícia

Fonte: Adaptado pelo autor.

A versão do *corpus* Fakepedia utilizada neste trabalho possui um total de 8.517 observações (notícias), sendo cada notícia descrita com os dados presentes no Quadro 1. A utilização desses dados é o processo-chave utilizado no presente trabalho para extração de informações importantes, por meio das análises realizadas.

1.2 OBJETIVO GERAL

O objetivo geral deste trabalho consiste em realizar uma análise textual de notícias falsas a partir de uma base de dados de notícias coletadas entre 2013 e 2021, com o uso de técnicas de processamento de linguagem natural. Portanto, a análise realizada no presente trabalho foca em analisar temporalmente as notícias falsas abrangendo aspectos, tais como: quantidade de publicações de notícias falsas, categorias (temáticas) que foram alvo de conteúdos de desinformação, entidades nomeadas citadas durante a construção das notícias falsas, bem como uma análise das entidades nomeadas entre as categorias presentes no *corpus*.

1.3 OBJETIVO ESPECÍFICO

De forma mais específica, o objetivo consiste em analisar as notícias do *corpus* Fakepedia com o intuito de verificar as evoluções das notícias falsas mais disseminadas ao longo do tempo, bem como identificar os padrões de escrita das

mesmas, fazendo uso da construção de um dicionário de tópicos e suas respectivas palavras mais frequentes utilizando técnicas de modelagem de tópicos, de forma a caracterizar as notícias do período vigente em estudo e também compreender o avanço desses padrões que, por sua vez, frequentemente se atualizam ao longo do tempo, e também de acordo com o vetor de disseminação (rede social, sites, blogs, etc). Adicionalmente, um dos objetivos específicos do trabalho visa partir da hipótese de que os períodos temporais, as categorias (temáticas) e as entidades nomeadas presentes nas notícias falsas são fatores determinantes na identificação de padrões nos textos dessas notícias. Portanto, ao longo do trabalho essa hipótese será testada.

1.4 COMPOSIÇÃO DOS CAPÍTULOS

O capítulo 1 deste trabalho apresenta uma introdução sobre as *fake news*, realizando toda a contextualização do problema, e trazendo dados e estatísticas extremamente pertinentes a problemática contextualizada ao cenário atual. Com a introdução de muitos conceitos e terminologias fundamentais, o Capítulo 2 tem como principal objetivo realizar a fundamentação teórica do trabalho, realizando a definição e a contextualização de aplicação de conceitos extremamente importantes, tais como: *Fake News*, Processamento de Linguagem Natural, Modelagem de Tópicos e *Web Scraping*. O capítulo 3 é responsável por descrever a revisão de literatura realizada no presente trabalho, trazendo as principais abordagens de combate, detecção e caracterização de *fake news* no cenário atual. Com isso, pode-se dizer que os 3 primeiros capítulos constroem todo o cenário teórico apropriado para o restante do trabalho.

No capítulo 4 são introduzidas as metodologias utilizadas para a realização do trabalho, bem como são descritos os dados utilizados no presente trabalho. Além disso, tanto são descritos os processos de tratamento textual do *corpus*, com o uso de técnicas de Processamento de Linguagem Natural, assim como são descritas todas as atualizações e os aperfeiçoamentos realizados nos dados presentes no *corpus* utilizado para o trabalho.

A análise exploratória dos dados é abordada no capítulo 5, onde é realizada a descrição das notícias falsas presentes no *corpus*, bem como é realizada a caracterização e a investigação de evolução dessas notícias. Nesse capítulo, o principal intuito consiste em analisar as principais informações presentes no *corpus*,

tais como: distribuição de quantidade de notícias, distribuição de categorias de notícias, distribuição de entidades nomeadas e distribuição de entidades nomeadas entre categorias de notícias, visando a caracterização temporal dessas distribuições. Com isto, é possível caracterizar a evolução das notícias tanto de maneira quantitativa, quanto de maneira qualitativa.

No capítulo 6 é realizada a construção do dicionário de tópicos referente as notícias presentes no *corpus*, com o objetivo de realizar a identificação dos principais padrões presentes nas notícias falsas, onde para tal construção são utilizadas técnicas de modelagem de tópicos, que foram anteriormente definidas no capítulo 2. Além disso, nesse capítulo também são realizadas: a descrição, a comparação e a avaliação dos dicionários construídos tanto por meio da análise quantitativa dos dicionários com a utilização da métrica de coerência, quanto por meio de análise qualitativa observando a qualidade e adequação dos tópicos gerados em relação ao contexto das notícias falsas presentes no *corpus*.

Por fim, no capítulo 7 é realizada a discussão final sobre o presente trabalho, bem como são levantadas as principais ideias para futuros trabalhos na área de identificação e caracterização de notícias falsas. Nesse sentido, é explicada a importância e as contribuições do presente trabalho no campo de pesquisa, bem como são descritos possíveis trabalhos futuros que aproveitam os processos e resultados alcançados no presente trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo expõe os principais temas tratados neste trabalho acadêmico. Inicialmente abordam-se os conceitos relacionados com a desinformação e as *fake news*. Após isto, são abordados os conceitos de processamento de linguagem natural, tais como: expressões regulares, normalização de textos, *Term Frequency – Inverse Document Frequency* (TF-IDF), entre outros. Em seguida, a teoria de modelagem de tópicos é apresentada, conceituando os principais métodos e técnicas utilizadas. Finalizando o capítulo, a teoria de *Web Scraping* é introduzida juntamente com as definições das principais técnicas de análise de dados textuais.

2.1 *FAKE NEWS*

As notícias falsas são objetos específicos de análise de um campo mais amplo de pesquisa referente ao estudo da desinformação. Em geral, o estudo da desinformação tem o objetivo de investigar as informações falsas. As informações falsas, por sua vez, podem ser separadas em duas diferentes categorias: *misinformation* e *disinformation*. Nesse caso, é importante trazer uma definição relativa à essas duas categorias de modo a realizar uma diferenciação entre ambas. Formalmente, temos que *misinformation* são “informações falsas ou enganosas, involuntariamente, apresentadas como fato. Quando alguém acredita que algo é preciso, mas na realidade não é.”(Purdue University, tradução nossa)[4], e. Por outro lado, temos que disinformation são “informações falsas ou enganosas distribuídas propositadamente. Quando alguém sabe que algo não é verdade, mas compartilha mesmo assim.” (Purdue University, tradução nossa)[5]. Portanto, é importante destacar que, de acordo com essas definições, temos que a grande diferença entre as duas categorias está na intenção ou no propósito de disseminação, ou seja, as informações falsas pertencentes a categoria de disinformation são caracterizadas pela intenção de disseminar, enquanto as informações falsas pertencentes a categoria de *misinformation* são disseminadas involuntariamente, isto é, sem intenção.

As *fake news* podem ser consideradas como um conceito cuja definição varia bastante na literatura de acordo com diferentes autores e contextos de estudo. Porém,

[4] Disponível em: https://www.lib.purdue.edu/misinformation-training/training-module/what-is-misinformation
[5] *Ibidem.*

é importante destacar definições de autores que se adequem ao contexto do presente trabalho. Desta forma, em (LAZER, 2018, p. 3, tradução nossa) podemos encontrar uma definição mais específica para as *fake news* em termos dessas duas categorias definidas anteriormente (*misinformation* e *disinformation*), isto é, notícias falsas são

> Informações fabricadas que imitam o conteúdo da mídia noticiosa na forma, mas não no processo ou intenção organizacional. Os veículos de notícias falsas, por sua vez, carecem das normas e processos editoriais da mídia noticiosa para garantir a precisão e a credibilidade das informações. Notícias falsas sobrepõe-se a outros distúrbios de informação, como *misinformation* (informações falsas ou enganosas) e *disinformation* (informações falsas que são propositadamente divulgadas para enganar as pessoas).

Além dessa definição, adicionalmente podemos elencar outra definição ainda mais específica referente à contextualização do conceito de notícias falsas em relação ao presente trabalho. Em (MUIGAI, 2019, p. 29, tradução nossa), temos que as *fake news* podem ser definidas como

> Qualquer informação falsa que é deliberadamente destinada a ser total ou amplamente falsa ou enganosa, espalhada pelas mídias sociais on-line, mas ocasionalmente encontrando seu caminho através da mídia tradicional impressa e de transmissão de notícias.

De maneira geral, o objeto de estudo do presente trabalho são notícias e informações falsas que são disseminadas através da rede (inclui mídias tradicionais e redes sociais), visto que o *corpus* Fakepedia engloba notícias e informações falsas checadas pelo site “boatos.org” que advêm tanto de redes sociais, quanto de sites e veículos de informação que disseminam informações falsas, sendo parte destas notícias pertencentes à categoria de *misinformation*, e outra parte pertencente à categoria de *disinformation*. Além disso, ao longo da presente pesquisa constatou-se que as notícias falsas que são objeto de estudo deste trabalho têm sua origem tanto em mídias sociais on-line, quanto em veículos de mídia tradicionais. Portanto, pode-se afirmar que ambas as definições de *fake news* descritas acima têm importância e estão contextualizadas com o presente trabalho.

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL

Nesta seção, será abordado o processamento de linguagem natural. O processamento de linguagem natural foi um meio essencial para a realização de diferentes etapas deste trabalho. Desta forma, se torna essencial realizar a sua

definição para a contextualização de sua aplicação. Para isso, serão descritas as principais definições dos conceitos de processamento de linguagem natural mais utilizados no presente trabalho, tais como: expressões regulares, normalização de textos e *Term Frequency – Inverse Document Frequency* (TF-IDF). Além disso, serão expressas as principais utilidades desses conceitos no trabalho, em outras palavras, as aplicações dos conceitos utilizados. Portanto, podemos definir o processamento de linguagem natural (PLN) como

> Uma área de pesquisa e aplicação que explora como os computadores podem ser usados para entender e manipular texto ou fala em linguagem natural para fazer tarefas. Os pesquisadores do PLN visam reunir conhecimento sobre como os seres humanos entendem e usam a linguagem para que ferramentas e técnicas apropriadas possam ser desenvolvidas para fazer com que os sistemas computacionais entendam e manipulem linguagens naturais para realizar as tarefas desejadas. (CHOWDHURY, 2005, p. 51, tradução nossa).

Portanto, a contextualização da utilização do conceito de processamento de linguagem natural com o presente trabalho é definida como o emprego de técnicas que visem a interação do computador com os textos das notícias falsas estudadas que, por sua vez, estão escritos em linguagem natural, com o intuito de realizar operações, automatizar processos e também para extrair informações desses textos para a realização de um estudo mais aprofundado.

2.2.1 Expressões Regulares

Um dos conceitos mais fundamentais empregados no contexto de realizar operações e automatizar processos no manuseio dos textos das notícias falsas estudadas refere-se ao conceito de expressões regulares que, por sua vez, foram ostensivamente utilizadas no presente trabalho. De forma mais precisa, podemos definir uma expressão regular como “uma *string* de letras, números e símbolos especiais para descrever uma ou mais *strings* de pesquisa. A *string* de pesquisa pode conter informações fixas ou variáveis” (BHATIA, 2005, p. 1, tradução nossa).

Em geral, a principal utilidade das expressões regulares no presente trabalho está associada à necessidade de realizar a limpeza do *corpus* utilizado para o estudo, de modo a realizar a remoção ou substituição de frações de dados não importantes para os fins deste trabalho, tais como: números de telefone, *stopwords* (incluindo sinais de pontuação), entre outras frações de dados não importantes. Para esse fim,

foram construídos padrões de expressões regulares que combinassem com as frações de dados às quais se desejava realizar limpeza ou substituição.

2.2.2 Normalização de Textos

Outro processo empregado no presente trabalho e que, por sua vez, tem estrita ligação com a teoria de processamento de linguagem natural refere-se à normalização de texto, que tem como principal importância a redução de aleatoriedade de padrões no texto, visando aproximar o texto das notícias de um padrão predefinido que, por sua vez, ajuda a reduzir a quantidade de informações despadronizadas com as quais o computador (algoritmos de modelagem de tópicos, etc) precisa lidar, o que consequentemente traz maior eficiência dos algoritmos. Portanto, de maneira mais formal, podemos definir a normalização de texto como o processo de

> Converter as palavras despadronizadas em seus formatos padrão (ou seja, no formato de *string* ou formato de texto simples). A normalização do texto é necessária para que palavras como datas, símbolos de moeda, abreviações, acrônimos e números, etc. sejam pronunciadas ou lidas pelo sistema (HARDE, 2019, p. 1, tradução nossa).

Portanto, deve-se ressaltar que a grande utilidade da normalização de textos está atrelada à necessidade de reduzir a aleatoriedade na construção de um texto, aproximando as informações presentes no texto de um “padrão” pré-definido. Essa ação, em geral, facilita o processo de homogeneização das informações com as quais os métodos de processamento de linguagem natural precisam lidar, dentre os quais podemos citar o processo de *stemming*, e o processo de lematização.

2.2.3 Tf-Idf

Outro conceito importante empregado no presente trabalho refere-se ao *Term Frequency – Inverse Document Frequency* (TF-IDF). Nesse caso, o principal intuito de sua utilização está atrelado ao objetivo de encontrar a importância de cada palavra presente no texto das notícias falsas estudadas. De maneira geral, podemos definir o TF-IDF como

> Uma medida, usada nos campos de recuperação de informação (IR) e aprendizado de máquina, que pode quantificar a importância ou relevância de representações de *strings* (palavras, frases, lemas, etc) em um documento entre uma coleção de documentos (também conhecido como *corpus*) (SIMHA, 2021, Online).

Com essa definição é possível compreender a ideia central dessa medida, faltando apenas sua definição matemática. Porém, antes de definir matematicamente essa medida, é também importante ressaltar que o TF-IDF é definido como o produto de outras duas medidas relevantes. Portanto, é imprescindível definir previamente essas duas outras medidas que, quando usadas em conjunto, resultam no TF-IDF. Essas duas medidas são: o *Term Frequency* (TF) e o *Inverse Document Frequency* (IDF).

Segundo Ganesan (2019), o *Term Frequency* reflete a frequência de ocorrência de um termo em um documento. Portanto, considerando um documento d_i com K termos, e definindo t_i como um termo específico nesse documento, e consequentemente $Q(t_i)$ como sendo a quantidade de vezes que o termo t_i aparece no documento, então calcula-se a frequência de um termo t_i nesse documento d_i, como:

$$TF(t_i, d_i) = \frac{Q(t_i)}{K}$$

Portanto, com relação ao *Term Frequency*, pode-se afirmar que se trata de uma medida que utiliza de informações obtidas localmente em um documento contido no *corpus* de estudo em questão, não sendo necessário obter informações além das contidas no próprio documento de análise.

Por outro lado, segundo Nettleton (2014), temos que o *Inverse Document Frequency* pode ser entendido como uma medida que atribui a um termo o quão comum ou raro é sua ocorrência em um documento específico de um *corpus* que, por sua vez, contém vários documentos. Para a obtenção dessa nova medida, deve-se calcular a razão entre o número total de documentos em um *corpus* e o número de documentos que contém o termo alvo no *corpus*. Matematicamente, considerando um *corpus* D, um termo t_i e um documento $d_j \in D$, temos que a medida *Inverse Document Frequency* para o termo t_i é definida como segue:

$IDF_i = log \frac{|D|}{|\{d: t_i \in d\}|}$, onde:

- $|D|$ é a quantidade de documentos no *corpus.*
- $|\{d: t_i \in d\}|$ é o número de documentos em que o termo t_i aparece.

Logo, no contexto do *Inverse Document Frequency*, pode-se entender que é uma medida calculada por meio de informações globais, isto é, o IDF não se atém a

um único documento, visto que utiliza para o cálculo quantidades relativas ao *corpus*, como um todo.

Consequentemente, após a definição conceitual e matemática dessas duas medidas (TF e IDF), então pode-se definir o conceito de *Term Frequency – Inverse Document Frequency* (TF-IDF) que, por sua vez, pode ser definido como a multiplicação das duas medidas definidas anteriormente, como segue:

$$TF - IDF = TF(t_i, d_i) \times IDF_i$$

Dessa forma, cabe ressaltar que a aplicação do TF-IDF é muito comum em sistemas de busca online, principalmente com o intuito de trazer o melhor resultado para consultas realizadas por usuários nesses sistemas. Além disso, é importante compreender que existem diferentes variantes dessa medida, ou seja, com formas de cálculo do TF e do IDF de diferentes maneiras, e com diferentes combinações.

Portanto, pode-se afirmar que a principal utilidade do TF-IDF neste trabalho foi a possibilidade de atribuir importância às principais palavras presentes nas notícias, e foi utilizado em combinação com uma das abordagens utilizadas para a modelagem de tópicos visando a construção do dicionário de tópicos.

2.2.4 Lematização

A lematização é uma técnica amplamente utilizada no processamento de linguagem natural e tem sido objeto de diversos estudos ao longo dos anos. Um dos trabalhos mais renomados sobre o tema foi apresentado por Martin Porter em 1980 (PORTER, 1980), que desenvolveu a técnica de *stemming*, processo de redução das palavras para suas formas mais simples. No entanto, enquanto o *stemming* é uma abordagem simplificada da lematização, outras técnicas mais avançadas têm sido desenvolvidas.

Uma abordagem tradicional da lematização envolve o uso de dicionários com informações sobre cada palavra, incluindo seu lema e outras informações lexicais relevantes. Trabalhos como o de Lovins em 1968 (LOVINS, 1968) e o de Wilks em 1972 (WILKS, 1972) exploraram essa estratégia em diferentes línguas. No entanto, esses métodos baseados em dicionários são limitados pela qualidade desses recursos.

Com o advento de algoritmos de aprendizado de máquina, tornou-se possível gerar lemas automaticamente. A abordagem baseada em modelos estatísticos é particularmente relevante nesse contexto. Dentre os trabalhos mais importantes baseados em técnicas de aprendizado de máquina para a lematização estão os de Schmid em 1994 (SCHMID, 1994) e o de Mikolov et al. em 2013 (MIKOLOV et al., 2013), que introduziram modelos neurais para resolver o problema.

A lematização é uma técnica importante em várias áreas de aplicação do processamento de linguagem natural, incluindo análise de sentimento, classificação de texto, recuperação da informação e tradução automática. Em geral, a lematização ajuda a melhorar a precisão do processamento de texto, permitindo que as palavras sejam agrupadas com mais eficácia e comparadas com outras palavras em um contexto específico.

Por fim, é importante ressaltar que a utilização da lematização no presente trabalho decorre da necessidade de realizar um pré-processamento nos textos usados no processo de modelagem de tópicos realizado. Desse modo, a lematização é aplicada de maneira a diminuir a redundância dos tópicos, e consequentemente trazendo menos ruído as modelagens realizadas.

2.3 MODELAGEM DE TÓPICOS

Nesta seção será abordado o conceito de modelagem de tópicos, trazendo inicialmente uma definição precisa da metodologia, e posteriormente definindo os dois principais métodos de modelagem de tópicos utilizados atualmente realizando uma diferenciação entre eles. De maneira geral, podemos definir a modelagem de tópicos como “métodos estatísticos que analisam as palavras dos textos originais para descobrir os temas que foram tratados no texto, como esses temas estão conectados uns aos outros e como eles mudam ao longo do tempo” (BLEI, 2012, p. 2, tradução nossa).

Se considerarmos alguns dos métodos mais utilizados para a realização da modelagem de tópicos, podemos citar dois métodos, em específico: o método *Latent Dirichlet Allocation* (LDA), e o método *Latent Semantic Analysis* (LSA). Ambos os métodos são comumente utilizados para esse fim, entretanto existem diferenças conceituais no funcionamento dos mesmos.

O método LDA é um dos métodos mais populares para realizar o processo de modelagem de tópicos. De maneira geral podemos definir o método LDA como

> Um método probabilístico generativo não supervisionado para modelar um *corpus*, é o método de modelagem de tópicos mais comumente usado. O LDA assume que cada documento pode ser representado como uma distribuição probabilística sobre tópicos latentes e que a distribuição de tópicos em todos os documentos compartilha um prior de Dirichlet comum. Cada tópico latente no modelo LDA também é representado como uma distribuição probabilística sobre palavras e as distribuições de palavras de tópicos também compartilham um prior de Dirichlet comum.(JELODAR et al., 2018, p. 5, tradução nossa).

Por outro lado, também temos o método LSA que, de certa forma, busca encontrar variáveis que possam representar um conjunto de palavras com o mesmo significado. Formalmente, segundo Wiemer-Hastings (2004, p. 1, tradução nossa) o método LSA é definido como “uma técnica para criar representações de textos baseadas em vetores que pretendem capturar seu conteúdo semântico”.

A principal diferença entre os dois métodos é o fato de que o LSA tem como saída uma matriz de similaridade de cossenos que, por sua vez, aponta as similaridades de cada documento. Já o método LDA tem como saída uma matriz, onde suas linhas representam todas as palavras no conjunto de dados, e suas colunas representam todos os documentos do *corpus* em estudo. Cada valor na matriz representa um tópico ao qual a palavra da linha e da coluna correspondente faz parte. Além disso, podemos afirmar que o funcionamento dos dois métodos é bem diferente. Segundo (BLEI; NG; JORDAN, 2003), no método LDA cada documento é uma mistura diferente de tópicos, e o LDA permite que cada palavra esteja em um tópico diferente. Esquematizando do ponto de vista matemático o funcionamento do algoritmo LDA aplicado a um *corpus* textual, e especificamente a um documento d contido no *corpus* de estudo, temos o seguinte processo:

1.Para um documento d_j:

2. Faça as distribuições $\Phi_k \sim \mathrm{Dir}(\Phi, \beta)$ para todo tópico k, como $0 \leq k \leq K$.

3. Faça uma distribuição $\theta_j \sim Dir(\theta, \alpha)$ para o documento d_j.

4. Para cada índice de posição i das palavras no documento d_j,

a) Escolha aleatoriamente um tópico $z_{j,i} \sim Multinomial(\theta_j)$.

b) Escolha aleatoriamente uma palavra $w_{j,i}$ com probabilidade $p(w_{j,i}|\Phi_{z_{j,i}})$

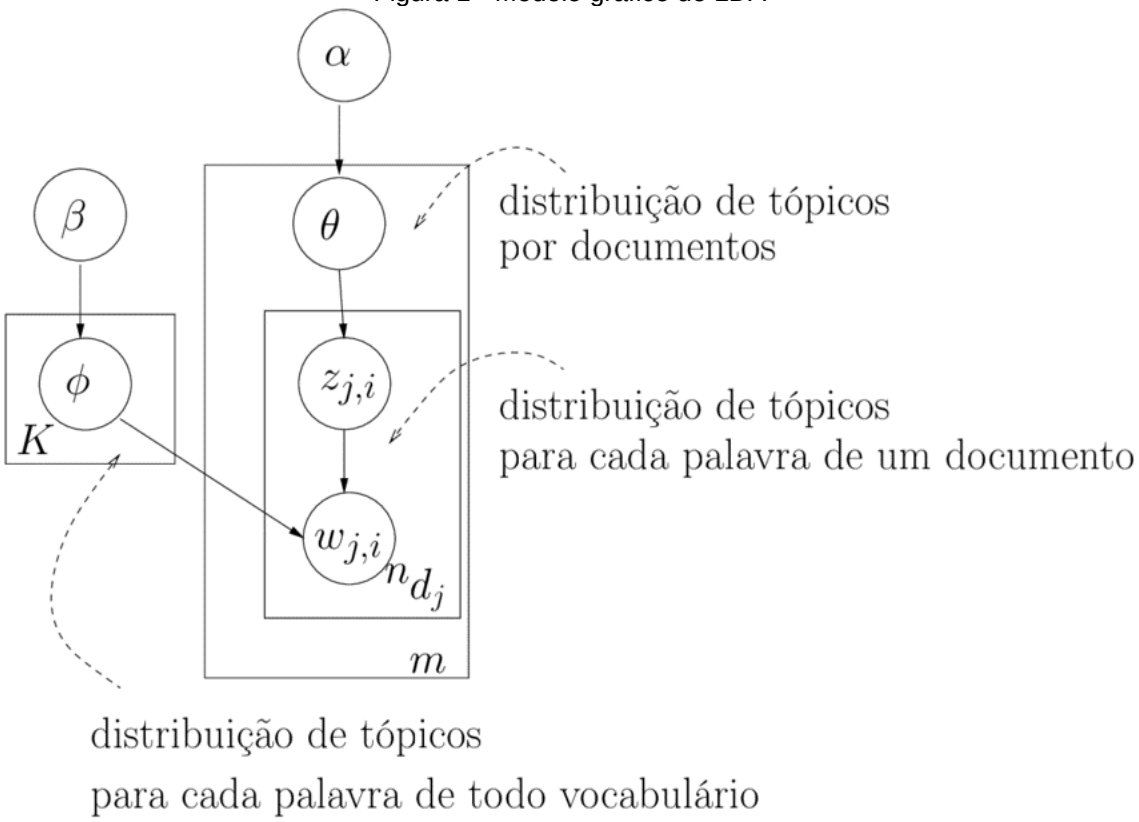
Observando a esquematização do funcionamento é possível perceber que existem três procedimentos principais: **geração dos tópicos**, **alocação de tópicos para documentos** e **geração de palavras**.

Nesse sentido, a distribuição responsável por realizar a amostragem da distribuição de tópicos é chamada de distribuição de Dirichlet. Como o processo é generativo, a amostragem da distribuição Dirichlet resultante é utilizada com o intuito de alocar as palavras em diferentes tópicos de maneira a abranger os documentos fornecidos ao algoritmo.

No processo de geração, geralmente são utilizadas duas variáveis para as distribuições que são geradas pela distribuição de Dirichlet. A variável Φ é n-dimensional visto que considera o número de palavras do vocabulário. Adicionalmente, a variável θ é K-dimensional, visto que leva em conta o número de tópicos. Cada uma dessas variáveis possuem hiperparâmetros α e β.

Após a obtenção de Φ e θ_j, gera-se um documento d_j com n_{d_j} termos representado por meio de uma *bag-of-words*. Para cada índice de posição i da *bag-of-words* é escolhida uma palavra obtida da distribuição de tópicos, isto é, escolhe-se um tópico k dos K tópicos existentes, e posteriormente é feita a correlação desse tópico ao índice de posição i escolhido do documento d_j. O tópico escolhido de acordo com a distribuição θ_j é representado pela variável $z_{j,i}$. Consequentemente, escolhe-se a palavra que irá ocupar o índice de posição de acordo com a distribuição Φ que, por sua vez, representa K distribuições de dimensão n, onde cada distribuição Φ_k, representa a proporção de palavras que descrevem semanticamente o assunto tratado no tópico k. Portanto, para definir $w_{j,i}$ é feita uma escolha no tópico $z_{j,i}$ de acordo com a distribuição de palavras $\Phi_{z_{j,i}}$. Sintetizando graficamente, temos na figura a representação do processo detalhado.

Figura 1 - Modelo gráfico do LDA



Fonte: (FALEIROS, 2016; BLEI; NG; JORDAN, 2003).

Por outro lado, segundo Stephanie Glen, no método LSA temos que as palavras são ligadas a conceitos, e após isto ambos são então comparados para chegar ao real significado do texto. O artigo[6] feito por ela define o método da seguinte forma:

1. O texto é convertido em matrizes de frequências para representar passagens. Cada célula na matriz contém o número de vezes que uma determinada palavra aparece em uma determinada passagem.

2. A matriz é fatorada usando o método *Singular Value Decomposition* (SVD) para que cada passagem seja representada como um vetor. O valor para cada vetor é a soma dos vetores que representam suas palavras componentes.

3. Após isso, temos que: produtos escalares, cossenos ou métricas semelhantes são usadas para representar semelhanças entre palavras e passagens.

Considerando a abordagem mais tradicional do LSA, temos que o mesmo utiliza de um modelo de espaço vetorial para representar os documentos de um *corpus* como vetores em um espaço vetorial onde termos de um dicionário são usados como

[6] Disponível em: https://www.statisticshowto.com/latent-semantic-analysis/

dimensões. Com isso, temos que uma coleção de documentos d em um espaço de t termos de dicionário são representados por uma matriz de frequência denotada por M que, por sua vez, é submetida ao processo de decomposição em valores singulares (SVD) onde é realizada a decomposição da matriz M em: autovetores de termos U, autovetores de documentos V e valores singulares Σ da seguinte forma:

$$M = U\Sigma V^T$$

A decomposição SVD representa M com a utilização de um espaço de dimensões semânticas latentes. Para explicar a variabilidade nas ocorrências de termos em documentos é importante observar os valores dos r elementos da matriz diagonal Σ, chamados de valores singulares, que são calculados como a raíz quadrada dos autovalores comuns tanto na análise de componentes principais considerando termos como variáveis (com documentos como observações), e também na análise de componentes principais considerando documentos como variáveis (com termos como observações). Portanto, observando as k dimensões mais importantes (ou seja, associadas com os k valores singulares mais altos) e descartando os restantes $r-k$ é possível construir uma versão truncada da matriz termo-frequência M_k, onde M_k é a melhor aproximação por método de mínimos quadrados da matriz original M de tal forma que seja possível minimizar o valor da norma de Frobenius de $Z = M - M_k$ que, por sua vez, é calculada como:

$$\|Z\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|z_{ij}\right|^2}$$

Portanto, M_k é a matriz cujo valor de $\|Z\|_F$ seja mínimo e transforma a matriz original M levando em conta a estrutura oculta dos tópicos onde os termos e documentos são construídos.

Neste trabalho, a principal utilidade da modelagem de tópicos é seu uso na construção de um dicionário de tópicos visando a caracterização das palavras mais importantes empregadas na construção das notícias falsas estudadas neste trabalho. Com isso, torna-se viável a identificação das palavras mais frequentes empregadas na construção de notícias falsas em língua portuguesa no período relativo às notícias estudadas, bem como compreender o padrão de escrita das mesmas.

2.4 *WEB SCRAPING*

Nesta seção será abordada a técnica de *Web Scraping* (raspagem da *Web*), também conhecida como extração ou coleta da *Web*, com o objetivo de trazer uma definição formal sobre essa técnica e descrever a principal utilidade da técnica no presente trabalho. De maneira formal, temos que a técnica *Web Scraping* pode ser definida como

> Uma técnica para extrair dados da *World Wide Web* (WWW) e salvá-los em um sistema de arquivos ou banco de dados para recuperação ou análise posterior. Normalmente, os dados da web são raspados utilizando o protocolo de transferência de hipertexto (HTTP) ou por meio de um navegador da web. Isso é feito manualmente pelo usuário ou automaticamente por um *bot* ou rastreador da web (ZHAO, 2017, p. 1, tradução nossa).

A principal utilidade da técnica de *Web Scraping* no presente trabalho está relacionada com a identificação de uma necessidade de obter dados sobre as notícias falsas do *corpus* Fakepedia que não foram agregadas em sua construção inicial, para a realização de análises mais detalhadas. Com isso, neste trabalho foi aplicada a técnica de *Web Scraping* por meio da construção de um *bot* para extrair dados das notícias, tais como: data de publicação das notícias e categorias das notícias. A aplicação dessa técnica ocorreu de forma a extrair esses dados do site fonte das notícias do *corpus* Fakepedia, e para isso foi construído um código (*bot*) responsável por realizar as requisições e extrair as informações das *tags* que constituem o código em Linguagem de Marcação de Hipertexto (HTML) do site. Mais detalhes sobre esse processo serão abordados no Capítulo 4, mais especificamente na seção 4.4 do presente trabalho.

3 REVISÃO DE LITERATURA

Este capítulo tem por finalidade apresentar e discutir o estado da arte em relação a técnicas de detecção automática de notícias falsas, bem como estudos aplicados à resolução do problema de disseminação das *fake news*, sendo esses estudos tanto focados em caracterização, identificação de padrões ou mitigação de efeitos de notícias falsas. Neste sentido, foram estudados tanto trabalhos no sentido da detecção automática, quanto trabalhos voltados à construção de *corpus* de notícias falsas, assim como trabalhos focados na identificação de padrões e na evolução de notícias falsas. Por fim, será realizada uma análise comparativa entre os trabalhos apresentados e apontados os déficits onde o presente trabalho busca realizar suas contribuições. Nesse caso, é importante informar que o objetivo desta revisão de literatura foi a de realizar uma revisão de literatura do tipo narrativa, não seguindo critérios sistemáticos para a busca e seleção dos artigos analisados.

No trabalho (MELO; FIGUEIREDO, 2021), é apresentada uma metodologia baseada em aspectos, tais como: modelagem de tópicos, reconhecimento de entidades nomeadas e análise de sentimentos, que visa obter os principais assuntos e temas em discussão na mídia jornalística e nas mídias sociais, e consequentemente é feita a aplicação dessa metodologia para análise do impacto da pandemia de COVID-19[7] no Brasil. Nesse trabalho foram coletados e analisados 18.413 artigos da mídia de notícias e 1.597.934 *tweets* postados por 1.299.084 usuários no Brasil. Os resultados alcançados com a metodologia apresentada nesse trabalho, mostraram ganhos na análise de sentimento de tópico ao longo do tempo, permitindo um melhor monitoramento da mídia da internet. Além desse resultado, esse trabalho também mostrou que o *Twitter* apresenta tópicos similares aos presentes nas notícias de mídia jornalística, sendo as principais entidades similares, havendo diferença na distribuição dos temas e na diversidade das entidades. Portanto, podemos afirmar que este estudo identificou os principais temas em discussão nas notícias e nas mídias sociais e como seus sentimentos evoluíram ao longo do tempo.

Na pesquisa (PÉREZ-ROSAS et al., 2017), foram construídos dois novos conjuntos de dados de notícias falsas, um obtido via *crowdsourcing* pela *Amazon Mechanical Turk* (AMT) que abrange seis domínios de notícias, com textos contendo

[7] ORGANIZAÇÃO MUNDIAL DA SAÚDE. Coronavírus. Disponível em: https://www.who.int/health-topics/coronavirus. Acesso em: 2 maio 2023.

principalmente a propriedade enganosa das notícias falsas, e o outro conjunto contendo notícias falsas sobre celebridades, que foram extraídas diretamente da *WEB*. Já no começo da coleta, foi percebido que a maior parte das notícias relacionadas a celebridades possuíam títulos sensacionalistas de fatos, tais como: divórcios, lutas, entre outros. Já em alguns outros casos de notícia ocorria a negação desses fatos. Nesse trabalho, foi realizada a construção de modelos de detecção de notícias falsas, e para isso foram extraídos vários conjuntos de características linguísticas (*features*). As principais *features* utilizadas, foram: N-gramas, pontuação, *features* psicolinguísticas (*Linguistic Inquiry and Word Count* (LIWC)), legibilidade e sintaxe. Com isso, foram conduzidos diversos experimentos, com diferentes combinações dessas *features*. Para o *dataset* de notícias falsas obtido por *crowdsourcing* via AMT, o resultado mais positivo, com 78% de acurácia, foi alcançado usando a *feature* de legibilidade, seguido pela combinação de todas as *features*, com 74% de acurácia. Já para o *dataset* de notícias falsas sobre celebridades, o modelo com maior acurácia sem a utilização de todas as *features*, foi obtido utilizando a *feature* de pontuação, com 70% de acurácia, seguido pela utilização da *feature* de N-gramas, com 67% de acurácia. Outro experimento realizado nesse trabalho foi uma análise entre domínios, usando os dois *datasets*. Para esse experimento, inicialmente foi utilizado o *dataset* de notícias de celebridades para o treino do modelo, e posteriormente o *dataset* de notícias via AMT para o teste do modelo. Nesse caso, em específico, a maior acurácia ocorreu na utilização da *feature* de legibilidade, o que resultou em 61% de acurácia. Após isto, foi utilizado o *dataset* de notícias via AMT para o treino, e o *dataset* de notícias de celebridades para o teste. Com essa configuração, o melhor resultado foi também de 61% de acurácia, obtido com o auxílio da *feature* LIWC, que engloba um conjunto de ferramentas para análise textual, tais como: categorias resumidas (por exemplo, pensamento analítico, tom emocional), processos linguísticos (por exemplo, palavras de função, pronomes) e processos psicológicos (por exemplo, processos afetivos, processos sociais).

No trabalho (REIS, 2020), foi utilizada uma base de dados que continha notícias divulgadas durante a eleição presidencial brasileira de 2018[8], advindas do *Whatsapp*. O trabalho objetivou a princípio, estudar características (*features*) que são importantes para detecção de *fake news* junto a métodos de aprendizado de máquina, e um dos

[8] Disponível em: https://www.tse.jus.br/eleicoes/eleicoes-2018

subprodutos do trabalho visou trazer uma pontuação de falsidade para as notícias, utilizando a extração dessas *features* estudadas como uma forma de treinar o modelo de aprendizado de máquina, e ranquear as notícias de acordo com suas pontuações. Para construir este ranqueamento foi utilizado o classificador *Extreme Gradient Boosting* (XGB), que realiza o processo de aprendizado de máquina com os dados de treino visando atribuir uma pontuação relativa à falsidade de notícias disseminadas no *Whatsapp* via imagens. Com esse modelo, e utilizando as *features* consideradas importantes para detecção de falsidade de notícias nesse trabalho, o autor conseguiu intervalos de confiança de 95%.

Em (MONTEIRO et al., 2018), foram coletadas e rotuladas amostras de notícias em língua portuguesa para a criação do *corpus* Fake.Br, visando manter o alinhamento de notícias falsas e verdadeiras, para realizar a construção de uma base de notícias balanceada. Além disso, o *corpus* foi construído, de maneira onde as notícias foram disponibilizadas em formatos de textos simples, e com tamanhos semelhantes, para evitar interferências em possíveis algoritmos de aprendizado de máquina que pudessem, posteriormente, serem aplicados na base de dados. No total, foram coletadas 7.200 notícias, com exatas 3.600 notícias verdadeiras e 3.600 notícias falsas. No geral, as notícias coletadas foram divididas em 6 grandes categorias: política, TV e celebridades, sociedade e notícias diárias, ciência e tecnologia, economia e religião. Além da construção do *corpus*, esse trabalho criou um classificador automático de *fake news* com uso de técnicas de aprendizado de máquina sobre o *corpus* construído e para isso foi aplicado o algoritmo *Suport Vector Machine* (SVM). O algoritmo foi aplicado incluindo diferentes *features* tanto separadamente, quanto em conjunto nos experimentos realizados, tais combinações são: classes gramaticais, classes semânticas, modelo *bag-of-words* (BOW), classes gramaticais + classes semânticas + modelo *bag-of-words*, pausalidade, emotividade, incerteza, não imediatismo, pausalidade + emotividade + incerteza, e por fim todas as *features* combinadas. Em um contexto geral, ao utilizar o modelo *bag-of-words* chegou-se a um resultado de 88% de assertividade utilizando o teste de acurácia *f-measure*. Além do método *bag-of-words*, também foi possível observar resultados bem positivos com o método *bag-of-words* combinado com a *feature* de emotividade, e também usando todas as *features*, onde para estes dois casos foi possível chegar a uma acurácia de 89%.

Por fins experimentais, esse trabalho também testou outros métodos de aprendizado de máquina, tais como: *Naive-Bayes*, *Random Forest* e *Multilayer Perceptron*. Além disso, foi utilizado o método *bag-of-words* com diferentes números mínimos de ocorrência no *corpus*, bem como outros valores para a ocorrência de palavras, como sua frequência (normalizada). Com isso, o método *Multilayer Perceptron* chegou a atingir 90% de precisão.

Um último experimento feito nesse trabalho foi executar o algoritmo sem truncar o tamanho do texto. Com isso, usando o texto completo chegou-se a 96% de precisão com o método *bag-of-words*, mas essa classificação é provavelmente tendenciosa, pois textos verdadeiros são significativamente mais longos do que os falsos.

Na pesquisa (NEWMAN et al., 2006), foi apresentada uma combinação de modelos estatísticos de modelagem de tópicos com reconhecedores de entidades nomeadas com o objetivo de fazer uma análise conjunta entre as entidades mencionadas em notícias com os tópicos discutidos nessas mesmas notícias. Para isso foi utilizada uma coleção de 330.000 artigos de notícias do jornal *New York Times* retirada do *corpus* (*Linguistic Data Consortium's English Gigaword Second Edition*), com uma filtragem de busca por notícias do tipo história datadas do ano 2000 até o ano de 2002. Para extração das entidades nomeadas dessas notícias foram avaliadas duas ferramentas inicialmente. A primeira ferramenta denominada *A Nearly-New Information Extraction System* (ANNIE) é um componente da ferramenta principal *General Architecture for Text Engineering* (GATE) baseada em regras e faz grande uso de dicionários geográficos. A segunda ferramenta utilizada denomina-se *Coburn's Perl Tagger*, e é baseada no *tagger part-of-speech Brill's* HMM. Como decorrência dessa avaliação, foi perceptível o perfil conservador da ferramenta ANNIE em identificar nomes próprios. Consequentemente, foi optado nessa pesquisa pela utilização da ferramenta denominada *Coburn's tagger* para extração das entidades nomeadas. O resultado dessa extração foram mais de 100.000 pessoas únicas, organizações e localizações. Desse resultado foi realizada uma filtragem com a condição de que uma entidade deve-se fazer parte de no mínimo 10 notícias diferentes, o que resultou em um total de 60.000 entidades.

Como experimento, essa pesquisa realizou a execução de modelagem de tópicos nesse *dataset* com uma escolha inicial de 400 tópicos, o que resultou em uma modelagem que trouxe certas associações com as entidades extraídas anteriormente, onde a modelagem pode inclusive inferir a probabilidade de um tópico particular a

partir de uma dada entidade. Além disso, nessa pesquisa foi utilizada a modelagem de tópicos para determinar as relações entre entidades baseada em tópicos, em outras palavras, a abordagem permite relacionar um par de entidades que nunca foram mencionadas em conjunto. Nesse caso, a relação é criada quando a afinidade das duas entidades está acima de um determinado limite. Essa afinidade entre duas entidades foi definida em termos das probabilidades condicionais envolvendo as duas entidades, conforme segue:

$$Afinidade(e_1, e_2) = \frac{\big(p(e_1|e_2) + p(e_2|e_1)\big)}{2}$$

Por fim, a importância desse trabalho consistiu em demonstrar que a modelagem de tópicos pode ter uma importante função em análises de grandes conjuntos de dados textuais. Além disso, foi demonstrada como as contribuições relativas dos tópicos mudaram ao longo do tempo, fazendo um paralelo com os principais eventos noticiados. Outro resultado interessante consiste no fato de que a modelagem de tópicos pode de maneira automática extrair conexões entre pessoas (entidades), que compartilham os mesmos tópicos.

No trabalho (PRITZKAU, 2022), foi realizada uma abordagem para a detecção de notícias falsas em textos. Nesse sentido, a tarefa foi especificada como um problema de classificação multiclasse. Como auxílio para identificar os principais padrões de conteúdo nos dados de treinamento, foi utilizada a modelagem de tópicos com o uso do método *Latent Dirichlet Allocation* (LDA). Além disso, no sentido de atribuir rótulos de classe aos documentos estudados foram utilizados os métodos RoBERTa (*A Robustly Optimized BERT Pretraining Approach*) e o método *Longformer* como uma arquitetura de rede neural para classificação sequencial. Nesse caso, o trabalho foi iniciado com um modelo pré-treinado para representação de linguagem, e posteriormente foram realizados ajustes no modelo de forma a adequá-lo ao problema de classificação com os dados para o treinamento supervisionado, e para isso o treinamento do classificador foi feito em nível de tópicos. O *dataset* utilizado para essa tarefa foi desenvolvido durante o evento CLEF-2021 *CheckThat!*, que consiste de 1.264 documentos. Além disso, outros dois *datasets* também foram coletados para a tarefa sendo eles: *Fake News Detection Challenge KDD 2020* e o *Fake News Classification Datasets*. Portanto, o resultado foi um *corpus* de treino com 51.148 documentos.

O processo de análise exploratória usado nesse trabalho é baseado em análise exploratória compreensiva de dados de treino. Como o conjunto de dados de treinamento inicial continha 1.264 documentos, então foram investigadas inicialmente as inconsistências nesse conjunto de dados, resultando na remoção de ambiguidades e duplicatas. Por fim, o conjunto de dados limpos passou a ter 1.096 documentos. Por outro lado, o *corpus* no total ficou com 44.910 documentos após limpezas nos outros *datasets*.

Além disso, foi constatado no conjunto de treino que a maior parte das sequências na contagem de *tokens* excedeu o limite dos modelos baseados em *Transformer*, consequentemente foi optado por mudar a arquitetura para o *Longformer*, que é uma arquitetura modificada derivada do *Transformer*. Além disso, o desbalanceamento dos dados foi outro problema encontrado no *dataset*, isto é, a classe de notícias verdadeiras possuía uma quantidade bem inferior em comparação com a classe de notícias parcialmente falsa.

Como forma de obter uma visão geral das informações que os dados continham foi utilizada a modelagem de tópicos, com o uso do algoritmo LDA para identificar os padrões de conteúdo nos dados de treinamento, ou seja, distinguir entre tópicos que são semanticamente interpretáveis e tópicos que são artefatos de inferência estatística. A abordagem desse trabalho foi baseada na suposição de que a diferenciação de vários pontos de vista geralmente ocorre de maneira relacionada aos tópicos e, consequentemente, um tópico é resultado de uma distribuição específica das palavras usadas. Com isso, através desta distribuição, diferentes tópicos podem ser distinguidos uns dos outros. Portanto, a ideia utilizada consistiu em basear a classificação automática de texto nos tópicos. Nesse sentido, o treinamento e a previsão realizada nesse trabalho foram precedidas pela modelagem de tópicos para primeiro dissecar padrões de conteúdo nos dados que foram estudados. Com isso foi aplicado o método LDA, e consequentemente um classificador específico foi treinado para cada tópico derivado.

De maneira geral, a modelagem de tópicos e a classificação de texto aplicada nesse trabalho seguiu o seguinte pipeline:

- *Input Embeddings* – Converte as entradas em sequências de *features*.
- *Word-Level Sentence Embeddings* – Cada sentença é dividida em palavras pelo *WordPiece tokenizer*.
- *Target Encoding* – Os rótulos alvos são codificados usando codificação de rótulo, assumindo a variável alvo como categórica.

Nesse trabalho, considerando a predição nos dados originais de treino com a sobre-amostragem foram alcançados resultados no processo de classificação das notícias com precisão de 59,33% para notícias falsas, nas notícias parcialmente falsas foi obtida uma precisão de 15,66%, e nas notícias verdadeiras foi alcançada uma precisão de 68,18%. Em uma visão macro, o efeito da sobre-amostragem no trabalho melhorou a média da métrica F1 de 0,2550 para 0,2736.

Em (NWANKWO; OKOLO; HABONIMANA, 2020), é debatida uma forma de conter a disseminação de desinformação por meios, tais como: *Facebook* e *Whatsapp*. A ideia central desse trabalho visa o combate a desinformação sobre a COVID-19 no contexto da África subsariana. Para isso, são realizadas recomendações no intuito do desenvolvimento de novas técnicas de *machine learning* e inteligência artificial para combater a desinformação. Segundo o trabalho, tanto o *WhatsApp*, quanto o *Facebook* têm desenvolvido certas funcionalidades nesse sentido, tais como:

- A limitação do encaminhamento de mensagens entre grupos para até cinco grupos no *WhatsApp*.
- A criação do *chatbot WHO Health Alert* que provê informações confiáveis sobre a COVID-19.
- *Facebook* e *Twitter* iniciaram a rotulação de posts como "*altered*" ou "*disproven*" como forma de indicar a não veracidade desses posts.

Porém, segundo o trabalho apesar desses esforços, as informações falsas continuam a ser disseminadas na região da África subsariana, e de uma maneira geral nos países do Hemisfério Sul. Segundo o artigo, técnicas como LDA ou *Bidirectional Encoder Representations from Transformers* (BERT) poderiam ser usadas para resolver o problema, contribuindo tanto com governo desses países, quanto com as organizações, na tarefa de identificar tópicos principais nas notícias falsas disseminadas, o que poderia potencializar o foco de combate das organizações.

Ainda segundo o artigo, a vantagem do LDA é a abordagem por *backtrack* que é útil na geração de palavras relacionadas a tópicos presentes em documentos ou formas mais curtas de texto. Já a vantagem do BERT é que o mesmo encontra relações contextuais entre palavras, o que permite alcançar bons resultados com análise de sentimentos, resposta a questões, e também no reconhecimento de entidades nomeadas.

Outra questão debatida no artigo refere-se a coleta dos dados, onde são propostos caminhos mais práticos de coleta de *datasets* com esse intuito específico de pesquisa por meio de *crowdsourcing*, ou seja, segundo o trabalho já existem exemplos de iniciativas trabalhando nessa questão, tais como: *Afara International* e *Africa Checks*. Portanto, a sugestão consiste na união de esforços entre as companhias para a criação de bases de dados úteis para o combate à desinformação.

Por fim, o artigo destaca o perigo que a desinformação sobre a COVID-19 e outros problemas relacionados à saúde podem representar para a sociedade. Desse modo, realça a importância de estudos que utilizem as sugestões de abordagem supracitadas, o que possibilita diversas vertentes de trabalho.

Com base no resumo expresso acima sobre os principais trabalhos relacionados, torna-se necessário realizar uma análise comparativa dos trabalhos relacionados em termos de conjuntos de dados, contexto de coleta, tarefa (ou objetivo) inicial designado e métodos utilizados para resoluções. Para isso, foi construído o Quadro 2 no presente trabalho, com o intuito de trazer essa análise comparativa trazendo as principais informações referentes a cada um dos trabalhos estudados.

Quadro 2 – Comparação entre os principais trabalhos relacionados concentrados na tarefa de caracterização e identificação de notícias falsas

(continua)

Trabalho	**Conjunto de dados**	**Contexto original**	**Tarefa**	**Observações verdadeiras**	**Observações falsas**	**Métodos utilizados**
(MELO; FIGUEIREDO, 2021)	Artigos de notícias e tweets coletados	*Websites*	Identificação de padrões(tweets/notícias)	18.413 notícias verdadeiras	0	Modelagem de Tópicos e Análise de Sentimentos
(MONTEIRO et al., 2018)	*Fake.Br*	*Websites*	Detecção de *Fake News*	3.600	3.600	*SVC, Naive-Bayes, Random Forest, Multilayer Perceptron* e *bag-of-words*.
(PÉREZ-ROSAS et al., 2017)	*FakeNewsAMT* e *Celebrity*	*Websites*	Detecção de *Fake News*	*FakeNewsAMT*(240) e *Celebrity*(250)	*FakeNewsAMT*(240) e *Celebrity*(250)	*Support Vector Machine* com uso de diferentes *features*
(REIS; BENEVENUTO, 2021)	2016 *US Election (BuzzFace dataset)* 2018 *Brazilian Election Dataset FakeHealth dataset*	*Facebook, Whatsapp* e *Twitter*	Detecção de *Fake News*	2016 *US Election (BuzzFace dataset)* (1.669); 2018 *Brazilian Election Dataset* (Aproximadamente 4.500 imagens) *FakeHealth dataset* (1.533)	2016 *US Election (BuzzFace dataset)* (349); 2018 *Brazilian Election Dataset* (Aproximadamente 4.500 imagens) *FakeHealth dataset* (763)	*XGB* com o uso de diferentes *features*
(NEWMAN et al., 2006)	Linguistic Data Consortium's English Gigaword Second Edition	*Websites*	Identificação de padrões em notícias por meio de modelagem de tópicos	330.000	0	Modelagem de tópicos e Extração de entidades (*Coburn's tagger*)

(conclusão)

Trabalho	Conjunto de dados	Contexto original	Tarefa	Observações verdadeiras	Observações falsas	Métodos utilizados
(PRITZKAU, 2022)	*Corpus* que abrange os seguintes Datasets: 1) Dataset desenvolvido durante o evento *CoCLEF-2021 CheckThat!*; 2) *Fake News Detection Challenge KDD 2020*; 3) *Fake News Classification Datasets*	*Websites*	Identificação de padrões em notícias por meio de modelagem de tópicos e identificação de notícias falsas	Aproximadamente 24.000 observações verdadeiras	Aproximadamente 20.000 observações falsas	*BERT*, *RoBERTa* e *Longformer*
(NWANKWO; OKOLO; HABONIMANA, 2020)	Sugestão da criação de *corpus* via *crowdsourcing* por meio de companhias, tais como: *Afara International* e *Africa Checks*	A depender do objetivo do estudo	Identificação de padrões de notícias falsas			Sugestão de utilização dos métodos *BERT* e *LDA*

Fonte: Elaborado pelo autor (2023).

Portanto, ao comparar os trabalhos estudados durante a revisão de literatura, temos que a maior parte deles, foca na tarefa de identificação de notícias falsas e também na detecção de padrões textuais das mesmas. Através do Quadro 2, observa-se que há uma lacuna no estudo e caracterização de conjuntos de dados não-balanceados contendo apenas notícias falsas com o intuito de identificar os padrões temporais, categóricos e das entidades nomeadas presentes nos textos de desinformação. Dessa forma, o presente trabalho tem por objetivo preencher essa lacuna deixada pelos principais trabalhos estudados. No presente trabalho é realizada uma análise exploratória considerando tanto o fator temporal, quanto o fator categórico para a caracterização de notícias falsas de um *corpus* não-balanceado contendo apenas notícias falsas, que será detalhado no Capítulo 5. Além disso, no presente trabalho também são realizadas modelagens de tópicos realizando divisões no *corpus* no Capítulo 6, de forma a compreender e caracterizar tanto temporalmente, quanto categoricamente as notícias falsas estudadas por meio da identificação de padrões nos textos estudados. Um dos pontos mais importantes do presente trabalho consiste na comprovação da hipótese de que os períodos temporais, as temáticas e as entidades nomeadas mencionadas nas notícias falsas estudadas são fatores importantes para a identificação de padrões nesses textos.

4 TRATAMENTO TEXTUAL DO CORPUS

Este capítulo demonstra todo o processo relativo ao tratamento textual do *corpus* Fakepedia realizado no presente trabalho. Em primeiro lugar é listada toda a arquitetura de software utilizada para o trabalho, listando os principais softwares utilizados no tratamento. Após isso, é realizada a descrição dos dados utilizados neste trabalho, trazendo uma visão geral do *corpus* Fakepedia. Em seguida, realiza-se a descrição geral do tratamento textual realizado nos dados utilizados neste trabalho, relatando os principais métodos de processamento de linguagem natural usados no tratamento para posterior análise. Por fim, é descrita a etapa de aperfeiçoamento dos dados que relata as principais necessidades de dados identificadas ao longo da pesquisa, e os processos envolvidos para suprir as principais necessidades de dados, que envolvem principalmente, a técnica de *Web Scraping*.

4.1 ARQUITETURA

Neste tratamento específico, a linguagem de programação Python[9] foi escolhida como base principal. Para realizar a análise de dados, foram utilizadas diversas bibliotecas, tais como: *Pandas*[10], *Matplotlib*[11], *Numpy*[12], *SpaCy*[13], *NLTK*[14], *Gensim*[15], além de outras bibliotecas mencionadas durante todo o processo. Com o intuito de garantir um ambiente adequado para executar as análises, foi optado por utilizar tanto o *Google Colab*[16], para realizar o tratamento de forma remota, quanto o *Jupyter Notebook*[17], para executá-lo localmente.

Adicionalmente, considerando a importância do versionamento do corpus e a necessidade de atualizar os dados conforme a limpeza e inserção de novas informações, também foi utilizado o *GitHub*[18] como uma solução eficiente para armazenar o corpus aprimorado. Essa abordagem permitiu a realização de um

[9] Disponível em: https://www.python.org
[10] Disponível em: https://pandas.pydata.org
[11] Disponível em: https://matplotlib.org
[12] Disponível em: https://numpy.org
[13] Disponível em: https://spacy.io
[14] Disponível em: https://www.nltk.org
[15] Disponível em: https://radimrehurek.com/gensim/
[16] Disponível em: https://colab.research.google.com
[17] Disponível em: https://jupyter.org
[18] Disponível em: https://github.com

controle preciso sobre as versões do corpus, garantindo que todas as modificações fossem registradas adequadamente.

É válido ressaltar que a escolha da linguagem Python e das bibliotecas mencionadas proporcionaram uma base sólida para a análise de dados realizada. A flexibilidade e variedade de recursos disponíveis no Python permitem analisar os dados de forma eficaz, atingindo os resultados desejados.

Por fim, a escolha de realizar a combinação entre o *Google Colab* e o *Jupyter Notebook*, viabilizaram a execução eficiente do tratamento tanto em ambientes remotos quanto locais. Essa flexibilidade proporcionou adaptar o fluxo de trabalho de acordo com as necessidades específicas de cada etapa do processo de tratamento de dados.

4.2 DADOS

Os dados usados neste trabalho são referentes ao *corpus* Fakepedia. O *corpus* Fakepedia possui uma dimensão de 8.517 observações (notícias) e 10 colunas (informações sobre as notícias), onde tais colunas são as seguintes:

- ***Title*** - Contém o título da notícia;
- ***Title_norm*** - Armazena o título da notícia normalizado, ou seja, o texto do título da notícia com um tratamento inicial usando técnicas de processamento de linguagem natural;
- ***Message*** - Contém o texto original da notícia, sem alterações;
- ***Message_norm*** - Armazena o texto da notícia com um tratamento inicial de normalização, que foi realizado por meio de técnicas de processamento de linguagem natural;
- ***Tokens*** - Separa as palavras do título da notícia em uma lista de tokens;
- ***Features*** – Responsável por armazenar as principais características sobre as observações;
- ***Entities*** - Contém as principais entidades nomeadas presentes nas notícias;
- ***Type*** - Determina se a notícia é falsa;
- ***Source*** - Armazena a fonte da notícia, isto é, nome da fonte.
- ***Url_review*** - Contém a URL específica que direciona para a notícia na *WEB*.

Como observação pertinente, é necessário ressaltar que todas as notícias presentes no *dataset* são falsas, em outras palavras, para todas as observações, temos na coluna ***type***, o valor “*false*”. Desse modo, vale ressaltar que um dos principais diferenciais do presente trabalho está no fato de operar em um *dataset* não-balanceado.

4.3 TRATAMENTO DOS DADOS

Nesta seção, será realizada a descrição do processo de tratamento dos dados realizado no *dataset* utilizado. Como o *dataset* utilizado possui notícias extraídas de páginas da *WEB*, foi perceptível desde o princípio do estudo, a presença de frações de dados que não são úteis para esta análise, tais como: presença de anúncios no texto das notícias, presença de informações inúteis, especificamente: *e-mails*, telefones, sinais de pontuação e outros dados não relevantes no texto das notícias.

Tendo em vista essa perspectiva, inicialmente optou-se por realizar um tratamento inicial dos dados, onde explorou-se o *dataset* com o intuito de remover esses dados ou frações de dados que não possuem utilidade para o objetivo do trabalho, isto é, remover dados que tanto não tragam informações relevantes para a análise temporal das notícias, quanto não sejam úteis para a construção do dicionário de tópicos, que servirá de base para futuros trabalhos de identificação e classificação de notícias falsas.

Neste tratamento inicial, optou-se por adicionar duas colunas ao *dataset* com o intuito de armazenar o resultado final do tratamento inicial. Uma das colunas contém o texto das notícias tratado com a remoção das frações de dados não relevantes descritos anteriormente, já a outra coluna contém o texto das notícias tratados com a remoção de dados não relevantes descritos anteriormente, e também com a remoção das *stopwords* visando posteriormente aplicar a modelagem de tópicos no texto das notícias. Portanto, pode-se utilizar duas fontes de dados: uma considerando o texto das notícias com a presença das *stopwords* e pontuações, e outra considerando o texto das notícias sem a presença das *stopwords* e pontuações.

Como citado anteriormente, o *corpus* possui a coluna “*message_norm*”, que originalmente é uma coluna que armazena o texto da notícia normalizado e tratado com técnicas de processamento de linguagem natural, que foram definidas e abordadas na seção 2.2 deste trabalho. Portanto, na abordagem deste trabalho optou-se por aproveitar esse pré-tratamento, e foi realizada a criação de uma nova coluna chamada “*message_norm_treatment*”, onde foi realizada a remoção de toda fração de dados que não sejam úteis para a análise posterior.

Inicialmente, foi realizada uma consulta na base a fim de procurar por informações não relevantes, e o primeiro problema encontrado no *corpus* foram amostras na coluna “*message_norm*” que continham valores *Not a Number* (NaN) por

padrão, ou seja, algumas das amostras do *corpus* Fakepedia não possuíam o texto das notícias. Para este caso foi realizada a substituição dos valores NaN por *string* vazia, com o objetivo de evitar os possíveis problemas advindos da manipulação destes valores na posterior análise dos dados, tendo em vista que não são valores úteis para a análise, e também não foi optado por eliminar essas amostras do *dataset*, visto que, a princípio, essas amostras não possuíam o texto da notícia, porém possuíam outras informações, tais como: título, URL, entre outras, que são dados úteis para o presente trabalho. Após isso, foram encontradas palavras e frases que denunciavam a falsidade da notícia, um exemplo foi a ocorrência da palavra “boato”, que tanto estava associada à propaganda do site “boatos.org” presente no próprio site de onde foram extraídas as notícias para esse *corpus*, como também estava associada à indicação na própria notícia de que a mesma era falsa. Para esse caso, foi optado por realizar o tratamento dessa ocorrência com a substituição por *string* vazia, tendo em vista que qualquer outro tipo de substituição, ou a não substituição poderia vir a interferir na análise posterior dos dados, isto é, como as etapas posteriores visam a construção do dicionário de tópicos, que pode futuramente ser usado como base na construção de um classificador probabilístico de notícias falsas, deve-se evitar a presença de palavras que indiciem a falsidade das notícias, visando não adicionar parcialidade ao dicionário final.

Após isso, também foi encontrado outro padrão que se repetia em uma quantidade expressiva das notícias, que se trata de uma mensagem automática criada no fim das notícias semelhante ao seguinte texto: “*Ps.: Esse artigo é uma sugestão de leitores do Boatos.org. Se você quiser sugerir um tema ao Boatos.org, entre em contato com a gente pelo site, Facebook e WhatsApp no telefone (61) 99458-8494.*”, por conseguinte efetuou-se a limpeza dos textos das notícias removendo esse padrão. Porém, foi possível perceber que esse padrão não somente ocorria no fim de notícias, como também em alguns casos, todo o texto da notícia era composto esse padrão. Além disso, com o decorrer da limpeza foram detectadas variações desse mesmo padrão com diferenças, tais como:

1. Presença da palavra “este” no lugar de “esse”
2. Presença da palavra “foi” no lugar de “é”
3. Presença do símbolo “:” no lugar de “.:”
4. Ausência de “ps:”, isto é, a notícia começa com “este” ou “esse”
5. Variações de números de telefone presentes no trecho de texto

Dessa maneira, optou-se por realizar a limpeza de todas essas variações substituindo esses padrões tanto nos finais das notícias, quanto nos casos onde toda a notícia equivalia a esse padrão. Para essa limpeza específica, foi realizada a substituição desse padrão por uma *string* vazia, de forma a preservar a consistência das informações de base presentes no *corpus*. Inicialmente essa substituição ocorreu com a utilização do método *built-in* de *string* em Python (*replace*). Porém, como esse padrão possuía diferentes variações, optou-se por construir uma expressão regular para captura dessas diferentes variações, e com isso substituiu-se o tratamento feito com o método *replace*, por um tratamento com o uso dessa expressão regular que foi construída para abranger todas essas variações, e a aplicação do tratamento em todo o *dataset* foi realizada por meio do método *apply* com o uso da função lambda.

Além disso, foi identificada tanto a presença de números de telefone em outros trechos de notícias, quanto a presença de *links* para o site “boatos.org”, que é a principal fonte das notícias do *corpus* Fakepedia. Considerando que os números de telefone e os *links* não seriam úteis para a análise posterior, e também com a percepção de que os *links* para o site “boatos.org” poderiam ser indícios que apontam a falsidade da notícia, o que traria parcialidade para a construção do dicionário. Adicionalmente, destacando o fato de que os telefones seriam capturados nas modelagens de tópicos que viriam a ser realizadas posteriormente devido à grande quantidade de ocorrências, o que não traria informação útil ao dicionário, então optou-se por realizar a limpeza destes números de telefone e *links*.

Além desses padrões, percebeu-se durante o tratamento, outro padrão presente em muitas notícias do *corpus*, que se trata de um texto descritivo presente ostensivamente no final de diversas notícias feito pelo editor-chefe do portal verificador de notícias falsas (Edgard Matsuki). Como o texto não possuía relação intrínseca com as notícias estudadas e continha indícios que apontavam um viés de falsidade, optou-se também por realizar a remoção desse padrão de texto do final do texto das notícias, com o intuito de preservar apenas o texto-base da notícia.

Outro padrão encontrado foi um resquício de texto contendo informações de *Whatsapp*. Nesse caso, o texto completo da notícia se resume a esse padrão, fato que pode demonstrar que o processo realizado para a construção do *corpus* Fakepedia pode ter eliminado o resto do texto da notícia fonte.

Posteriormente, ao realizar consultas nas notícias, percebeu-se a presença de outro padrão no fim de notícias do site que, por sua vez, está contido em uma tabela

do site “boatos.org” solicitando aos leitores que sigam redes sociais. Esse padrão possui links gerados pelo encurtador de links *Bitly*[19]. Desse modo, foi necessário construir expressões regulares para eliminar esses padrões contendo links encurtados com o referido encurtador tendo em vista que os *links* encurtados não acrescentam informações úteis para a análise de dados, e podem vir a ser capturados para o dicionário durante a modelagem de tópicos, o que não trará informação relevante.

Em uma etapa subsequente, foi optado por criar um novo *dataframe* com o intuito de realizar uma série adicional de limpezas considerando o *corpus* sem as notícias duplicadas, o que facilitou as consultas em buscas de novas frações de dados que não possuíam utilidade para a análise das notícias. Com essa nova abordagem, removendo as notícias duplicadas utilizando como parâmetro o texto original da notícia como fator verificador de duplicidade, constatou-se que o novo *dataframe* possuía 4.209 observações, isto é, 4.209 notícias, o que equivale à menos da metade das notícias presentes no *corpus* original, que é 8.517 notícias. Porém, após isso foi realizada a averiguação de duplicidade novamente, dessa vez considerando o texto original da notícia, e o título da notícia como fatores verificadores de duplicidade. Com essa nova metodologia, constatou-se que o novo *dataframe* possuía 5.201 notícias. Portanto, com essas duas metodologias foi detectada a existência de notícias com mesmo texto, porém com títulos diferentes, o que explica o fato da redução de notícias ser menor ao considerar o texto e o título da notícia. Para fins de consolidar este resultado, foram mesclados novos fatores verificadores de duplicidade considerando outras colunas do *dataset*, tais como: URL, *tokens* e entidades. Porém, o resultado obtido foi mantido com um total de 5.201 notícias, o que reforça o resultado da análise feita, em outros termos, corrobora o fato de haverem notícias com mesmo texto e títulos diferentes presentes no *corpus* Fakepedia.

Utilizando esse novo *dataframe* foi possível verificar um novo trecho de texto presente nas notícias que começa com o seguinte texto: “*print de notícia falsa...*”. Como este trecho se repete para um número considerável de notícias, e a palavra “falsa” denuncia a falsidade da notícia, foi optado por realizar a remoção da palavra “falsa” do texto da notícia com o intuito de evitar que isto venha a interferir no objetivo de trazer imparcialidade ao dicionário que será construído, e que posteriormente poderá ser utilizado por classificadores. Além disso, foi identificado no final de uma

[19] Disponível em: https://bitly.com

quantidade considerável de notícias o seguinte trecho de texto: “*Ps2: Confira a nossa nova seção “Oportunidades” clicando aqui. Na página, você pode acesso a promoções, descontos e sites que dão brindes.*”, que não faz parte do texto da notícia, e sim de uma propaganda do próprio site. Consequentemente, optou-se por realizar a remoção deste trecho de texto, visto que o mesmo não pertence à notícia.

Outro problema verificado no texto das notícias, tem relação com valores monetários. No geral, em notícias contendo valores monetários, foi possível identificar que determinados números continham espaçamento entre as casas decimais, por exemplo: 2. 500. 000 (Dois milhões e quinhentos mil). Como posteriormente na análise das notícias serão aplicados métodos de modelagem de tópicos no *corpus*, o que requer a *tokenização* das notícias, e como existem muitas ocorrências de valores monetários, os métodos podem vir a considerar cada parte do número como uma palavra diferente e, consequentemente, ocorre a geração de agrupamentos de partes como: “000” como palavras frequentes em tópicos da modelagem, o que não é um fato, visto que “000” não corresponde à uma palavra frequente, e sim a um trecho de número. Além disso, como a *tokenização* considera vírgulas e pontos, como separadores de palavras, foi necessário realizar a remoção das vírgulas e pontos nos valores monetários, a fim de evitar a inclusão de partes de valores monetários como palavras frequentes em tópicos por parte do modelo.

Após isso, foi realizada a construção da coluna “*message_norm_treatment_ssw*”, onde foi criada uma cópia da coluna em que foi realizada a limpeza do texto das notícias. Dessa forma, foi realizada a *tokenização* dos textos das notícias, transformando cada texto de notícia em uma lista de *tokens*. Após a *tokenização* em cada amostra da coluna “message_norm_treatment_ssw” foi realizada a remoção das *stopwords*. Para a remoção das *stopwords*, a princípio foi utilizado o *corpus* da biblioteca NLTK, que possui diversas *stopwords* em sua base. Posteriormente, foi utilizada a função *FreqDist* da biblioteca NLTK, com o intuito de observar as palavras mais frequentes presentes na coluna, de modo a verificar a existência de *stopwords* ainda não removidas. Logo, constatou-se uma grande presença de *stopwords* com 3 ou menos caracteres. Consequentemente, foi construída mais uma lista de *stopwords*, e além disso foi feita a remoção de *stopwords* com 3 ou menos caracteres e, adicionalmente foi realizada a remoção manual de mais *stopwords* específicas. Posteriormente, o resultado final foi armazenado em uma nova coluna denominada “*message_norm_treatment_ssw2*”, que tem o tratamento mais

completo em relação a remoção de *stopwords*. O código do processo de tratamento textual do *corpus* realizado no presente trabalho pode ser visto de forma ainda mais detalhada no Apêndice A.

4.4 APERFEIÇOAMENTO DOS DADOS

Esta seção tem por objetivo explicar os aperfeiçoamentos realizados no *corpus* Fakepedia, especificando as inserções de dados úteis que foram realizadas para a viabilização e o prosseguimento do trabalho, bem como para a realização da análise exploratória e análise de evolução das notícias falsas.

Para a posterior etapa de análise dos dados foi identificada a necessidade de criar um novo *corpus* com base nas limpezas descritas na seção anterior. Além disso, foi possível identificar a necessidade de aprimorar o conjunto de dados do *dataset* Fakepedia, com o intuito de obter mais informações para estudar a evolução das notícias falsas ao longo do tempo. Uma das informações mais necessárias refere-se à data de publicação das notícias que não existem no *dataset* Fakepedia, visto que o presente trabalho também tem por objetivo realizar uma análise das notícias feita com base no tempo. Portanto, para obter as datas de publicação das notícias optou-se por realizar a técnica de *Web Scraping* com o intuito de extrair esses dados do site "boatos.org", que é a fonte das notícias presentes no *dataset* Fakepedia. Para isto, foram utilizadas as seguintes bibliotecas da linguagem Python: *requests*[20] (responsável por realizar as requisições de conteúdo das páginas *Web* referentes às notícias), *urllib*[21] (responsável por realizar a manipulação referente às URL's usadas na técnica de *Web Scraping*) e *BeautifulSoup*[22] (responsável pela extração dos dados das páginas HTML). Com a utilização destas bibliotecas foi possível realizar a identificação e captura das *tags* da página HTML do site "boatos.org" que continham os dados relativos as datas de publicação das notícias, bem como realizou-se a requisição dessas informações por meio das URL's de cada notícia presentes no *dataset* referentes às páginas HTML. Por fim, foi realizado o armazenamento destas informações em uma nova coluna do novo *corpus* denominada "***datetime***".

Após sanar esta primeira necessidade de dados, também foi identificada a ausência das categorias referentes a cada notícia no *corpus* Fakepedia, o que

[20] Disponível em: https://pypi.org/project/requests/
[21] Disponível em: https://docs.python.org/3/library/urllib.html
[22] Disponível em: https://pypi.org/project/beautifulsoup4/

impossibilitaria a identificação dos temas das notícias, como por exemplo: esportes, política, entretenimento, etc. Tendo em vista essa necessidade, optou-se por realizar no presente trabalho uma nova aplicação da técnica de *Web Scraping* fazendo uso das mesmas bibliotecas usadas na extração das datas de publicação das notícias, e por conseguinte foi possível agregar a categoria de cada notícia no novo *corpus* em uma nova coluna chamada "*category*".

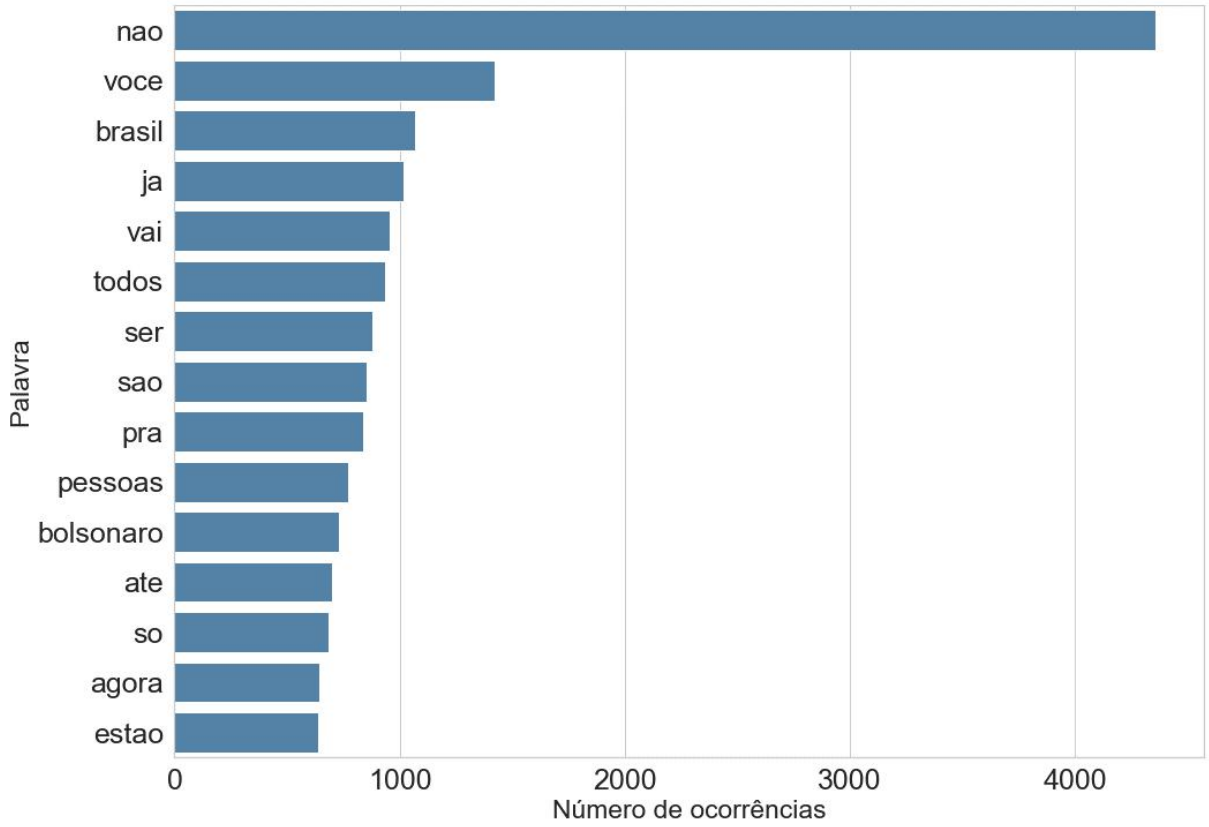
Os algoritmos de *Web Scraping* foram desenvolvidos no presente trabalho e estão dispostos como apêndices. O algoritmo para extração das datas de publicação das notícias está no Apêndice B, bem como o algoritmo para extração das categorias das notícias se encontra no Apêndice C. O *corpus* com os dados advindos do aperfeiçoamento está disposto no Apêndice H.

5 ANÁLISE EXPLORATÓRIA E EVOLUÇÃO DAS NOTÍCIAS FALSAS

Neste capítulo serão discutidas as análises de dados realizadas para o estudo da evolução temporal das notícias do *corpus* Fakepedia, bem como serão apresentados os principais resultados advindos dessas análises, com foco em apresentar tanto graficamente quanto estatisticamente a evolução das notícias falsas. É importante ressaltar que o escopo do trabalho se restringe ao *dataset* Fakepedia, isto é, em geral as análises tendem a refletir resultados advindos das notícias presentes nesse *dataset*, em específico. Portanto, não deve-se realizar generalizações que fujam ao escopo do objeto de estudo deste trabalho. O código referente a análise exploratória completa pode ser visto no Apêndice D.

O primeiro aspecto a ser considerado na análise textual está relacionado à frequência de palavras nas notícias de um modo geral. Para isso, foi utilizado a função *FreqDist* na coluna que contém o texto das notícias ainda com a presença de *stopwords*, e após isto foi aplicado a função *FreqDist* na coluna que contém o texto das notícias sem a presença de *stopwords*. No Gráfico 1, é possível observar a distribuição de frequência das palavras na coluna com a presença de *stopwords*.

Gráfico 1: Palavras mais frequentes no *corpus* incluindo *stopwords*

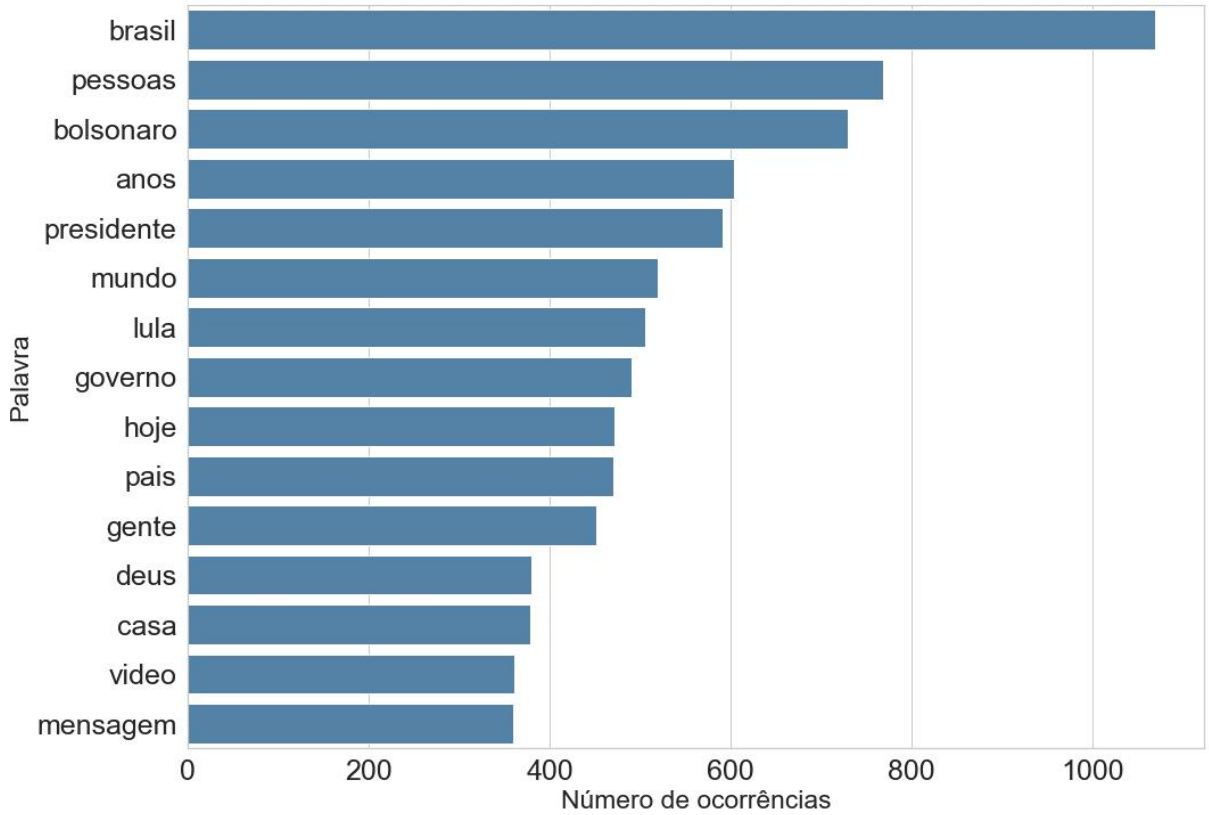


Fonte: Elaborado pelo autor (2023).

Com base no Gráfico 1 é possível constatar que a palavra mais frequente é o advérbio de negação “não”, o que tende a reafirmar a falsidade das notícias, isto é, trabalhos anteriores mostraram que enganos e declarações falsas podem ser detectadas a partir do estilo de escrita dos autores ou da linguística e, às vezes, podem ser usadas para inferir suas personalidades (PENNEBAKER; KING, 2000). Alguns autores mostraram que os mentirosos podem ser detectados enquanto contam histórias complexas, onde fazem menos auto-referências - para se dissociar da história, e tendem a ter uso mais frequente de emoção negativa nas palavras – como sinal de culpa (NEWMAN; PENNEBAKER, 2003). Portanto, é lógico considerar as emoções dentro dos textos postados como uma sugestão em relação ao ato de disseminação de notícias falsas.

No Gráfico 2 está caracterizada a distribuição de frequência das palavras na coluna sem a presença de *stopwords*.

Gráfico 2: Palavras mais frequentes no *corpus* sem *stopwords*



Fonte: Elaborado pelo autor (2023).

Com base no Gráfico 2, identifica-se que palavras, tais como: “Brasil”, “pessoas”, “Bolsonaro”, “Lula”, “presidente”, “governo”, entre outras são as mais frequentes nas notícias. Portanto, a remoção das *stopwords* auxiliou na obtenção de

um indício geral para o conteúdo mais frequente envolvido em notícias falsas, ou seja, em geral assuntos nacionais e que abordam questões políticas são os mais frequentes entre as notícias do *corpus*.

Além do gráfico de distribuição de frequência de palavras, também foi plotada a nuvem de palavras referente aos *tokens* mais frequentes contidos na amostragem geral, isto é, considerando todas as notícias presentes no *dataset*. No Gráfico 3, temos uma nuvem de palavras que, por sua vez, traz uma melhor percepção da frequência das principais palavras presentes nas notícias falsas, considerando o tamanho relativo à presença das principais palavras no texto.

Gráfico 3: Nuvem de palavras das notícias do *corpus*

Fonte: Elaborado pelo autor (2023).

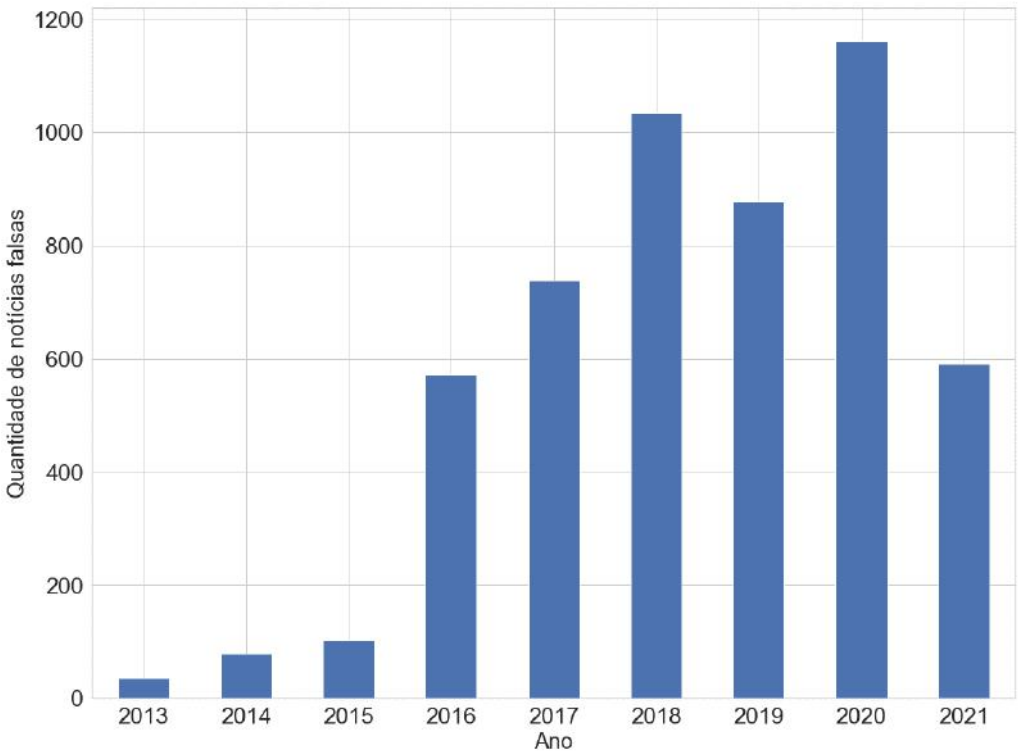
Podemos verificar que no conjunto geral, a palavra mais recorrente é “Brasil”, o que nos remete também ao fato de que uma parcela relevante das notícias falsas presentes no *dataset* de estudo trata de conteúdo desinformacional associado a assuntos nacionais.

Outra etapa da análise foi realizada com o intuito de identificar a frequência de notícias por período de tempo, utilizando os novos dados inseridos na etapa de aperfeiçoamento dos dados que, por sua vez, foram introduzidos na seção 4.4 do presente trabalho.

5.1 ANÁLISE TEMPORAL DAS PUBLICAÇÕES DAS NOTÍCIAS

O *dataset* Fakepedia aprimorado utilizado como base neste trabalho possui notícias compreendidas no período desde o ano de 2013 até o ano de 2021. No Gráfico 4, é possível observar a distribuição anual de notícias falsas no *dataset*.

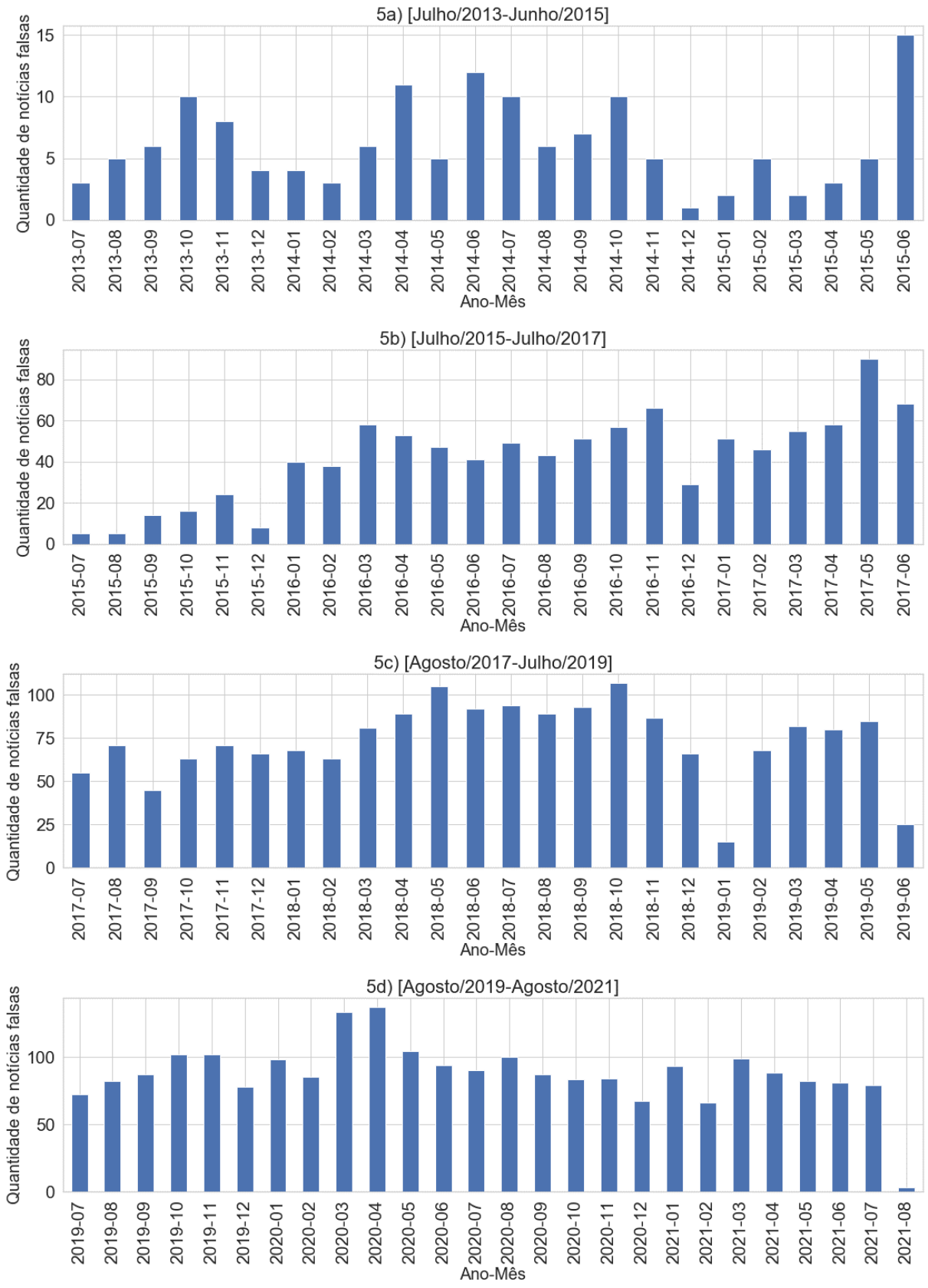
Gráfico 4: Quantidade de notícias falsas por ano



Fonte: Elaborado pelo autor (2023).

Podemos verificar que os anos de 2013, 2014 e 2015 tiveram quantidades relativamente inferiores de notícias falsas em comparação aos anos posteriores. Inclusive, ao realizar a soma das notícias desses três respectivos anos, temos um total de 220 notícias, ou seja, não se alcança a quantidade de notícias de nenhum dos anos posteriores considerados individualmente. Além disso, percebe-se no Gráfico 5, que mesmo considerando a média entre os meses dos anos de 2013, 2014 e 2015 não há um quantitativo considerável de notícias falsas ao realizar a comparação com o quantitativo dos anos posteriores. Essa quantidade baixa de notícias tende a trazer poucas informações, se considerarmos estes três anos de maneira individual na análise.

Gráfico 5: Quantidade de notícias falsas ano/mês. 5a) Julho/2013-Junho/2015. 5b) Julho/2015-Julho/2017. 5c) Agosto/2017-Julho/2019. 5d) Agosto/2019-Agosto/2021

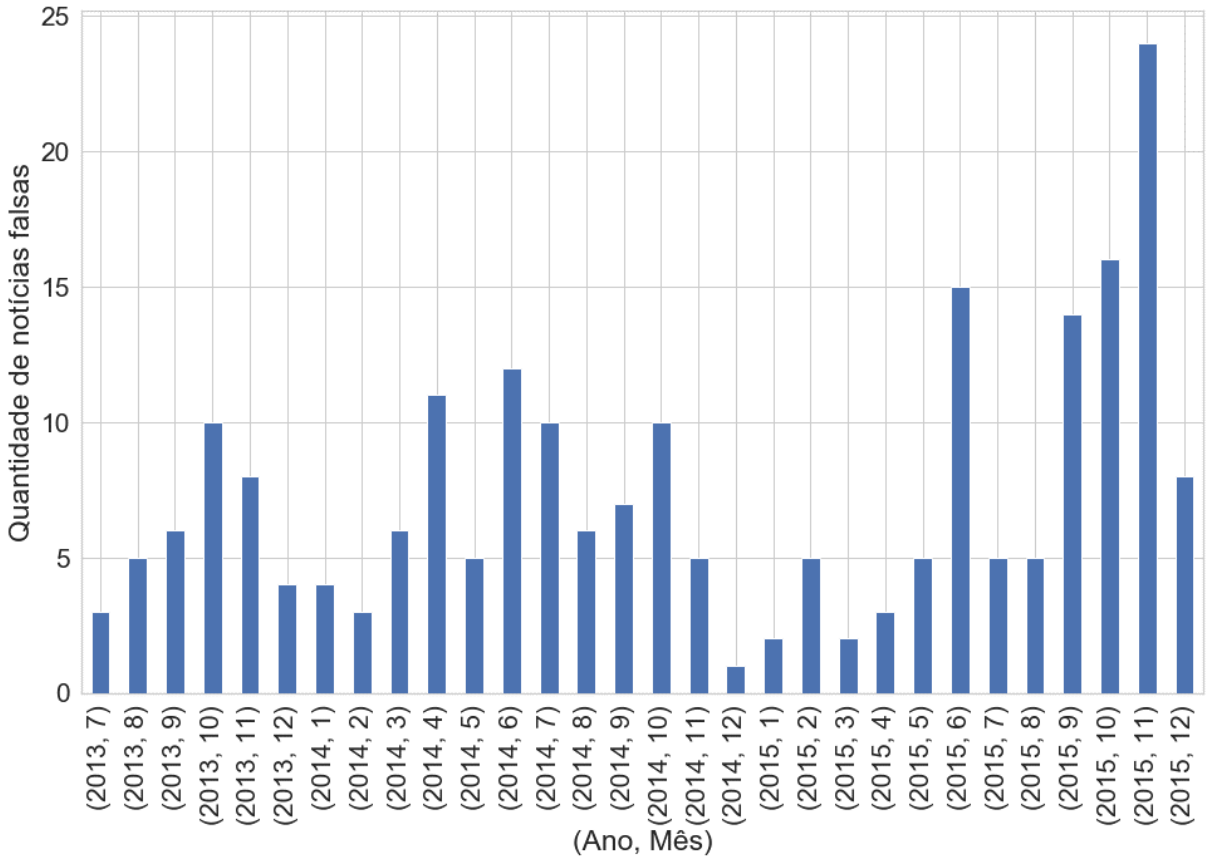


Fonte: Elaborado pelo autor (2023).

Considerando essa reflexão, optou-se ao longo do trabalho por realizar a análise temporal considerando as notícias de 2013, 2014 e 2015 como um único *dataframe*, e considerando as notícias dos anos posteriores, isto é, de 2016 até 2021, em *dataframes* individuais separados. Esse caminho foi escolhido visto que, considerando as notícias do triênio (2013-2015) em um único conjunto de observações, é possível obter uma quantidade de notícias razoável tanto para a análise, quanto para a posterior etapa de modelagem de tópicos, visando a construção de tópicos mais coesos. Adicionalmente, foi optado por considerar os anos posteriores em *dataframes* individuais, devido a vasta quantidade de notícias em cada ano, e devido ao fato de que a distribuição de acontecimentos sociais históricos desses períodos podem ser a chave para o entendimento da evolução das notícias falsas.

Após realizar a divisão, obtém-se uma observação mais específica relativa a cada ano, conforme pode ser observado no Gráfico 6. Em geral, o triênio (2013-2015) se caracteriza com um total de 220 notícias, e com estabilidade na maior parte dos meses, com exceção de alguns meses onde ocorrem picos de quantidades de notícias, em outras palavras, quantidades acima da média de notícias do respectivo ano. No geral, tem-se que a média de notícias do triênio é de 7,3 notícias por mês, sendo que o número mínimo de notícias falsas ocorre em dezembro de 2014 com apenas 1 notícia, e o número máximo de notícias falsas ocorre em novembro de 2015 com 24 notícias falsas.

Gráfico 6: Quantidade de notícias falsas do triênio ano/mês



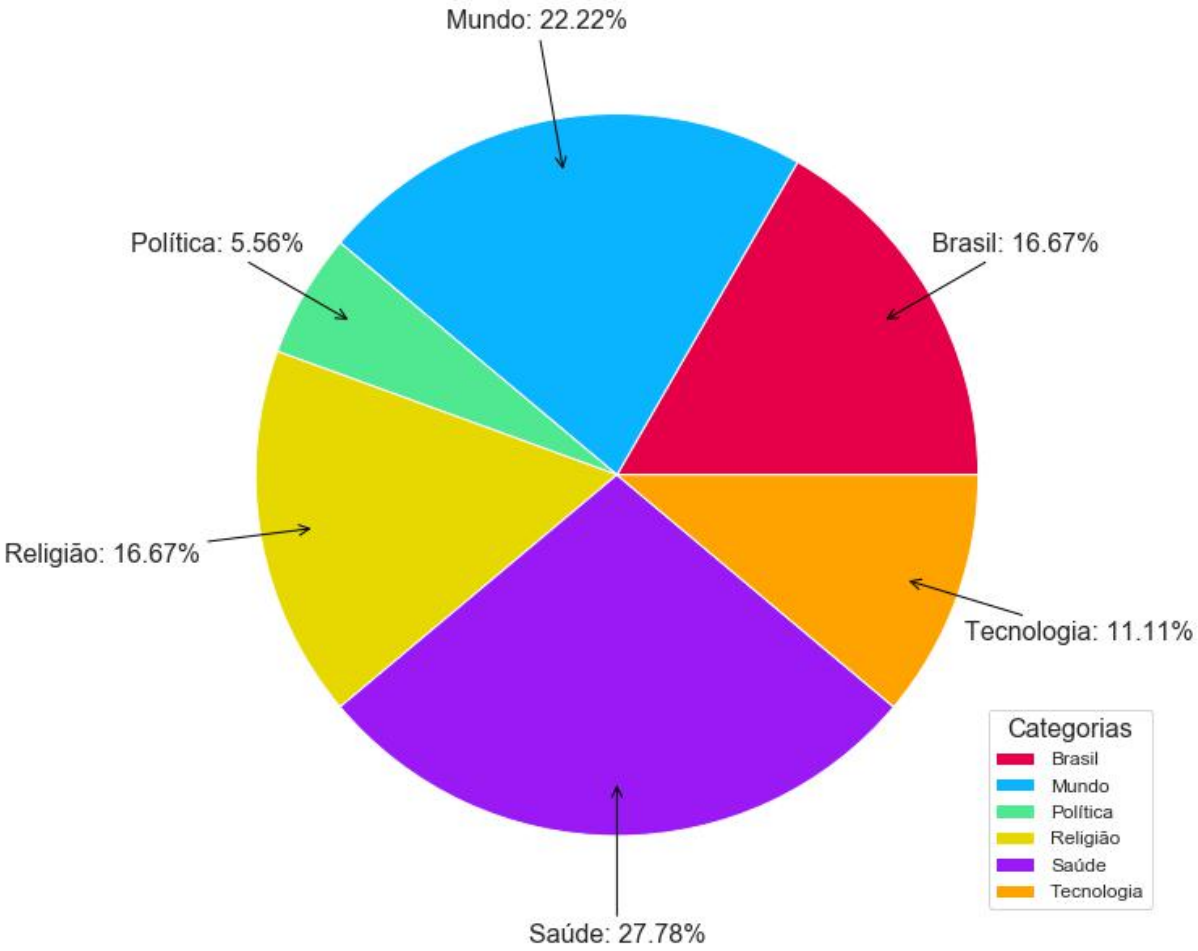
Fonte: Elaborado pelo autor (2023).

Como foram percebidos momentos de picos de notícias falsas ao longo de todo o *corpus*, foi realizado um estudo mais detalhado visando entender os principais assuntos tratados nesses momentos onde ocorrem maior quantidade de notícias, visto que trazem a possibilidade de compreender o que ocorre (assuntos com maior quantidade de *fake news*) nos momentos onde há maior quantidade de desinformação disseminada. Desse modo, é importante observar cada ano do *corpus* separadamente e identificar as principais categorias de notícias falsas disseminadas nos momentos de alta quantidade de notícias, isto é, nos momentos de pico. Adicionalmente, temos que uma análise mais elaborada sobre as categorias de notícias mais frequentes ao longo de todo o período temporal contido no *corpus* será detalhada na seção 5.2.

Considerando o ano de 2013, e analisando mais especificamente os meses de pico de notícias, constata-se que os mesmos ocorrem nos meses de outubro e novembro. Dessa forma, observando as categorias de notícias tratadas nesses meses, contabiliza-se um total de seis categorias. Além disso, é possível perceber que

a maior parte das notícias falsas estão relacionadas aos temas sobre saúde, assuntos internacionais, assuntos nacionais e religião como pode ser observado no Gráfico 7.

Gráfico 7: Proporção de cada categoria nos momentos de pico de 2013



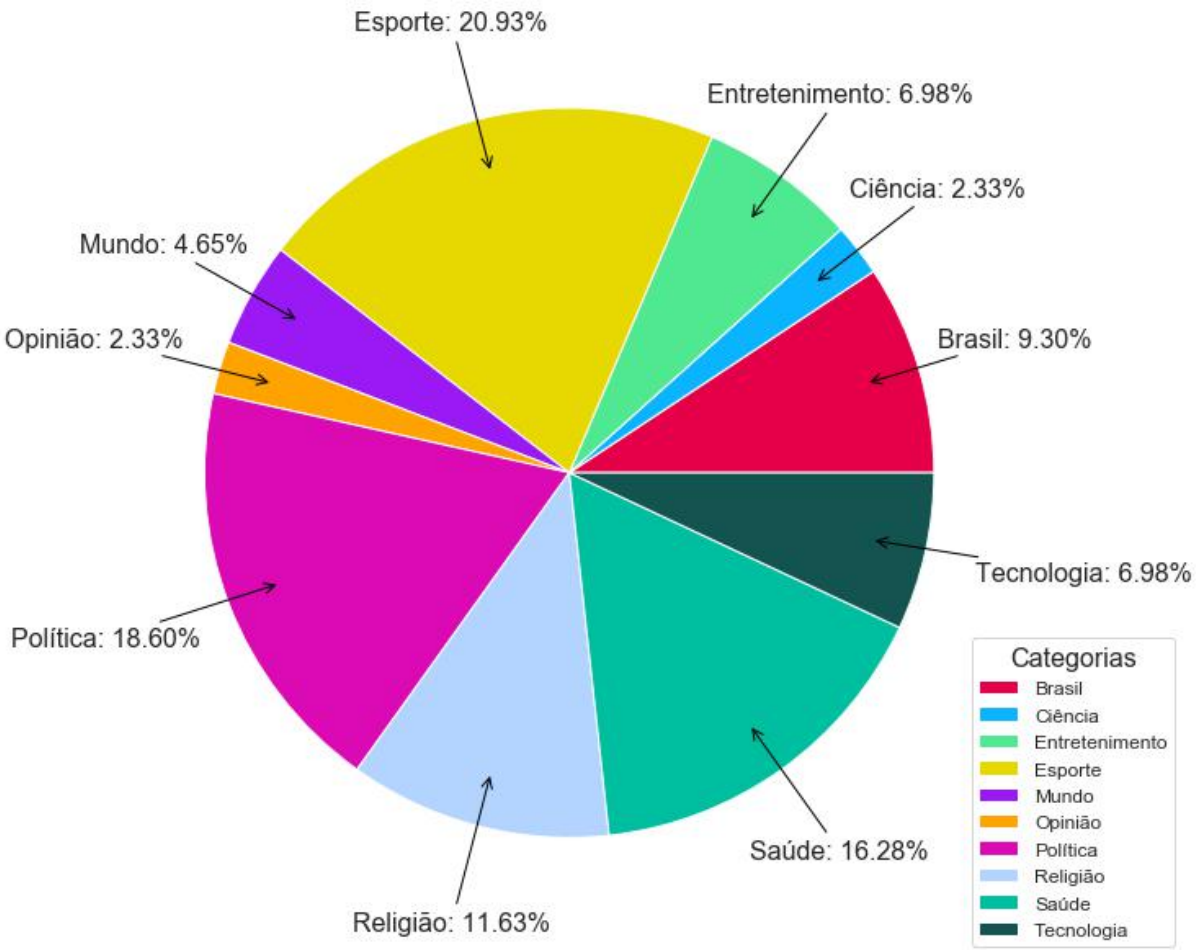
Fonte: Elaborado pelo autor (2023).

Como exemplo de assuntos tratados para cada uma das categorias mais frequentes nos momentos de pico do ano de 2013 listadas acima, podemos citar: "*Mito sobre consumo de alimentos de micro-ondas fazerem mal a saúde*", "*Vídeo falso de criatura escalando prédio na Rússia*", "*Boato sobre empresa Royal Canin fazer parceria com Instituto Royal*" e "*Texto falso sobre a Arca de Noé encontrada por chineses*", respectivamente.

No ano de 2014, temos que os picos de notícias ocorrem nos seguintes meses: abril, junho, julho e outubro. Consequentemente, observando as categorias das notícias, temos que as mesmas abrangem dez categorias distintas, conforme pode

ser visto no Gráfico 8. Nesse caso, temos que as categorias mais presentes nos meses de pico estão atreladas aos temas: esporte, política, saúde e religião.

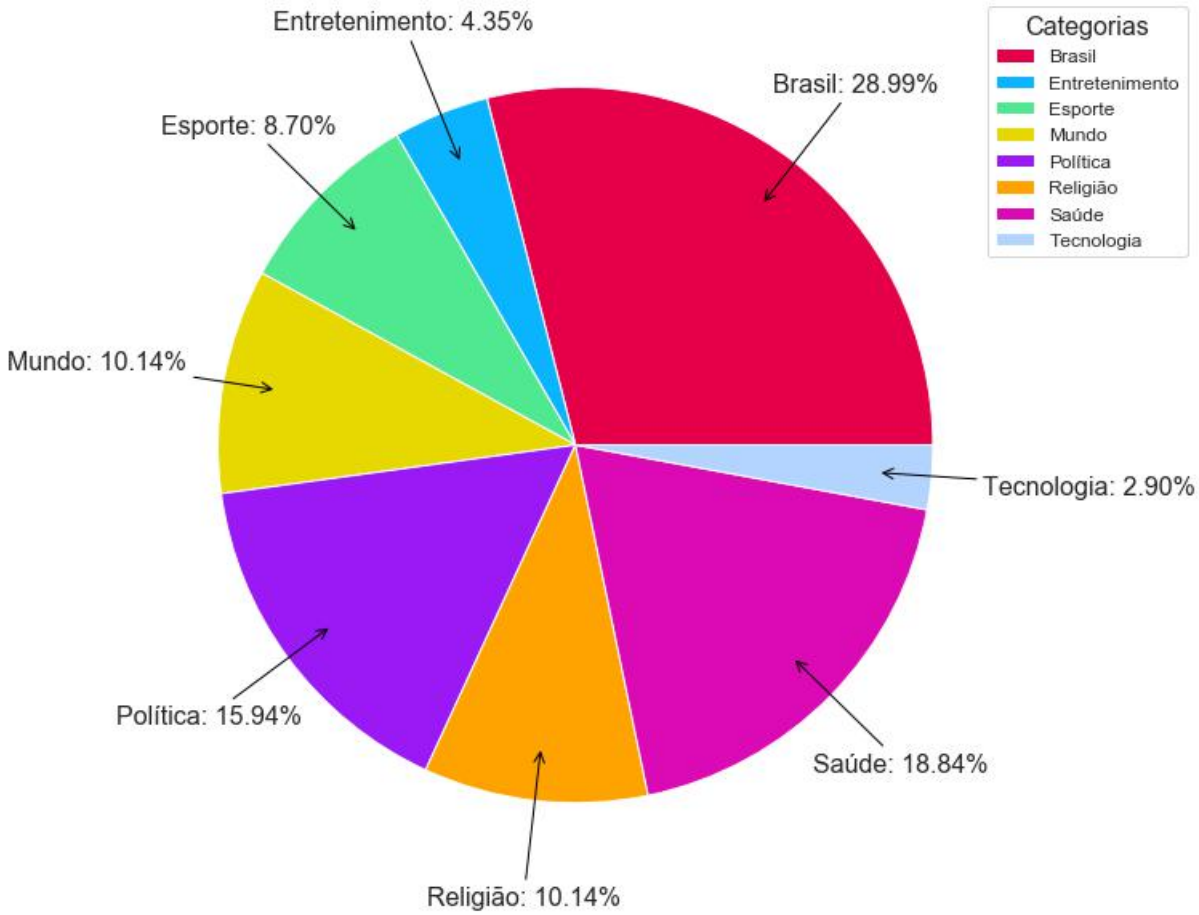
Gráfico 8: Proporção de cada categoria nos momentos de pico de 2014



Fonte: Elaborado pelo autor (2023).

Dessa forma, como exemplos de assuntos tratados em cada uma das categorias mais frequentes supracitadas, podemos elencar: "*Notícia falsa sobre queda do avião do clube de futebol Real Madrid*", "*Notícia falsa sobre câmara dos deputados aprovarem o fim do 13° salário*", "*Notícia falsa sobre o Ebola ser uma farsa criada pela Cruz Vermelha*" e "*Notícia falsa sobre o filme Malévola*", respectivamente. Portanto, considerando os momentos de pico do triênio, temos que no ano de 2015 ocorrem picos nos meses de junho, setembro, outubro e novembro. Logo, temos que os momentos de pico abrangem oito categorias distintas, conforme pode ser visto no Gráfico 9. Considerando as categorias mais frequentes nos momentos de pico desse ano, podemos citar: assuntos nacionais, saúde e política.

Gráfico 9: Proporção de cada categoria nos momentos de pico de 2015

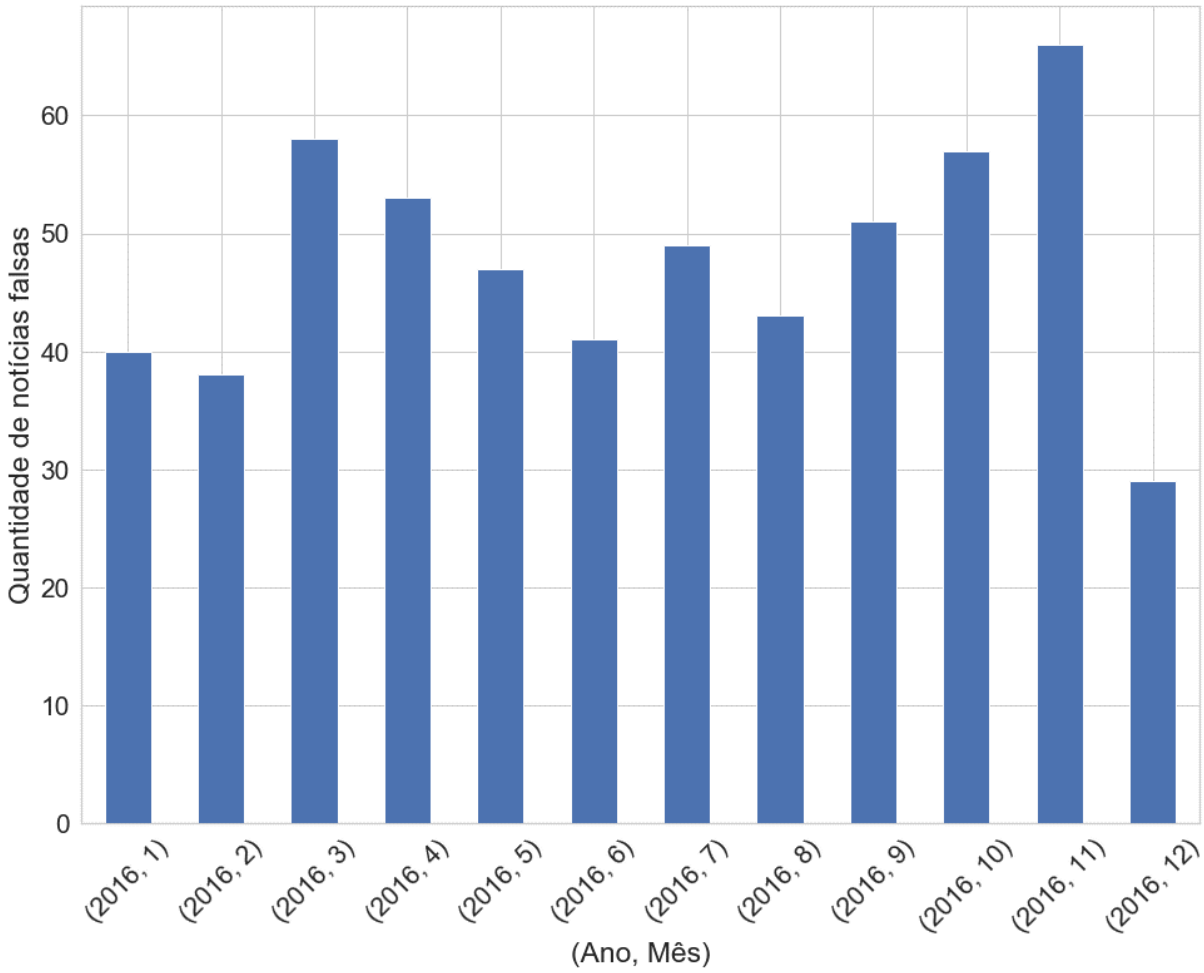


Fonte: Elaborado pelo autor (2023).

Como exemplo de assuntos atrelados as principais categorias citadas acima nos momentos de pico em 2015, podemos exemplificar com notícias, tais como: "*Boato sobre falta de nota fiscal garantir pedágio gratuito*", "*Notícia falsa sobre desenvolvimento de uma vacina contra o câncer*" e "*Notícia falsa sobre greve geral*", respectivamente.

O ano de 2016 possui um total de 572 notícias falsas no *dataset*, o que caracteriza um aumento de 160% na quantidade de notícias desse ano em relação ao triênio (2013-2015). Além disso, o ano de 2016 tem uma média de 47,6 notícias falsas por mês, com menor número de notícias falsas em dezembro de 2016 com 29 notícias falsas, e com o maior número de notícias falsas em novembro com 66 notícias falsas, conforme pode ser visto no Gráfico 10.

Gráfico 10: Quantidade de notícias falsas em 2016 ano/mês



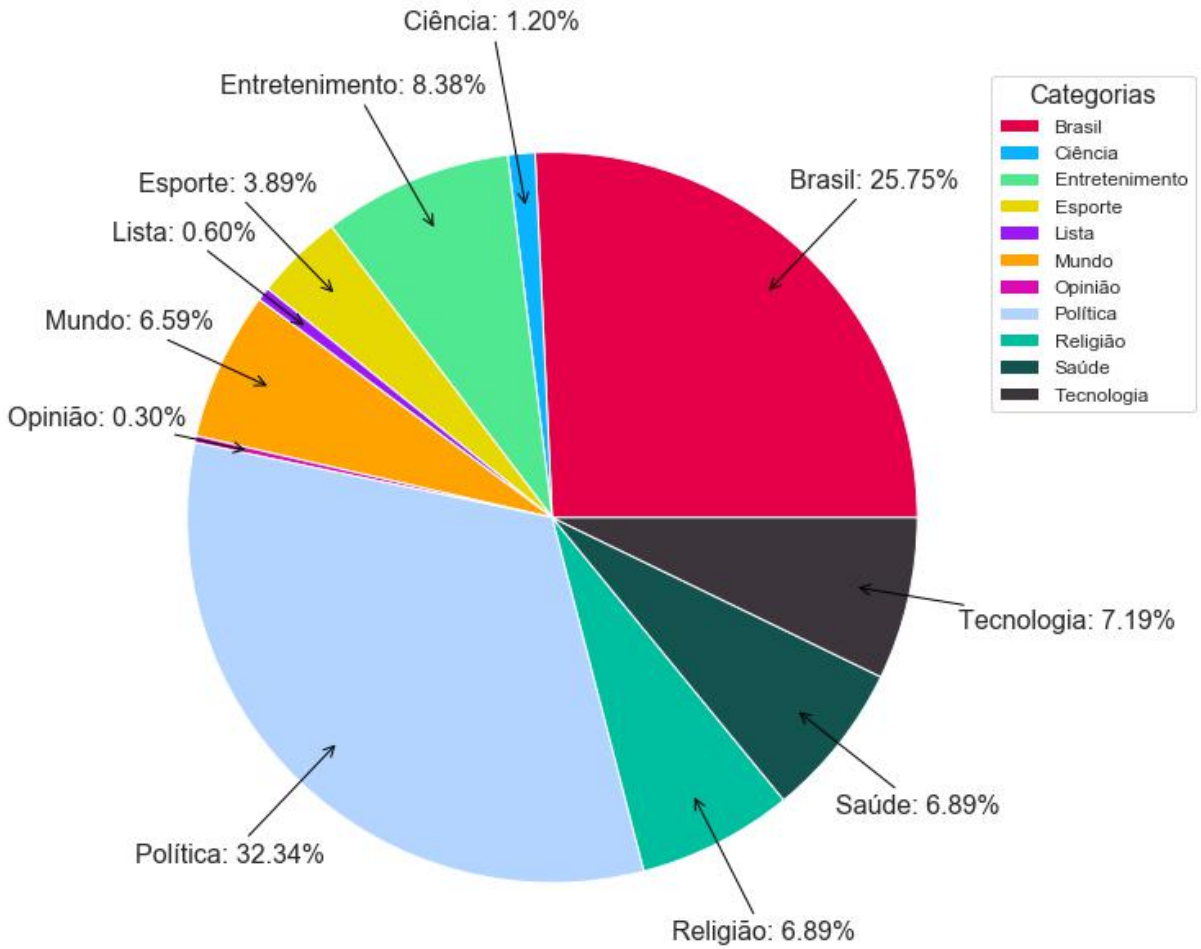
Fonte: Elaborado pelo autor (2023).

De maneira geral, o aumento de 160% pode ser entendido de forma mais especifica ao analisar as principais categorias de notícias falsas disseminadas no ano de 2016 nos meses com quantidade de notícias acima da média, isto é, principalmente nos meses de março, abril, julho, setembro, outubro e novembro. No Gráfico 11, compreende-se que os temas mais tratados nesses momentos estão relacionados a temas como política e assuntos nacionais. Além disso, ao observar as notícias desse período, verifica-se que muitas notícias estão ligadas aos acontecimentos políticos ocorridos no referido ano, tais como o processo de Impeachment[23]. Portanto, dessa forma é possível entender os principais motivos do aumento, ou seja, boa parte das

[23] Disponível em: https://www12.senado.leg.br/noticias/arquivos/2016/08/31/veja-a-sentenca-de-impeachment-contra-dilma-rousseff

notícias falsas desse período contemplam as principais questões do cenário político em alta na mídia tradicional, assim como boatos com temáticas mais diversificadas.

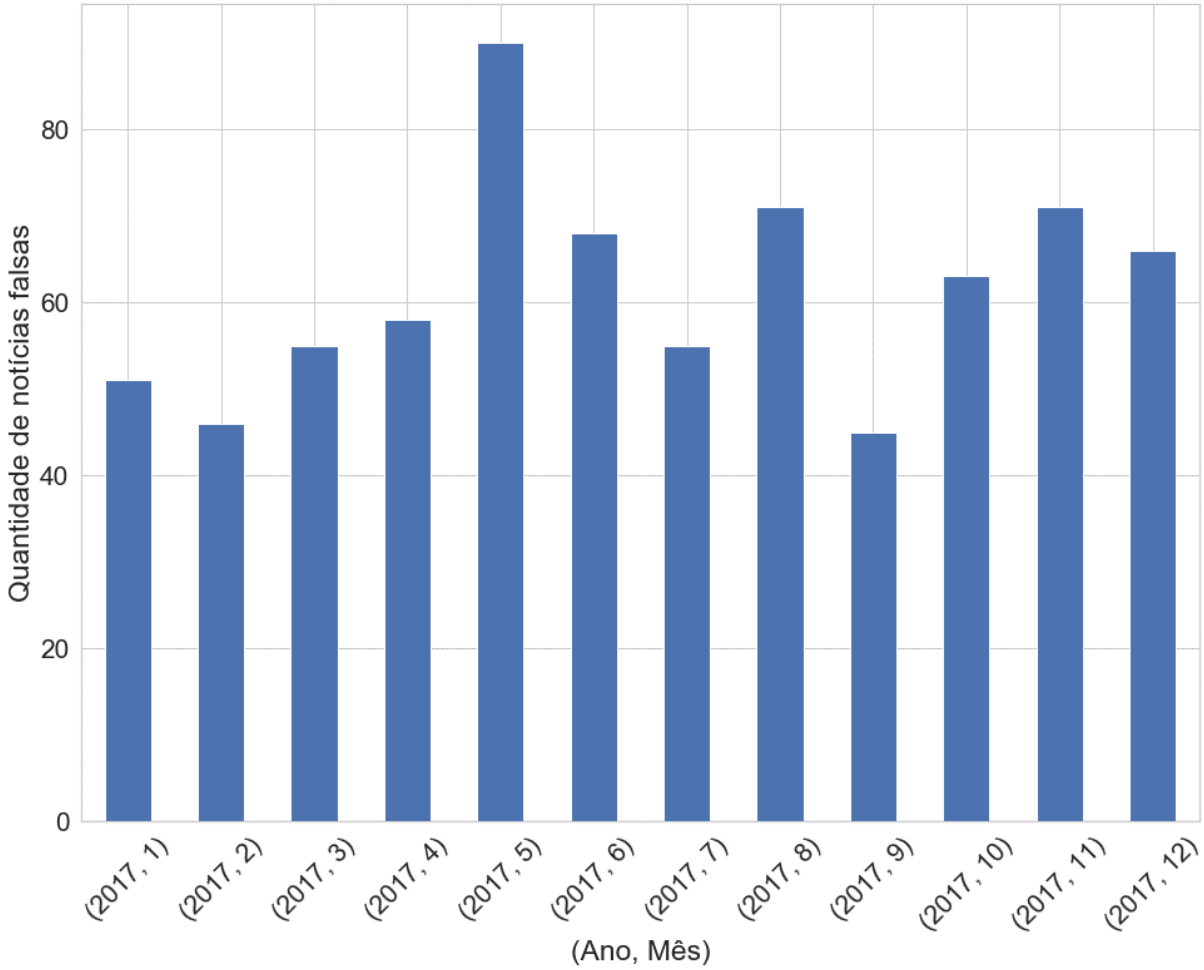
Gráfico 11: Proporção de cada categoria nos momentos de pico de 2016



Fonte: Elaborado pelo autor (2023).

O ano de 2017 tem um total de 739 notícias falsas pelo site “boatos.org”, o que caracteriza um aumento de 29,19% em relação ao ano de 2016. Além disso, o ano de 2017 possui uma média de 61,5 publicações de notícias falsas por mês, sendo o menor número de notícias registrado em setembro com 45 notícias, e o maior número de notícias registradas, em maio, com 90 notícias. No Gráfico 12, caracteriza-se a distribuição de notícias no ano de 2017.

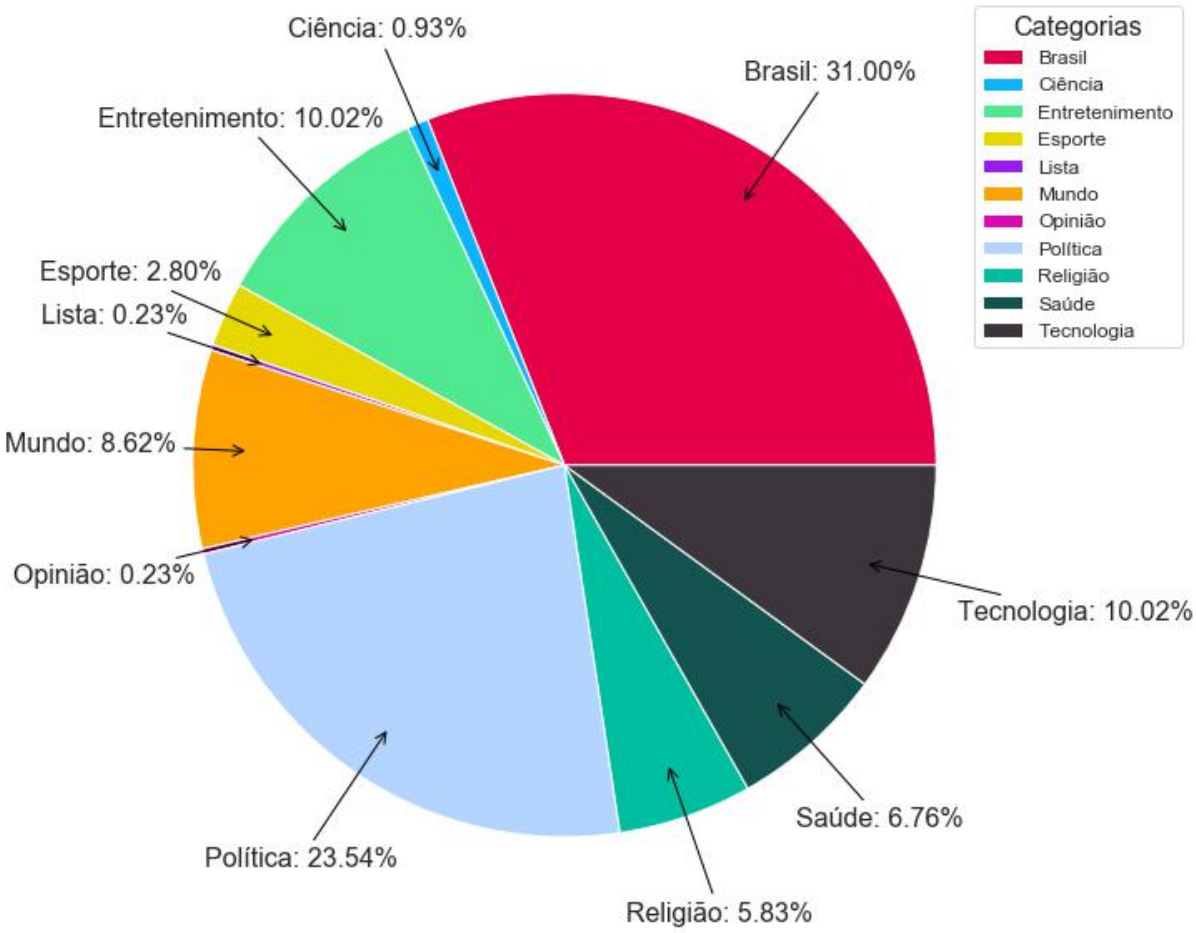
Gráfico 12: Quantidade de notícias falsas em 2017 ano/mês



Fonte: Elaborado pelo autor (2023).

Observando os meses com quantidade de notícias acima da média no ano de 2017, temos os seguintes meses: maio, junho, agosto, outubro, novembro, dezembro. Considerando as categorias mais frequentes nesses momentos, temos que assuntos nacionais e política são predominantes conforme pode ser visto no Gráfico 13. Analisando as notícias desse período de maneira mais específica, temos principalmente notícias falsas envolvendo nome de políticos e pessoas públicas, tais como: Lula, Bolsonaro, Sérgio Moro, Dilma e Michel Temer.

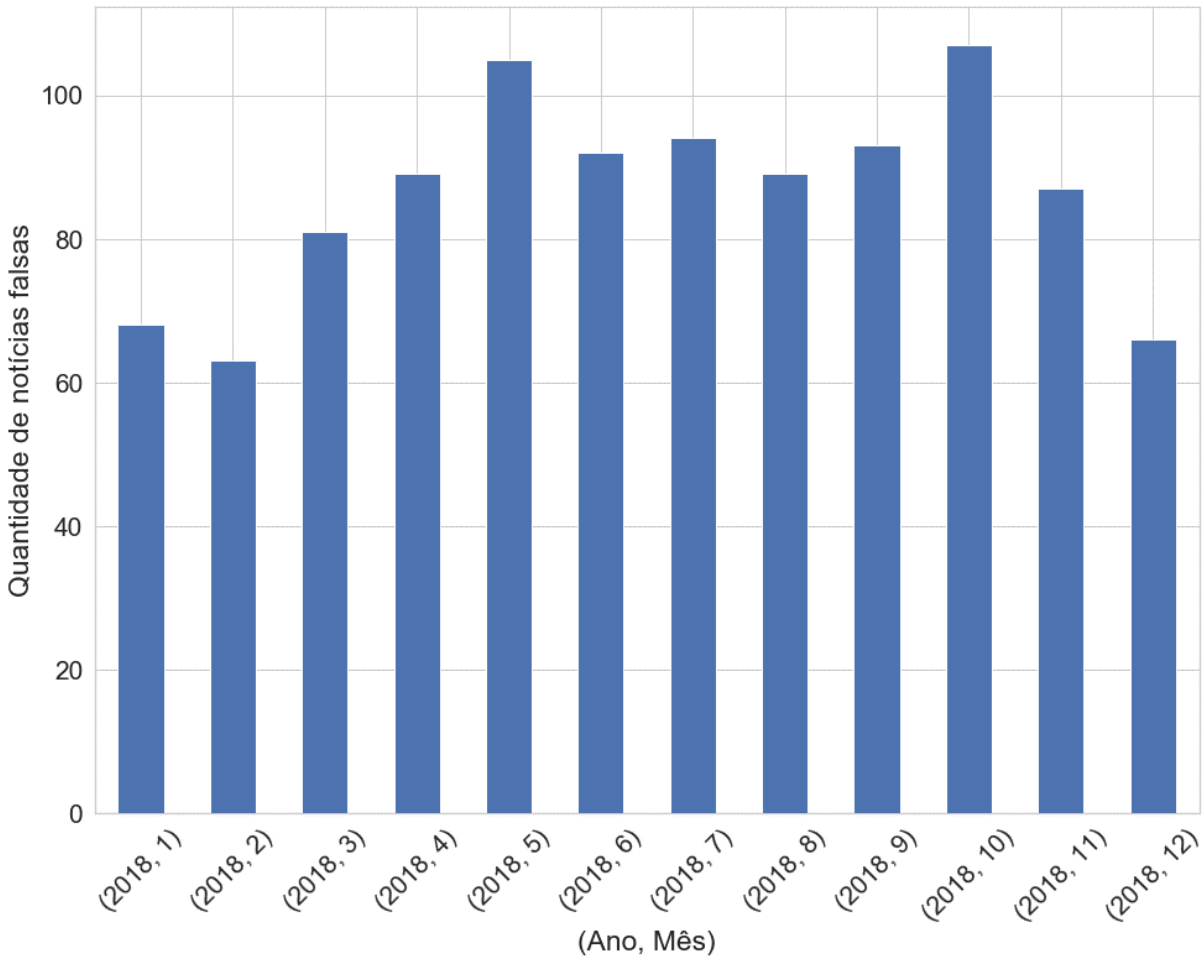
Gráfico 13: Proporção de cada categoria nos momentos de pico de 2017



Fonte: Elaborado pelo autor (2023).

O ano de 2018 tem um total de 1.034 notícias falsas no *dataset*, o que caracteriza um aumento de 39,91% em relação ao ano de 2017. Além disso, o ano de 2018 tem uma média de 86,1 notícias publicadas por mês, sendo o menor número de publicações realizadas em fevereiro, e o maior número de publicações realizadas em outubro. Como observação, temos que o principal acontecimento social ocorrido em 2018 está atrelado às eleições presidenciais, o que traz um indício de que esse aumento decorreu com o maior volume de notícias relacionadas à categoria de política, o que será corroborado nas seções posteriores do trabalho, onde foram realizadas análises voltadas para distribuição de categorias ao longo do tempo. No Gráfico 14, é possível observar a distribuição de notícias no ano de 2018.

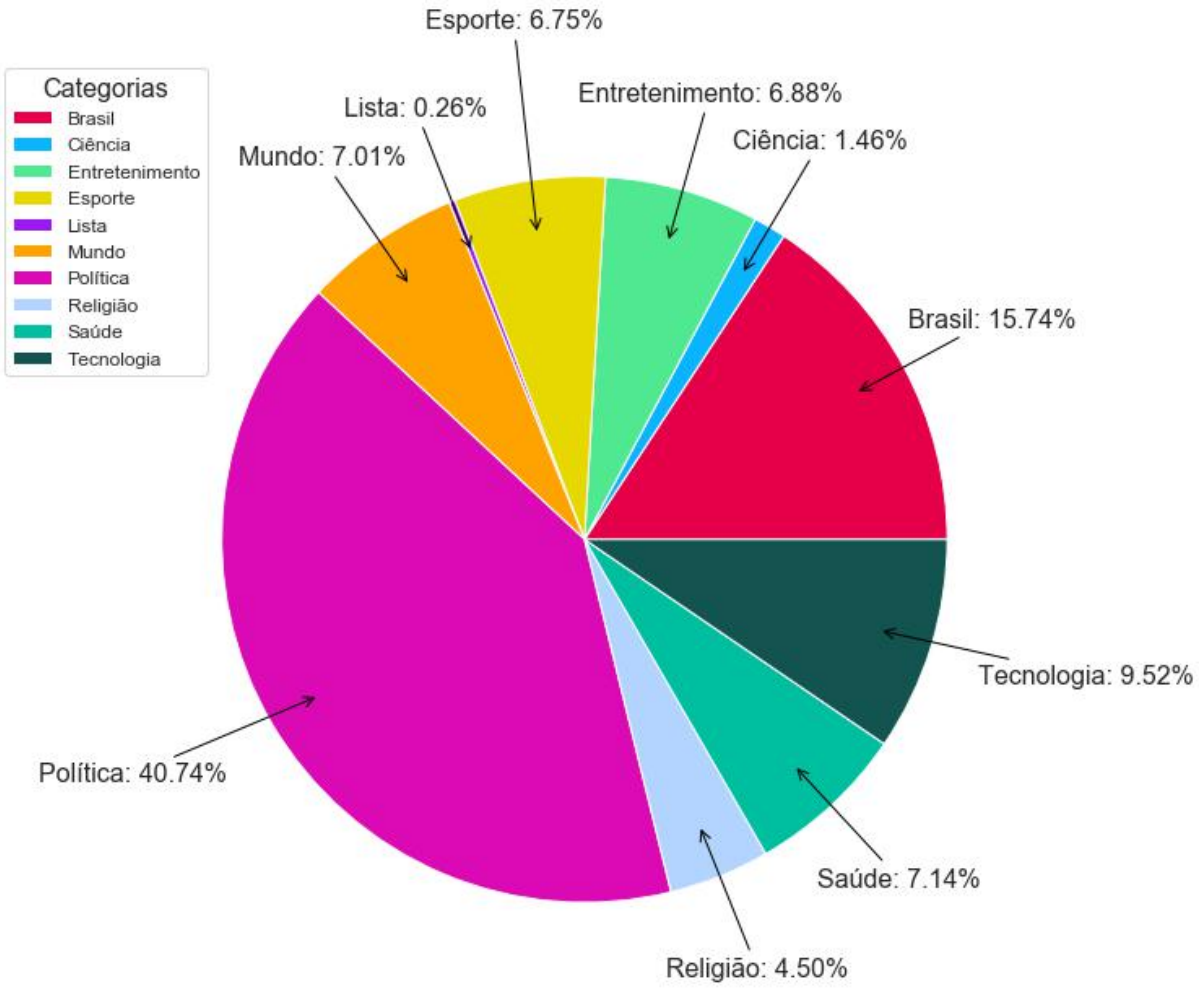
Gráfico 14: Quantidade de notícias falsas em 2018 ano/mês



Fonte: Elaborado pelo autor (2023).

Especificamente no ano de 2018, existe uma mudança considerável ao observar os meses com quantidades acima da média, isto é, abril, maio, junho, julho, agosto, setembro, outubro e novembro. Essa mudança está ligada ao fato de que o tema sobre política assume uma maior parcela de notícias em relação aos momentos de pico dos anos anteriores, conforme pode ser visto no Gráfico 15. Nesse caso, ao observar as notícias desse período foi constatado que a maior parte das mesmas tinham relação com as eleições de 2018, tanto envolvendo a disputa presidencial, quanto envolvendo os outros cargos políticos em disputa. Em geral, os nomes mais citados nas notícias são: Bolsonaro, Lula e Haddad. Portanto, temos a eleição como um possível motivo a ser considerado para o aumento das notícias falsas no ano de 2018 em relação ao ano de 2017, o que será corroborado nas seções posteriores do trabalho, onde foram realizadas análises voltadas para distribuição de categorias e entidades ao longo do período temporal estudado.

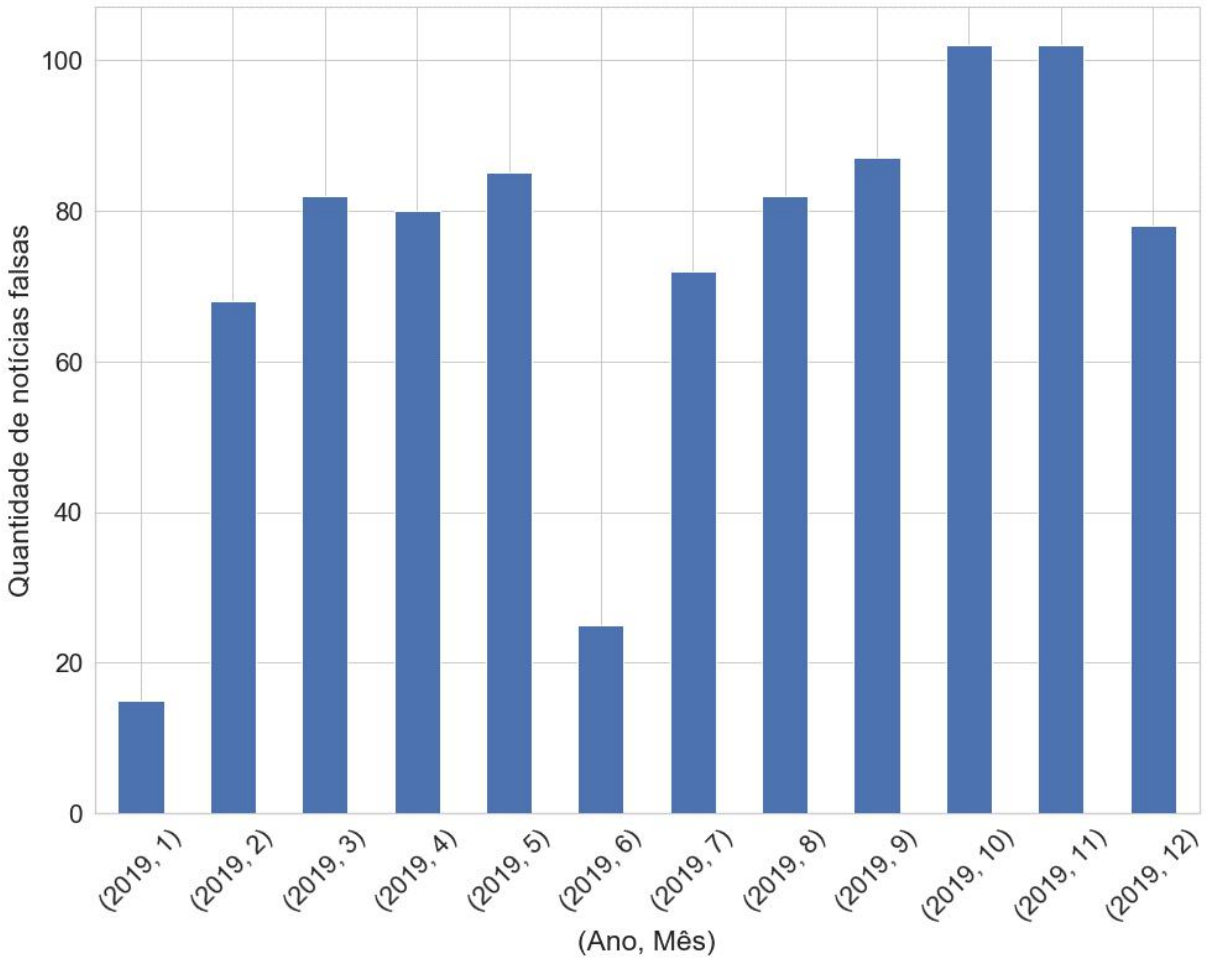
Gráfico 15: Proporção de cada categoria nos momentos de pico de 2018



Fonte: Elaborado pelo autor (2023).

O ano de 2019 possui um total de 878 notícias falsas, o que caracteriza uma diminuição de 15,08% em relação ao ano de 2018. Além disso, o ano de 2019 tem uma média mensal de 73,1 notícias. Sendo 15 notícias no mês de janeiro, ou seja, o mês com a menor quantidade de notícias falsas, e os meses de outubro e novembro são os meses com maior quantidade de notícias falsas (ambos os meses com 102 notícias cada). No Gráfico 16, é possível observar a distribuição de notícias no ano de 2019.

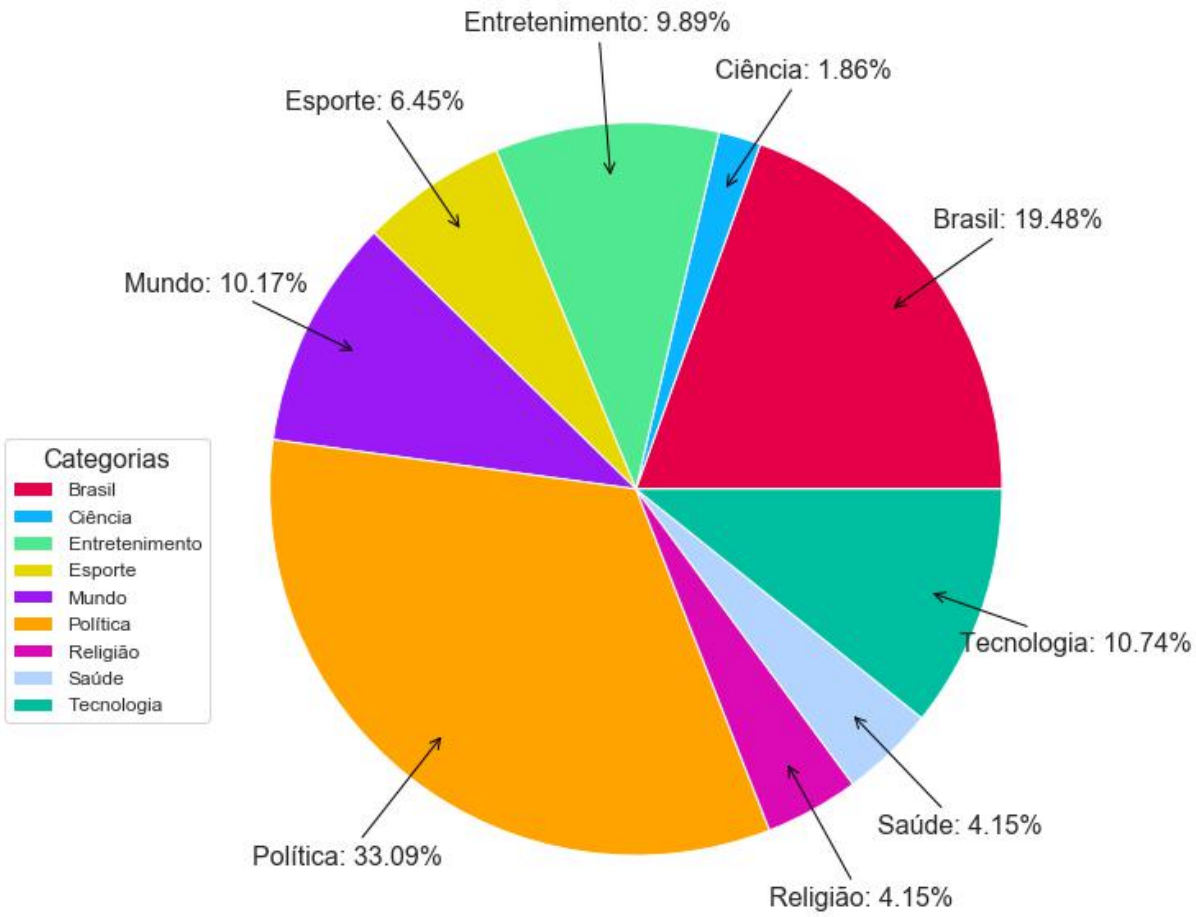
Gráfico 16: Quantidade de notícias falsas em 2019 ano/mês



Fonte: Elaborado pelo autor (2023).

Analisando o ano de 2019, temos que os meses onde há disseminação acima da média são: março, abril, maio, agosto, setembro, outubro, novembro e dezembro. Dessa forma, ao observar esses meses, novamente temos política como tema mais frequente, porém há uma diminuição em relação aos meses de pico do ano anterior, conforme pode ser visto no Gráfico 17. Ao observar as principais notícias relativas a política nesse ano, constata-se ainda grande presença de nomes como Bolsonaro e Lula. Porém, a diminuição da quantidade de notícias em relação ao ano anterior é justificada ao observar o fato de que as notícias relacionadas a política não tem tanta relação com o tema sobre eleição, que foi amplamente utilizado no ano de 2018 para veiculação de notícias falsas. Portanto, apesar do tema política continuar sendo o mais tratado, tem-se que a falta de um acontecimento histórico de grandes proporções tal como a eleição, implica em um menor número de notícias falsas.

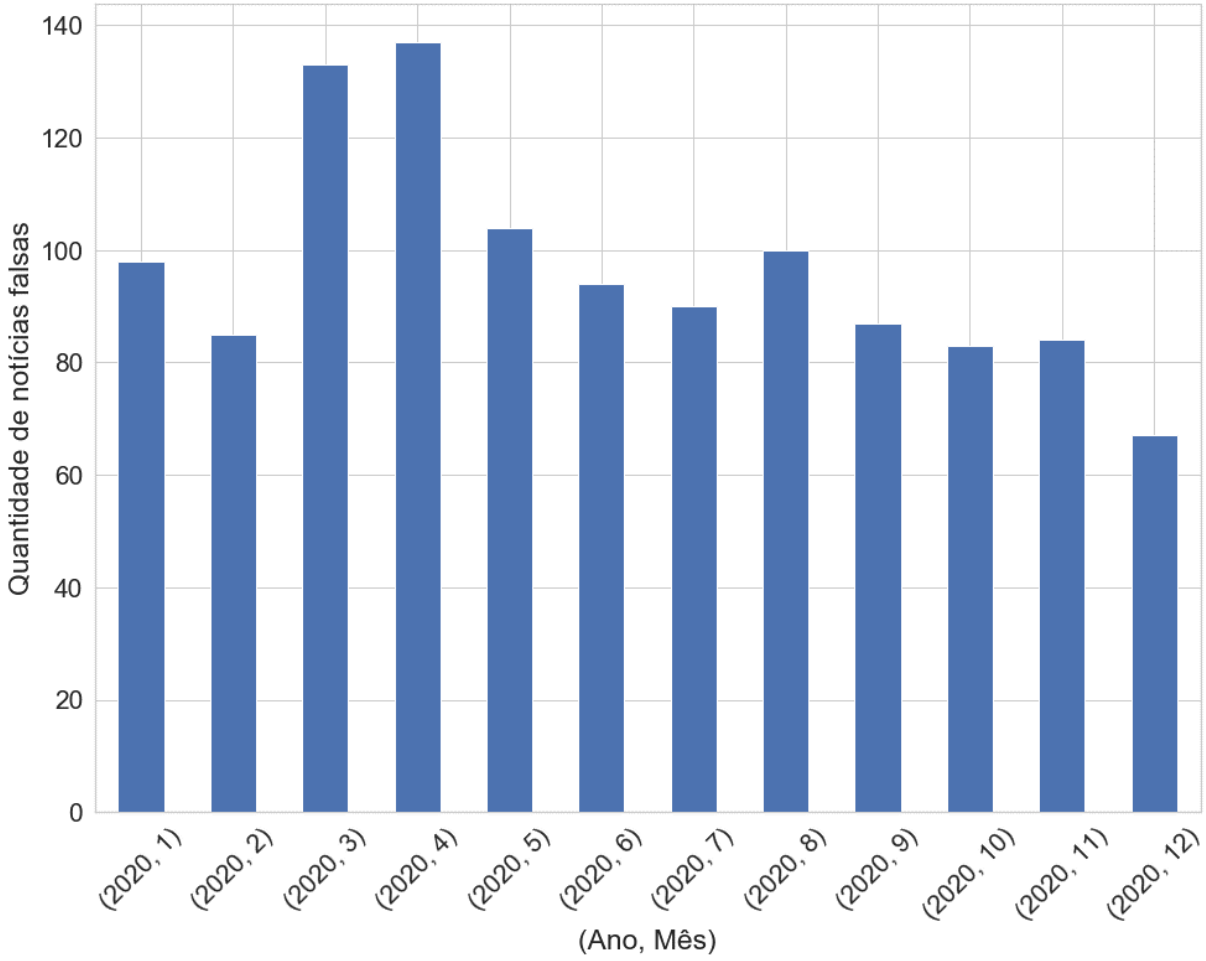
Gráfico 17: Proporção de cada categoria nos momentos de pico de 2019



Fonte: Elaborado pelo autor (2023).

O ano de 2020 possui um total de 1.162 notícias falsas no *dataset*, o que caracteriza um aumento de 32,34% em relação ao ano de 2019. Além disso, o ano de 2020 tem uma média de 96,8 notícias falsas mensalmente. Sendo o menor número de notícias falsas registradas em dezembro, com 67 notícias falsas. Além disso, tem-se que o maior número de notícias falsas ocorreu em abril com 137 notícias falsas. Como observação, temos que o principal acontecimento histórico ocorrido em 2020 se relaciona com a ocorrência da pandemia de COVID-19. No Gráfico 18, se consolida a caracterização da distribuição de notícias no ano de 2020, onde constata-se estabilidade em altas quantidades de publicações de desinformações na maior parte dos meses.

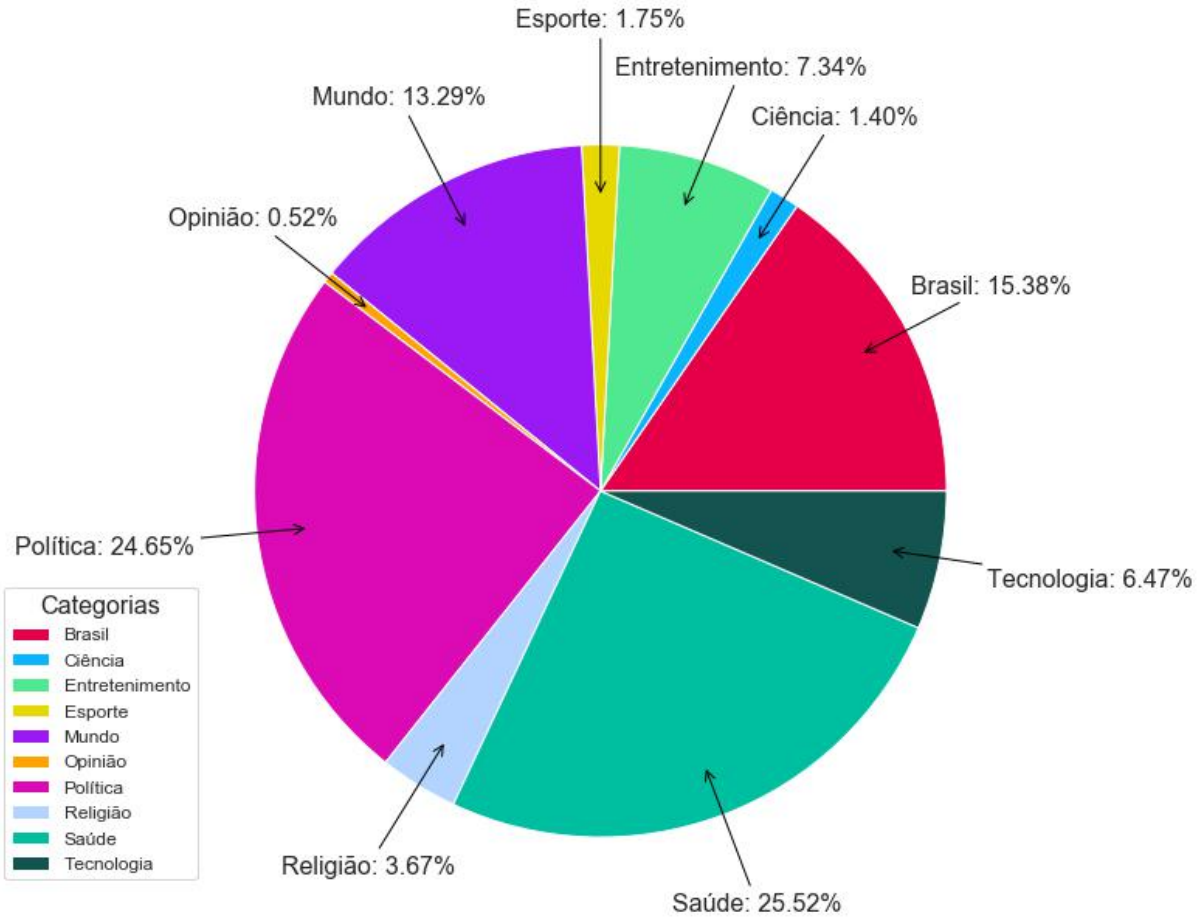
Gráfico 18: Quantidade de notícias falsas em 2020 ano/mês



Fonte: Elaborado pelo autor (2023).

Analisando o ano de 2020, constata-se um possível motivo de aumento da quantidade de notícias falsas em relação ao ano anterior se observarmos os meses de pico, isto é, janeiro, março, abril, maio e agosto. Considerando os meses citados, torna-se perceptível uma mudança de padrão expressiva em relação aos anos anteriores, ou seja, o tema saúde passa a ser o destaque das notícias, seguido pelo tema política, conforme pode ser visto no Gráfico 19. Considerando esse fato, ao ler as principais notícias nesse período foi identificado um número considerável de notícias sobre a COVID-19, incluindo boatos sobre efeitos da doença, falsos métodos de tratamento da doença, entre outros. Com isso, é possível identificar que a pandemia pode ser um dos principais motivos para o aumento de notícias falsas, o que será corroborado nas seções posteriores, se de fato a categoria relacionada a Saúde for a mais frequente ao analisar todas as notícias do referido ano, visto que nessa primeira análise o foco está associado ao período de pico de notícias.

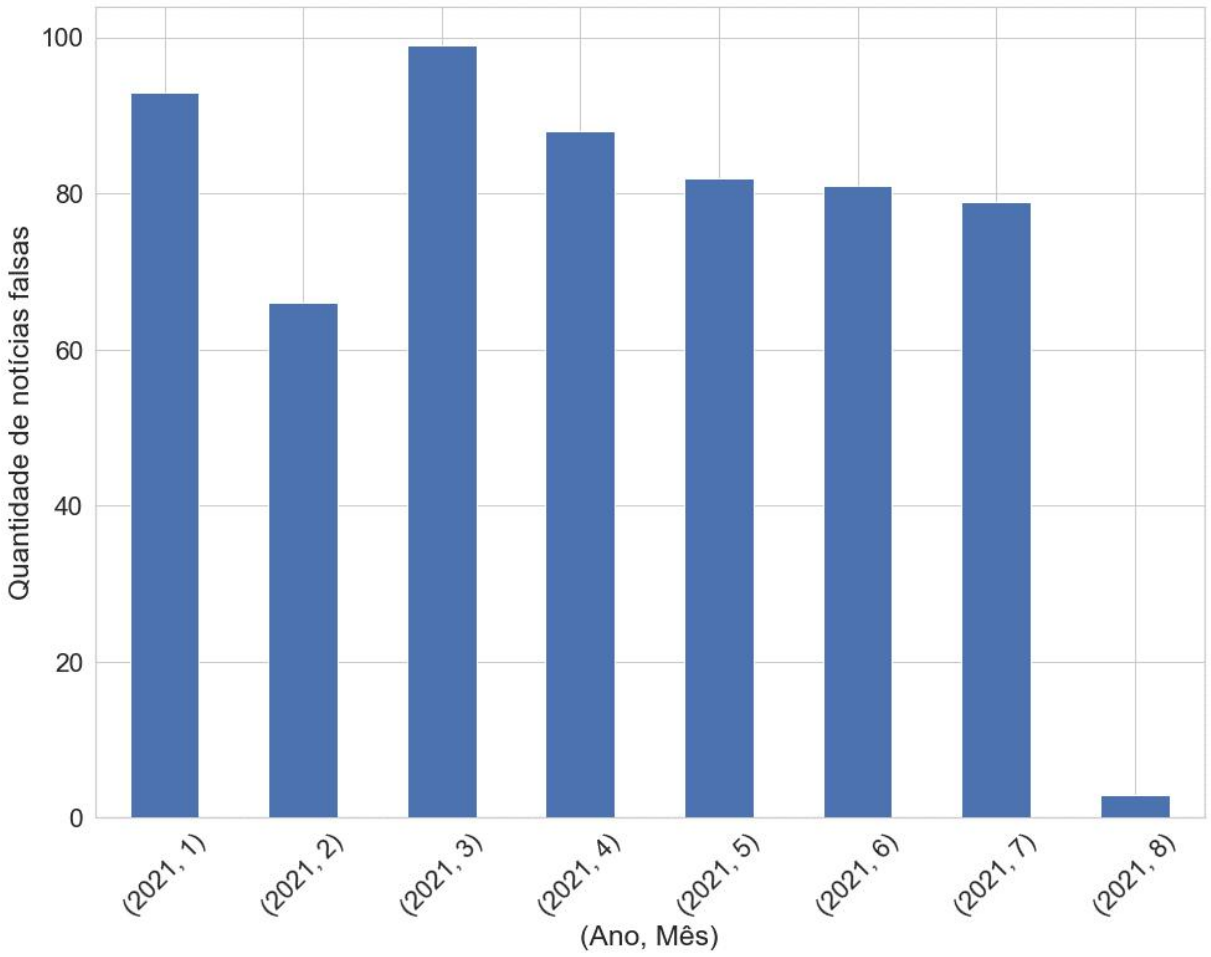
Gráfico 19: Proporção de cada categoria nos momentos de pico de 2020



Fonte: Elaborado pelo autor (2023).

O ano de 2021 possui um total de 591 notícias falsas no *dataset*, o que caracteriza uma diminuição de 49,13% em relação ao ano de 2020. Além disso, o ano de 2021 tem uma média de 73,8 notícias falsas mensalmente. Sendo o menor registro feito em agosto com 3 notícias, e o maior registro feito em março com 99 notícias. Como observação, temos o fato de que a construção do *corpus* Fakepedia utilizado como base neste trabalho foi realizada utilizando coletas até o mês de agosto, sendo coletadas apenas 3 notícias em agosto, o que ajuda a explicar o fato da diminuição de notícias. No Gráfico 20, caracteriza-se a distribuição de notícias no ano de 2021.

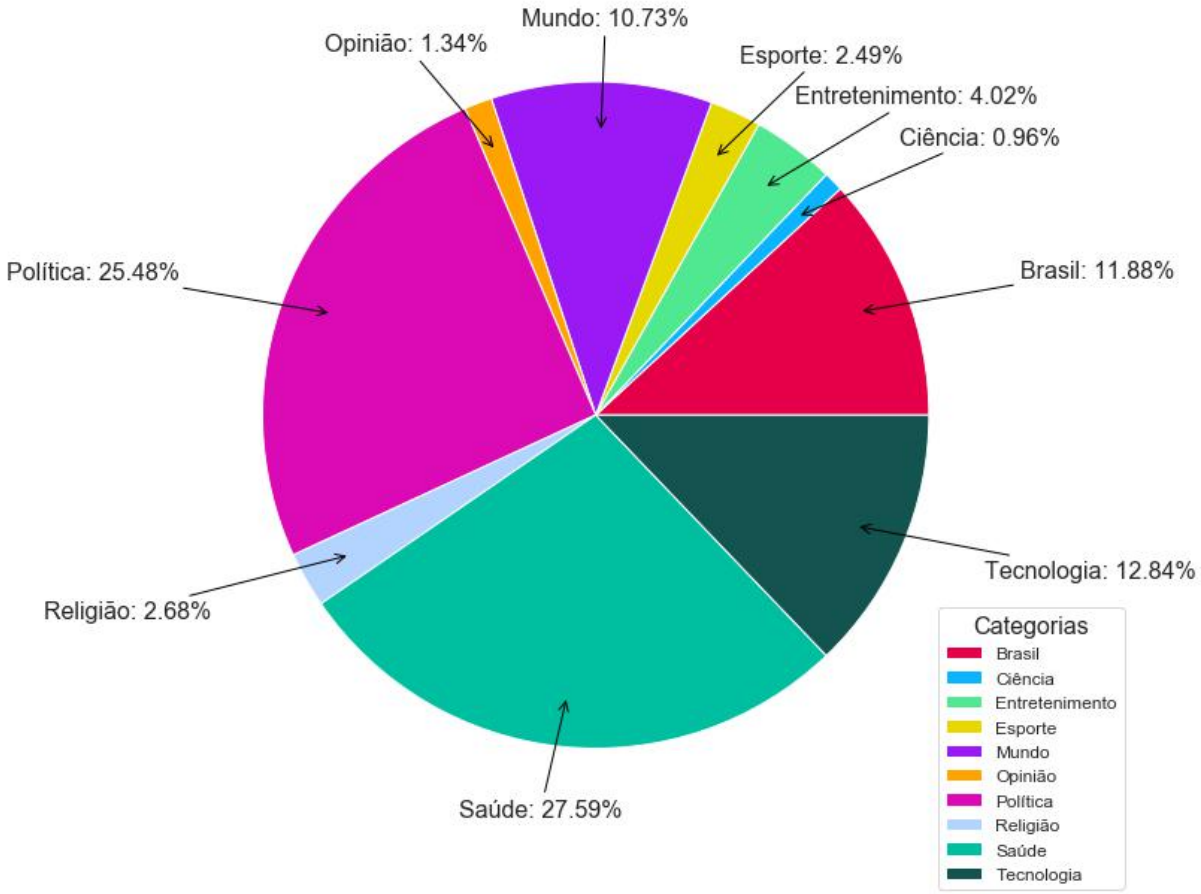
Gráfico 20: Quantidade de notícias falsas em 2021 ano/mês



Fonte: Elaborado pelo autor (2023).

Observando o ano de 2021, temos que os meses acima da média são: janeiro, março, abril, maio, junho e julho. Além disso, ao analisar, observa-se que as categorias de saúde e política aumentam discretamente em relação ao ano anterior conforme mostra o Gráfico 21. Porém, ao ler as principais notícias foi identificado um aumento no número de notícias que envolvem o tema sobre COVID-19 tendo relação com medidas políticas adotadas em relação a pandemia, e também relacionadas as vacinas contra a COVID-19 (falsos efeitos colaterais, etc). Porém, há uma diminuição discreta no número de notícias com informações falsas sobre questões intrínsecas a doença, isto é, boatos sobre efeitos da doença, entre outros. Portanto, devido a esse fato é possível compreender um dos possíveis motivos para a diminuição da quantidade de notícias falsas.

Gráfico 21: Proporção de cada categoria nos momentos de pico de 2021



Fonte: Elaborado pelo autor (2023).

Com esta caracterização temporal, é possível relacionar alguns dos números obtidos com os acontecimentos históricos ocorridos nos períodos supracitados. Estas correlações foram corroboradas nas seções a seguir onde as distribuições das categorias das notícias, das entidades nomeadas presentes nas notícias e das entidades nomeadas entre categorias de notícias ao longo do tempo foram abordadas com mais detalhes.

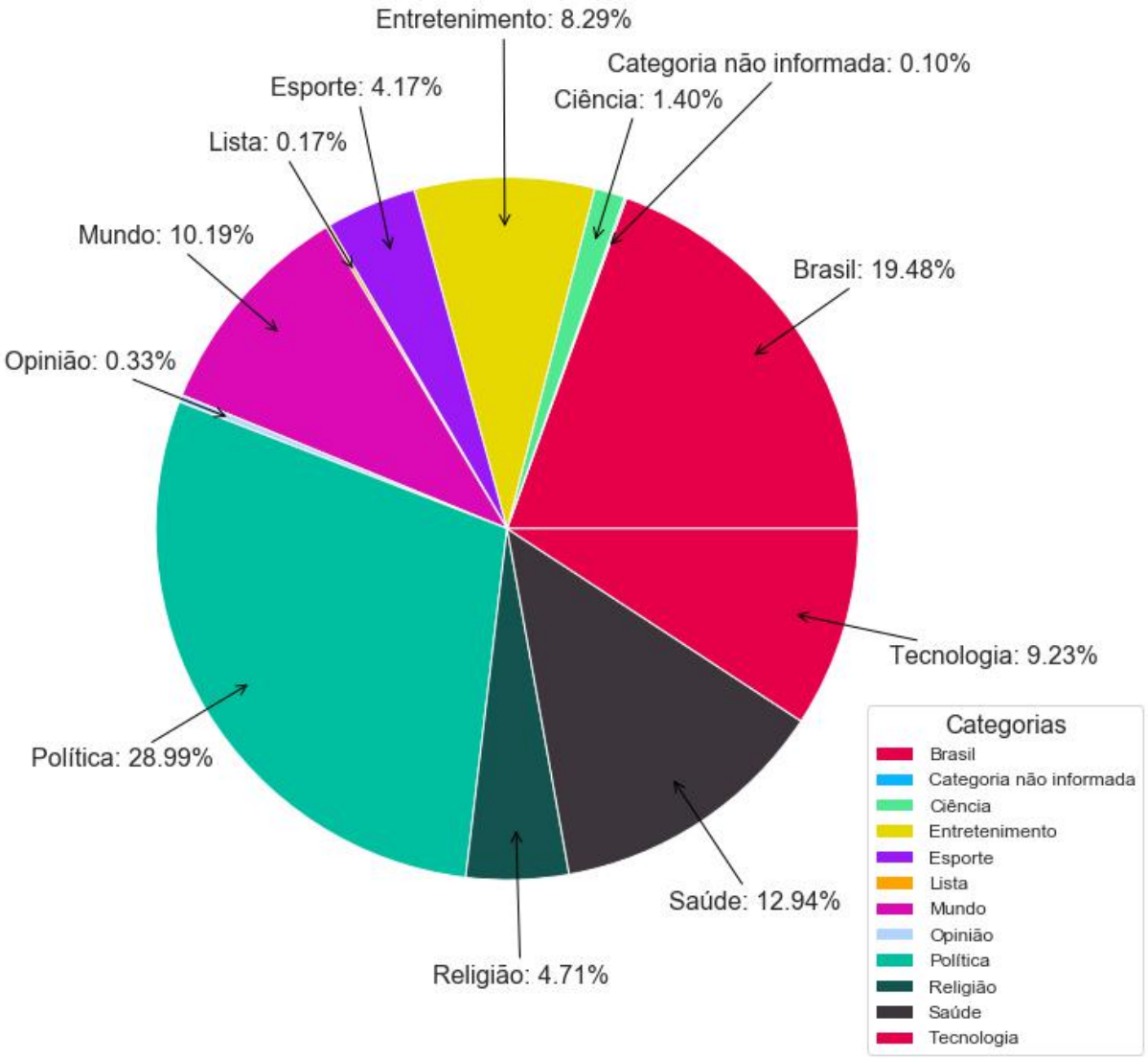
5.2 ANÁLISE TEMPORAL DAS CATEGORIAS DAS NOTÍCIAS

Após a análise temporal relativa à distribuição da quantidade de notícias falsas publicadas e análise relativa as categorias das notícias falsas presentes nos momentos de pico do *corpus*, realizou-se a análise voltada para a distribuição temporal, observando as categorias referentes às notícias falsas, de forma a viabilizar neste trabalho a identificação dos temas mais tratados em notícias falsas ao longo do tempo. No decorrer desta seção são realizadas inferências por meio da análise com

base nos dados relativos à frequência das categorias das notícias. Consequentemente, a maior parte dessas inferências foram corroboradas com mais detalhes nas próximas seções, onde foram realizadas análises temporais relativas à frequência de entidades nomeadas presentes nas notícias, bem como análises temporais relativas à frequência de entidades nomeadas presentes em cada categoria de notícia.

Considerando todos os anos do *corpus*, compreende-se no Gráfico 22, que a maior parte das notícias tem relação com política e assuntos nacionais. Além disso, torna-se perceptível que o tema saúde é o terceiro tema mais presente no *corpus*, e isso ocorre principalmente devido a pandemia de COVID-19 nos anos de 2020 e 2021, onde o número de notícias relacionadas a saúde aumentou de forma significativa, e além disso temos que nos anos anteriores a quantidade de notícias relacionadas a saúde era consideravelmente menor se comparada em relação aos anos vigentes da pandemia.

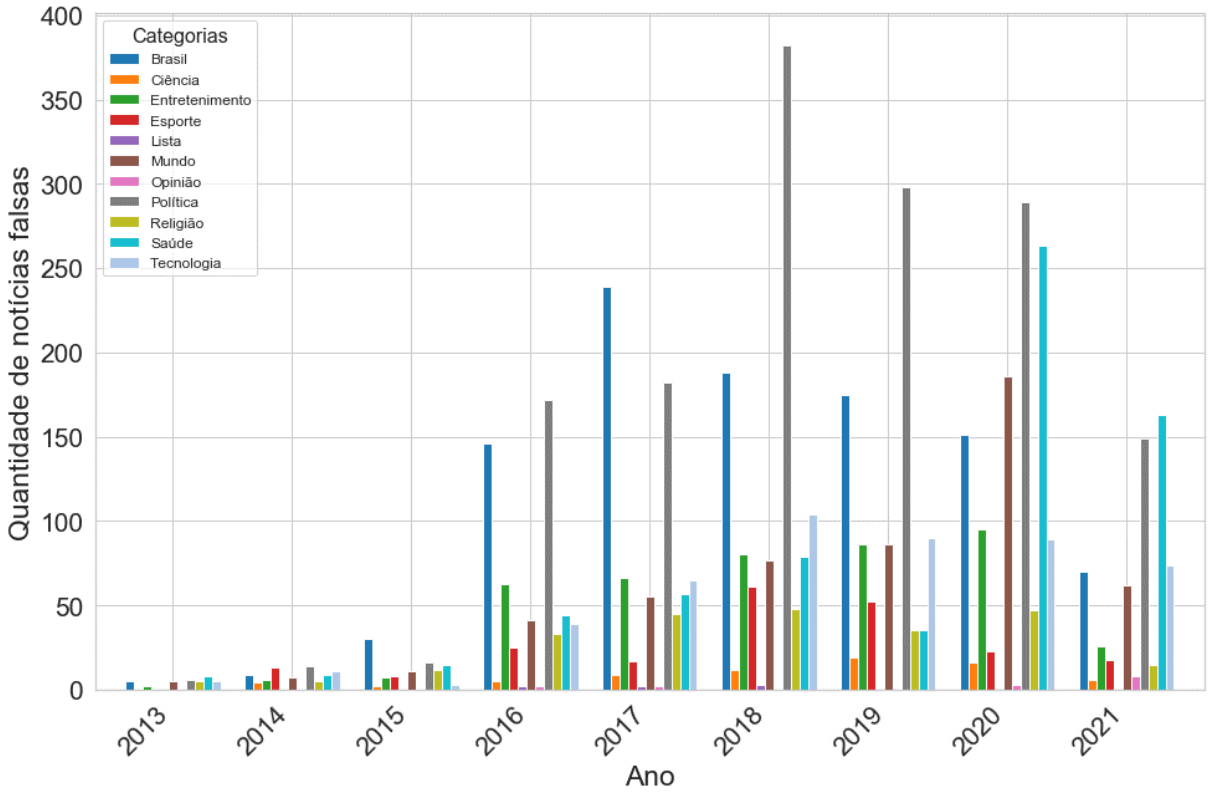
Gráfico 22: Proporção de cada categoria em todos os anos do *corpus*



Fonte: Elaborado pelo autor (2023).

A princípio realizou-se uma amostragem geral do *corpus* considerando todos os anos de forma a identificar a categorização geral das notícias. Podemos observar no Gráfico 23, a distribuição geral referente às categorias das notícias falsas presentes em todo o *corpus*. Portanto, percebe-se claramente o avanço quantitativo de notícias, bem como o aumento da diferença entre as categorias mais frequentes em relação as outras categorias. Além disso, é fácil observar que o maior pico de uma determinada categoria em todo o *corpus* Fakepedia ocorre com a categoria política em 2018, sendo esse pico associado as notícias falsas disseminadas no ano das eleições.

Gráfico 23: Quantidade de notícias falsas por categoria ao longo de todos os anos do *corpus*



Fonte: Elaborado pelo autor (2023).

Considerando o cenário geral, temos as categorias mais frequentes em cada ano do *corpus*. Em 2013, a categoria de notícias mais frequente no *dataset* refere-se ao tema saúde com um total de 8 notícias falsas, onde maior parte das notícias trazem boatos sobre variados assuntos relacionados a saúde, não havendo nenhum assunto mais frequente entre as notícias. Em 2014, a categoria de notícias mais frequente refere-se ao tema política com um total de 14 notícias falsas, entre as notícias há presença considerável de notícias falsas relacionadas a eleição[24] de 2014, incluindo notícias falsas envolvendo o nome do candidato presidenciável Eduardo Campos. Após isso, em 2015, a categoria de notícias com maior número de ocorrências refere-se ao tema sobre assuntos nacionais denominado Brasil no *corpus* com um total de 30 notícias falsas, com a presença de notícias falsas versando sobre diferentes assuntos, isto é, não foi encontrado nenhum assunto predominante. Adicionalmente, considerando a média do triênio (2013-2015), pode-se afirmar que a categoria mais frequente tem relação ao tema sobre assuntos nacionais. Posteriormente, em 2016, a

[24] Disponível em: https://www.tse.jus.br/eleicoes/eleicoes-anteriores/eleicoes-2014

categoria de notícias mais frequente refere-se ao tema política com um total de 172 notícias falsas, fato que pode estar atrelado aos acontecimentos políticos ocorridos no Brasil em 2016, tais como: *impeachment*, operação lava jato[25], eleições municipais[26], etc. Em 2017, a categoria de notícias com maior presença no *dataset* refere-se ao tema sobre assuntos nacionais com um total de 239 notícias falsas. Após isso, em 2018, a categoria de notícias mais frequente refere-se ao tema política, com um total de 382 notícias falsas, o que equivale a mais de uma notícia desmentida por dia durante o ano de 2018. Esse fato pode ter estreita relação com o processo político vigente no ano de 2018 relativo às eleições. Em 2019, a categoria de notícias mais presente no *dataset* refere-se ao tema política com um total de 298 notícias falsas, onde há grande quantidade de notícias falsas envolvendo o nome de Bolsonaro e Lula. No ano de 2020, a categoria de notícias mais frequente refere-se ao tema política com um total de 289 notícias falsas. Porém, é necessário nesse ano se atentar ao segundo tema mais tratado nas notícias falsas, ou seja, o tema saúde com 263 notícias falsas, que apresenta grande relação com a pandemia de COVID-19. Por fim, em 2021, a categoria de notícias mais frequente refere-se ao tema saúde com um total de 163 notícias falsas, fato que também pode estar associado com a pandemia de COVID-19.

Ao mesmo tempo em que foram analisadas as categorias mais frequentes, também foram analisadas as categorias menos frequentes nas notícias falsas nos anos referentes às notícias do *corpus*, caracterizando-se da seguinte forma:

Em 2013, as categorias de notícias falsas com menor número de ocorrências no *dataset* referem-se aos temas: Ciência, Esporte, Lista e Opinião, ambas com nenhuma notícia desmentida. No ano de 2014, às categorias de notícias falsas com menor número de ocorrências no *dataset* referem-se aos temas: Lista, que basicamente trata-se de uma categorização do site fonte das notícias atribuída a publicações que envolvem listas e rankings de notícias falsas; e o tema Opinião, ambos com 1 notícia falsa. Após isso, em 2015, as categorias de notícias falsas com menor número de ocorrências no *dataset* referem-se ao tema Lista e o tema Opinião, ambos com nenhuma notícia falsa. Posteriormente, em 2016, as categorias de notícias falsas com menor número de ocorrências no *dataset* referem-se ao tema Lista

[25] Disponível em: https://www.mpf.mp.br/grandes-casos/lava-jato/entenda-o-caso
[26] Disponível em: https://www.tse.jus.br/eleicoes/eleicoes-anteriores/eleicoes-2016

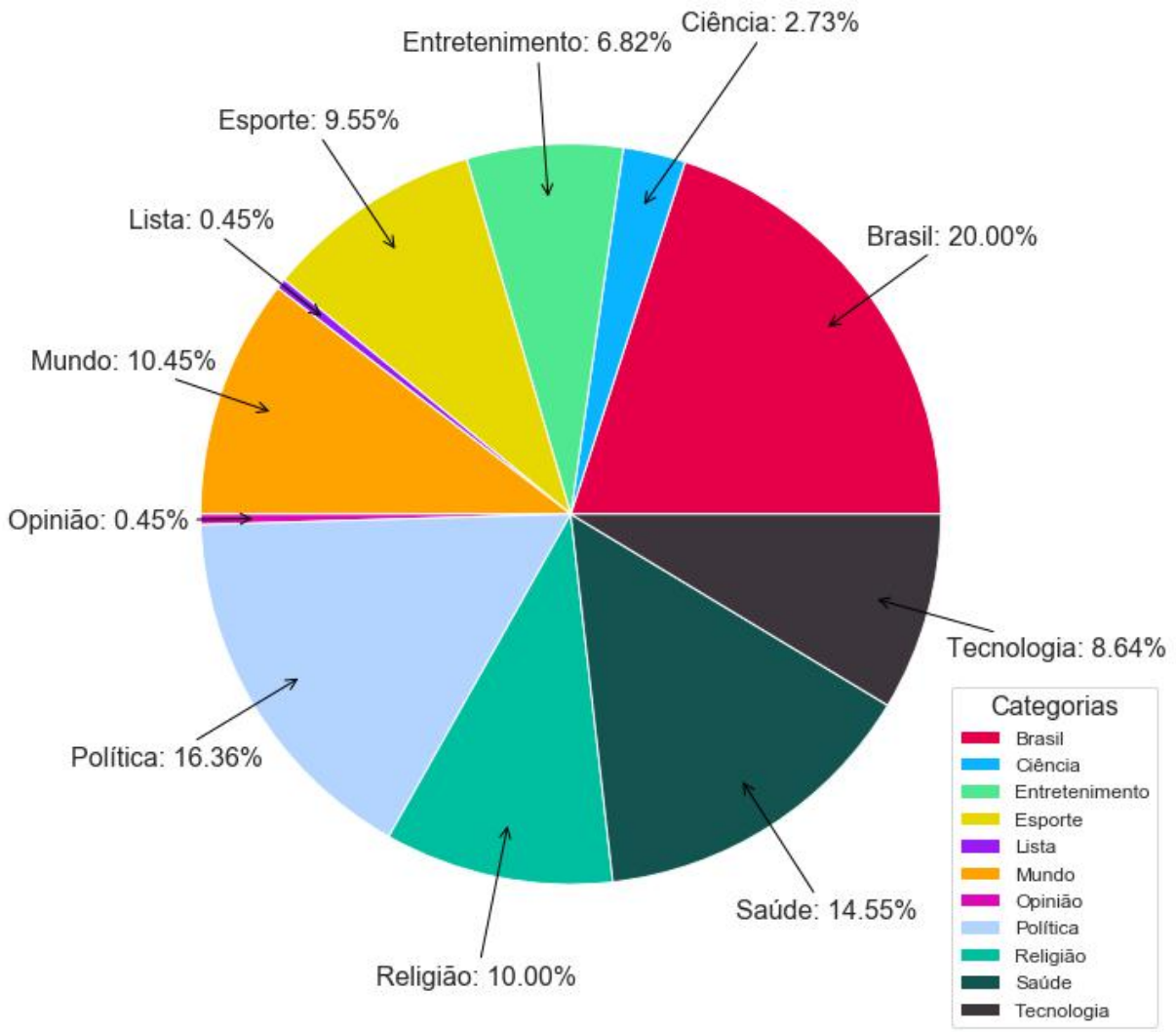
e o tema Opinião, ambos com 2 notícias falsas. Em 2017, as categorias de notícias falsas com menor número de ocorrências no *dataset* referem-se ao tema Lista e o tema Opinião, ambos com 2 notícias falsas, em cada. No ano de 2018, a categoria de notícias com menor ocorrência refere-se ao tema Opinião com nenhuma notícia falsa.

Em 2019, as categorias de notícias falsas com menor número de ocorrências no *dataset* referem-se ao tema Lista e o tema Opinião, ambos com 1 notícia falsa, em cada. Em 2020, a categoria de notícias com menor ocorrência refere-se ao tema Lista, com nenhuma notícia falsa. Por fim, em 2021, a categoria de notícias menos frequente refere-se ao tema Lista.

Após esta macroanálise relativa às categorias das notícias considerando o *dataset* Fakepedia completo, torna-se necessário realizar uma análise mais específica relativa a cada ano presente no *dataset*, visto que oscilações ao longo dos meses de cada ano podem determinar qual o assunto mais comentado, o que pode ter relação com acontecimentos históricos e sociais referentes a cada período de tempo.

Considerando o triênio 2013-2015 observa-se que os assuntos mais tratados estão relacionados com política, assuntos nacionais e saúde. De forma mais específica, o tema mais tratado tem relação com assuntos nacionais, conforme pode ser visto no Gráfico 24.

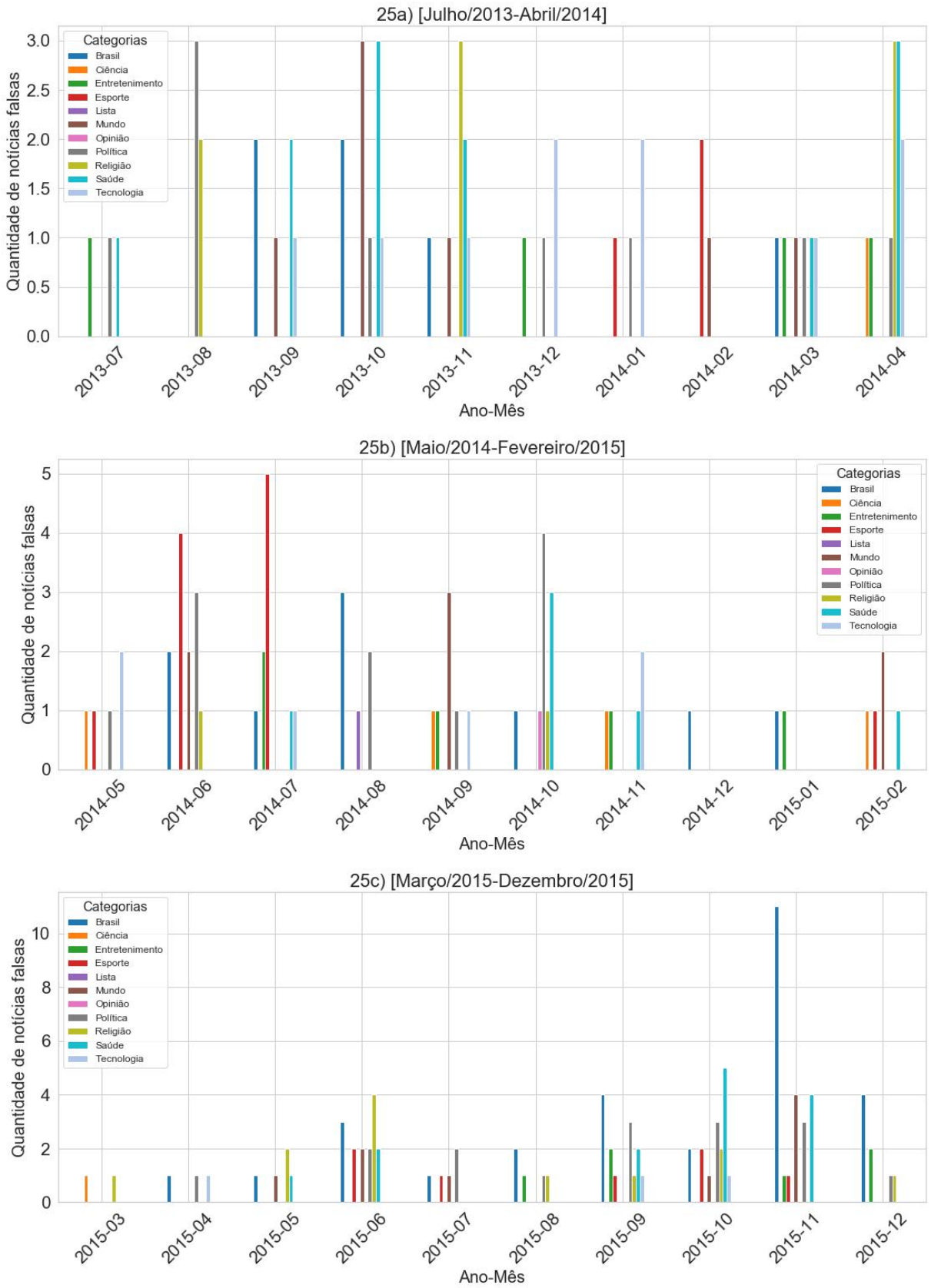
Gráfico 24: Proporção de cada categoria no triênio (2013-2015)



Fonte: Elaborado pelo autor (2023).

No Gráfico 25 é possível observar uma caracterização geral do triênio 2013-2015 ao longo dos meses em relação a distribuição de categorias mais frequentes. Dessa forma, compreende-se as variações quantitativas entre as categorias. A maior variação entre categorias pode ser percebida em novembro de 2015 quando a categoria sobre assuntos nacionais tem uma grande alta, o que alavanca o crescimento de notícias nesse mês. Porém, entre as notícias tratadas não há nenhum assunto predominante, ou seja, o mês de novembro apresenta notícias falsas variadas com boatos e rumores envolvendo diferentes temáticas.

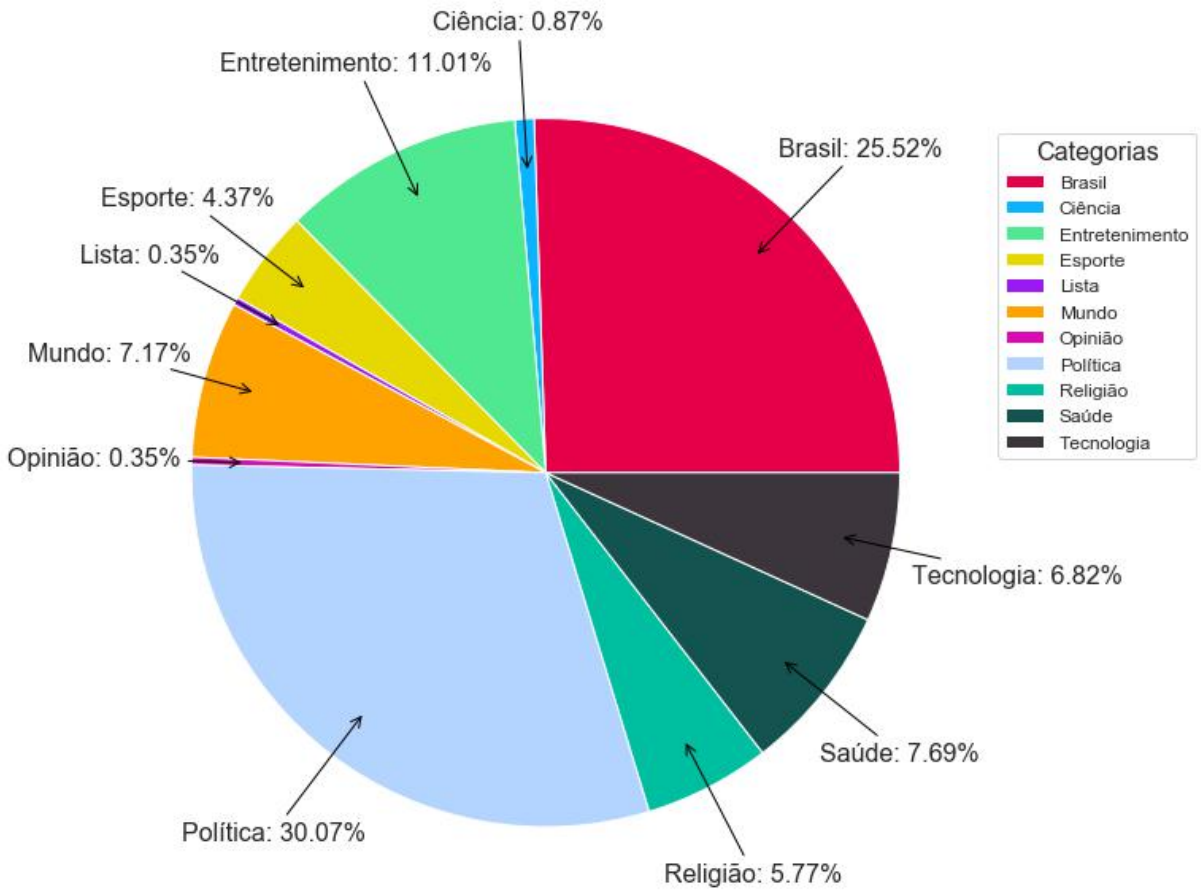
Gráfico 25: Quantidade de notícias falsas por categoria ao longo do triênio (2013-2015). 25a) Julho/2013-Abril/2014. 25b) Maio/2014-Fevereiro/2015. 25c) Março/2015-Dezembro/2015.



Fonte: Elaborado pelo autor (2023).

No ano de 2016, identifica-se que os assuntos mais tratados estão relacionados com política, assuntos nacionais e entretenimento. Observando detalhadamente, constata-se que o tema mais abordado nas notícias falsas tem relação com assuntos relacionados a política conforme pode ser visto no Gráfico 26.

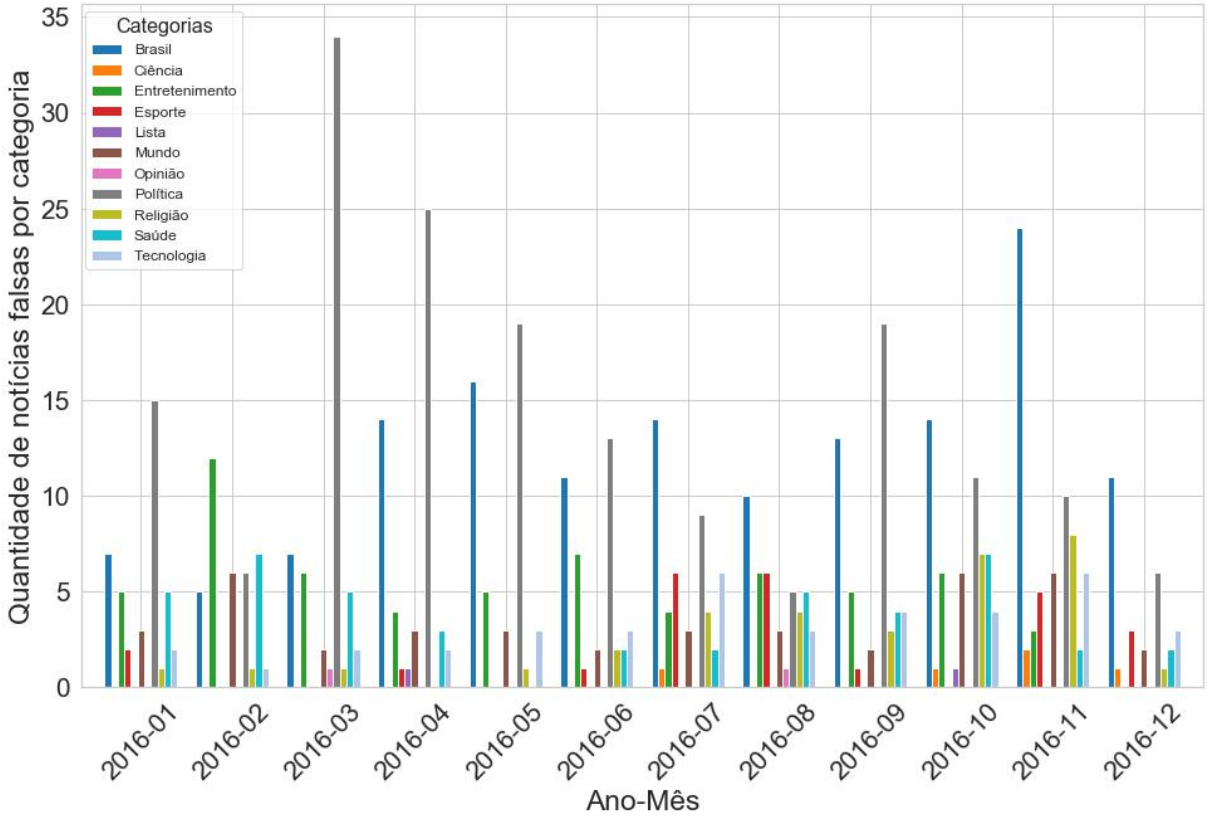
Gráfico 26: Proporção de cada categoria no ano de 2016



Fonte: Elaborado pelo autor (2023).

Se observarmos a divisão entre os meses do ano de 2016 no Gráfico 27, temos que o primeiro semestre do ano de 2016 tem o tema política como categoria de notícia mais frequente. Após isso, em julho e agosto, há maior quantidade de notícias relacionadas com o tema assuntos nacionais (Brasil), que são notícias falsas variadas que tratam de assuntos do âmbito nacional. Já em setembro, o tema sobre política volta a ser a categoria mais frequente no *dataset*. Por fim, os meses de outubro, novembro e dezembro tem como categoria mais frequente, o tema sobre assuntos nacionais (Brasil) novamente.

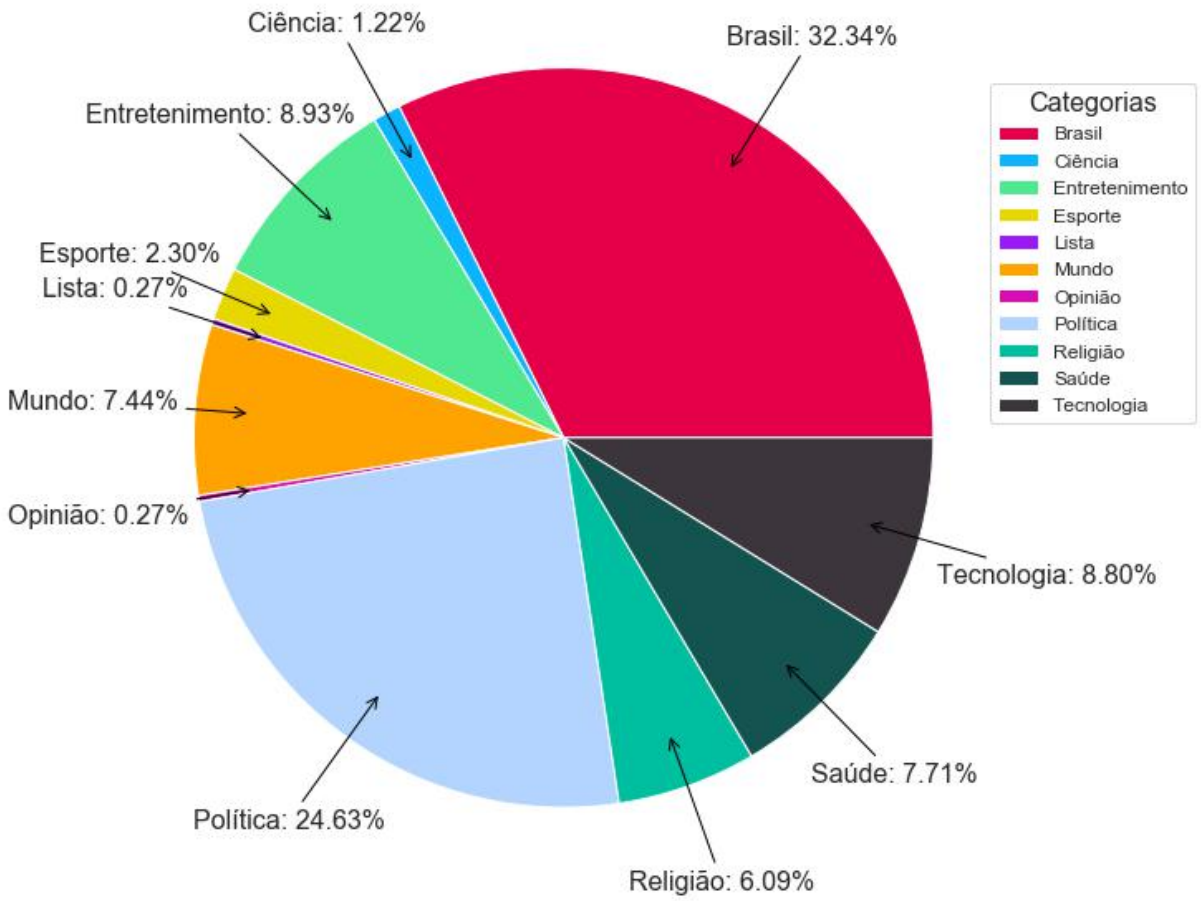
Gráfico 27: Quantidade de notícias falsas por categoria ao longo do ano de 2016



Fonte: Elaborado pelo autor (2023).

No ano de 2017 há destaque para notícias falsas relacionadas com assuntos nacionais e política. De maneira mais específica, o tema mais abordado nas notícias falsas tem relação com assuntos nacionais conforme pode ser visto no Gráfico 28.

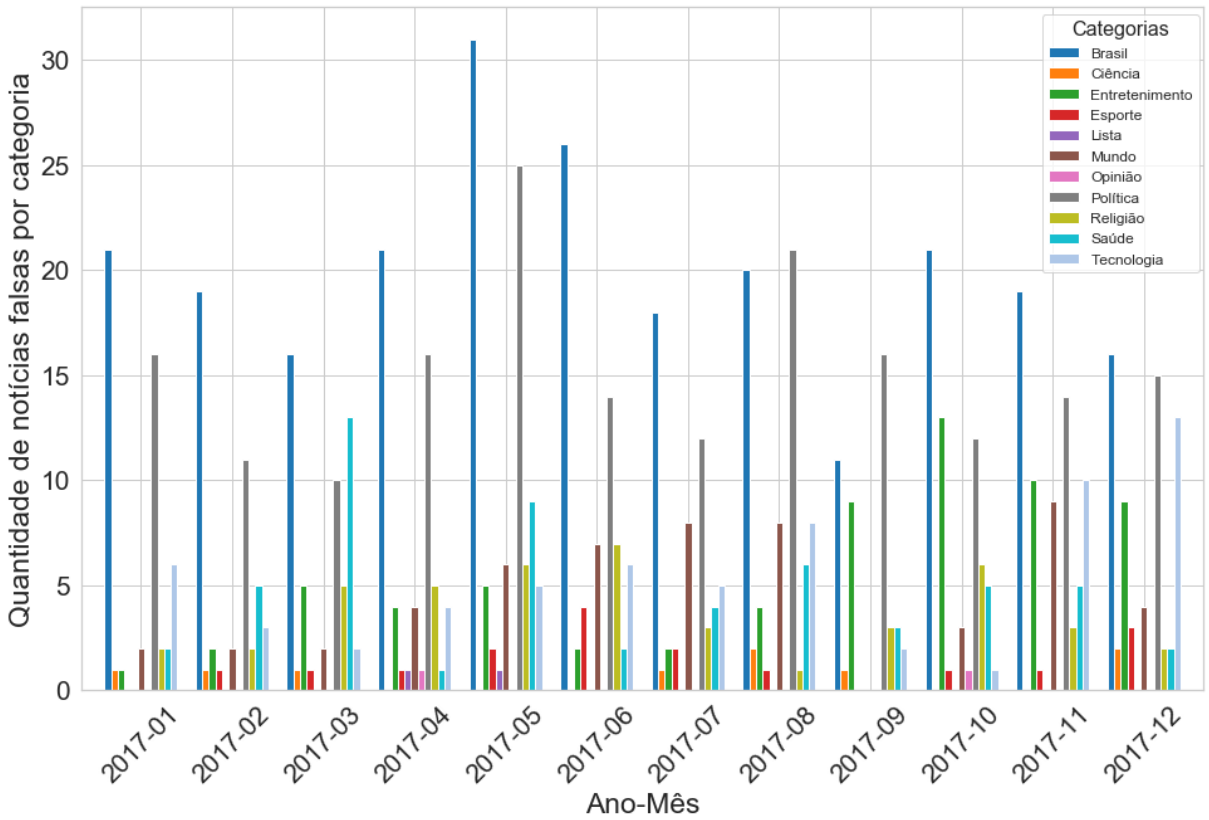
Gráfico 28: Proporção de cada categoria no ano de 2017



Fonte: Elaborado pelo autor (2023).

Analisando o ano de 2017 com relação aos meses no Gráfico 29, temos que a categoria predominante nas notícias falsas ao longo dos meses refere-se ao tema sobre assuntos nacionais. Com exceção dos meses de agosto e setembro, onde a política se torna a categoria mais frequente. Em geral, boa parte das notícias falsas sobre política relativa aos meses de agosto e setembro envolvem nomes de diferentes políticos na criação de diversas informações falsas, ou seja, tanto abordando informações falsas sobre assuntos que existem, quanto sobre assuntos e situações inexistentes, isto é, situações criadas por disseminadores de notícias falsas utilizadas com o intuito de atingir a reputação de pessoas.

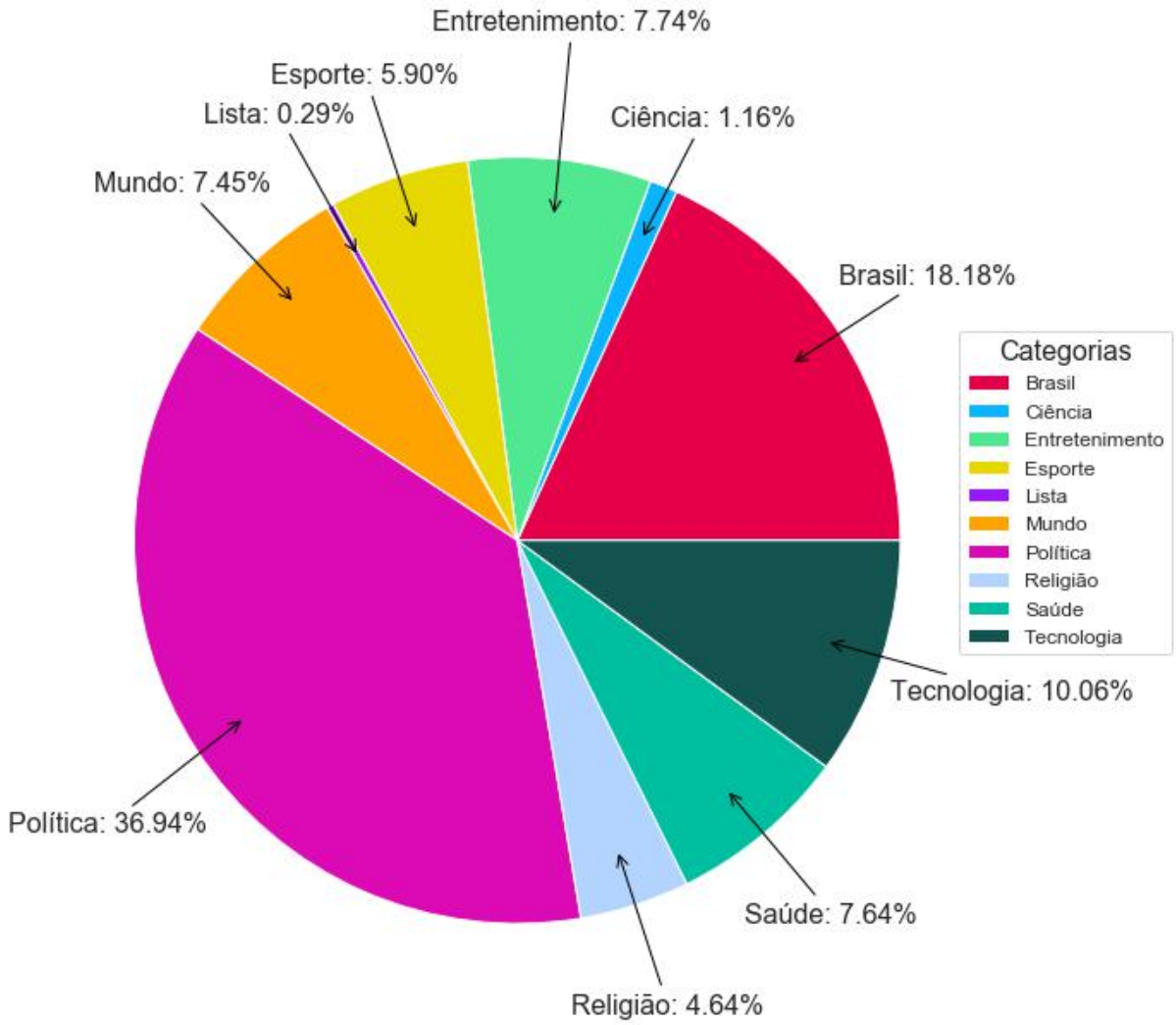
Gráfico 29: Quantidade de notícias falsas por categoria ao longo do ano de 2017



Fonte: Elaborado pelo autor (2023).

Em 2018, pode-se destacar o aumento de notícias relacionadas à política em detrimento das notícias relativas à assuntos nacionais. Dessa forma, o tema mais abordado nas notícias falsas tem relação com política, em seguida temos assuntos nacionais como segundo tema mais abordado nas notícias falsas, conforme pode ser visto no Gráfico 30.

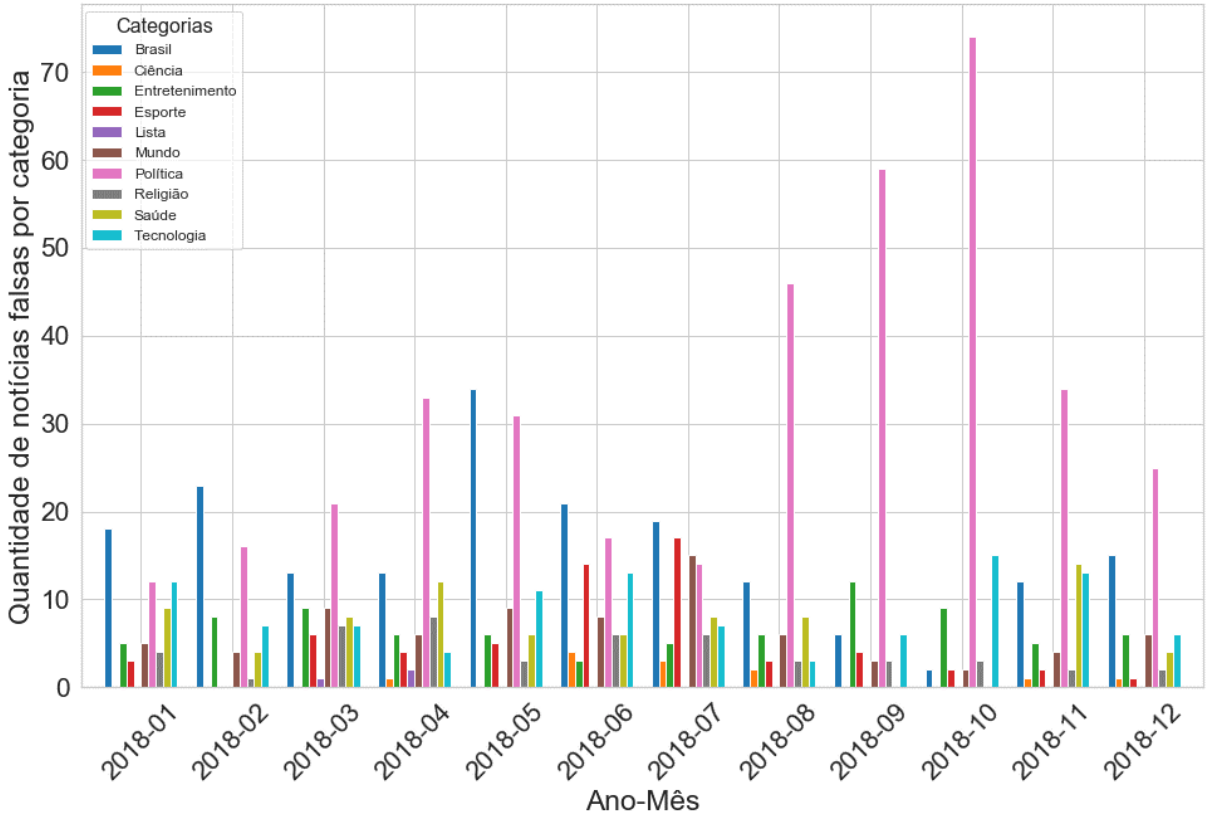
Gráfico 30: Proporção de cada categoria no ano de 2018



Fonte: Elaborado pelo autor (2023).

Observando os meses do ano de 2018 individualmente conforme está ilustrado no Gráfico 31, torna-se perceptível o fato de que o tema sobre assuntos nacionais (Brasil) é a categoria mais frequente, em janeiro e fevereiro. Após isto, em março e abril, o tema mais frequente passa a ser política. Posteriormente, em maio, junho e julho, o tema Brasil volta a ser a categoria mais frequente. Por fim, de agosto a dezembro, a categoria mais frequente se refere à política, e isso se deve principalmente ao fato de que esse momento corresponde ao período de acirramento da disputa das eleições presidenciais que, por sua vez, pode ser observada com frequência ao ler as principais notícias desse período.

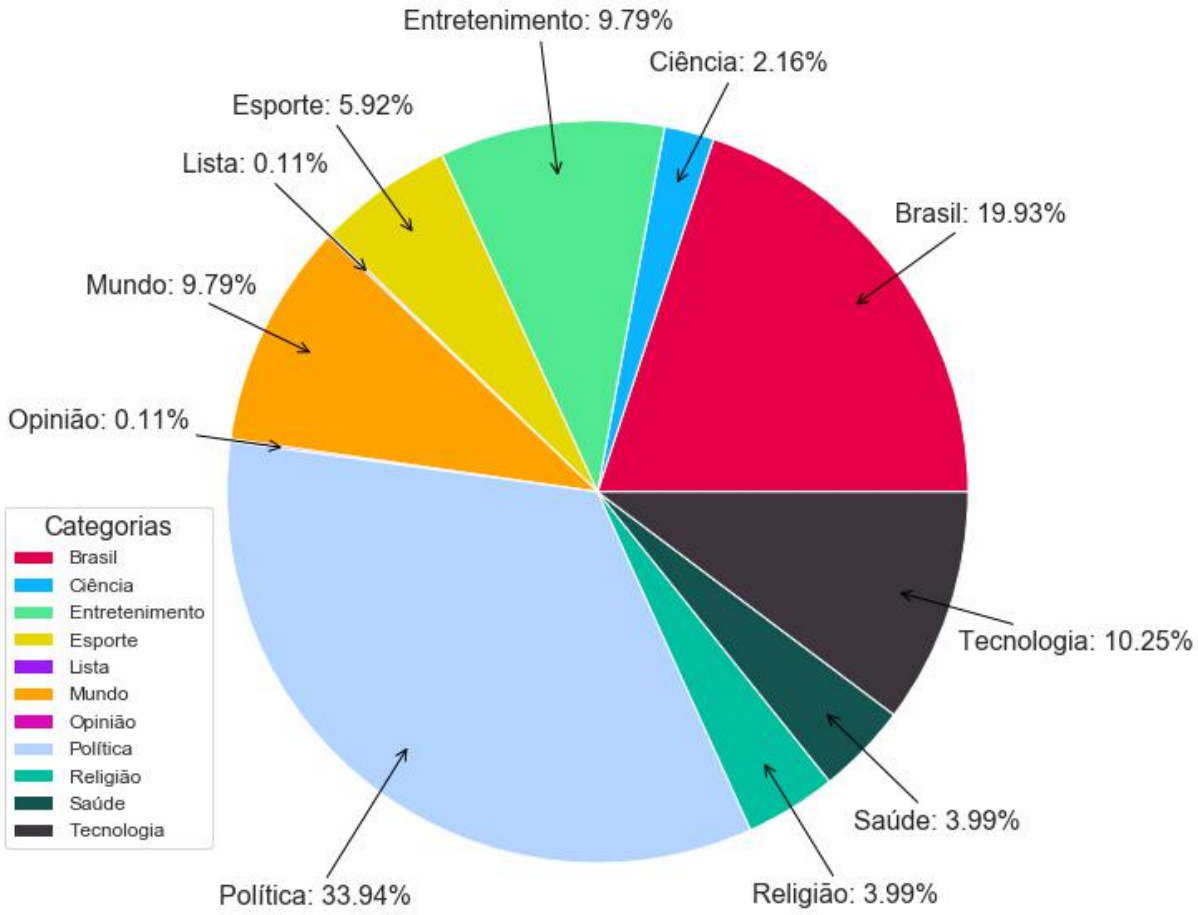
Gráfico 31: Quantidade de notícias falsas por categoria ao longo do ano de 2018



Fonte: Elaborado pelo autor (2023).

Considerando o ano de 2019, ocorre um pequeno aumento de notícias falsas relativas a assuntos nacionais e uma leve diminuição de notícias falsas relacionadas à política, conforme pode ser visto no Gráfico 32. Observando as principais notícias desse período é perceptível a diminuição de notícias que tratem sobre assuntos ligados a eleições que, por sua vez, foram amplamente disseminadas no ano anterior. Portanto, um dos motivos da diminuição supracitada tem relação com esse fato.

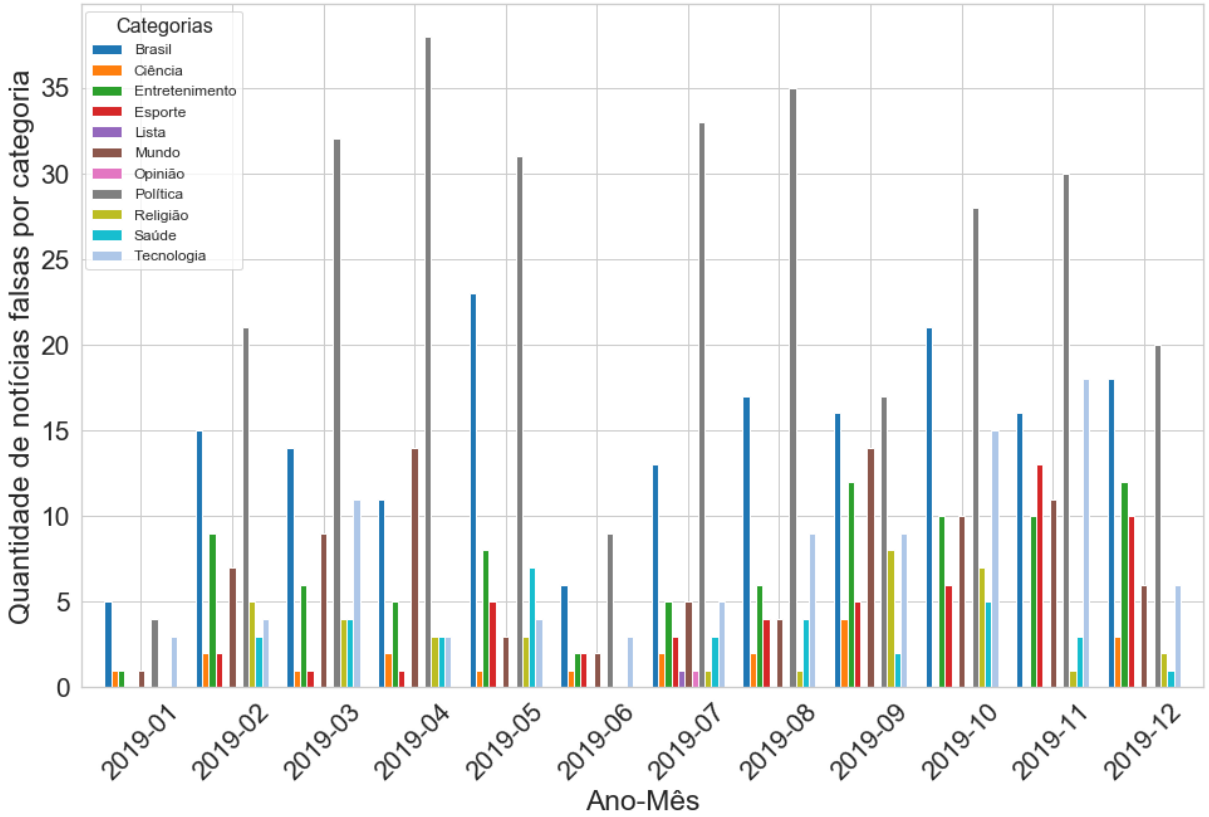
Gráfico 32: Proporção de cada categoria no ano de 2019



Fonte: Elaborado pelo autor (2023).

Observando o ano de 2019 com relação à sua distribuição mensal no Gráfico 33, pode-se constatar que quase todos os meses tem como categoria principal, o tema sobre política, com exceção do mês de janeiro, que tem o tema sobre assuntos nacionais (Brasil) como categoria mais frequente.

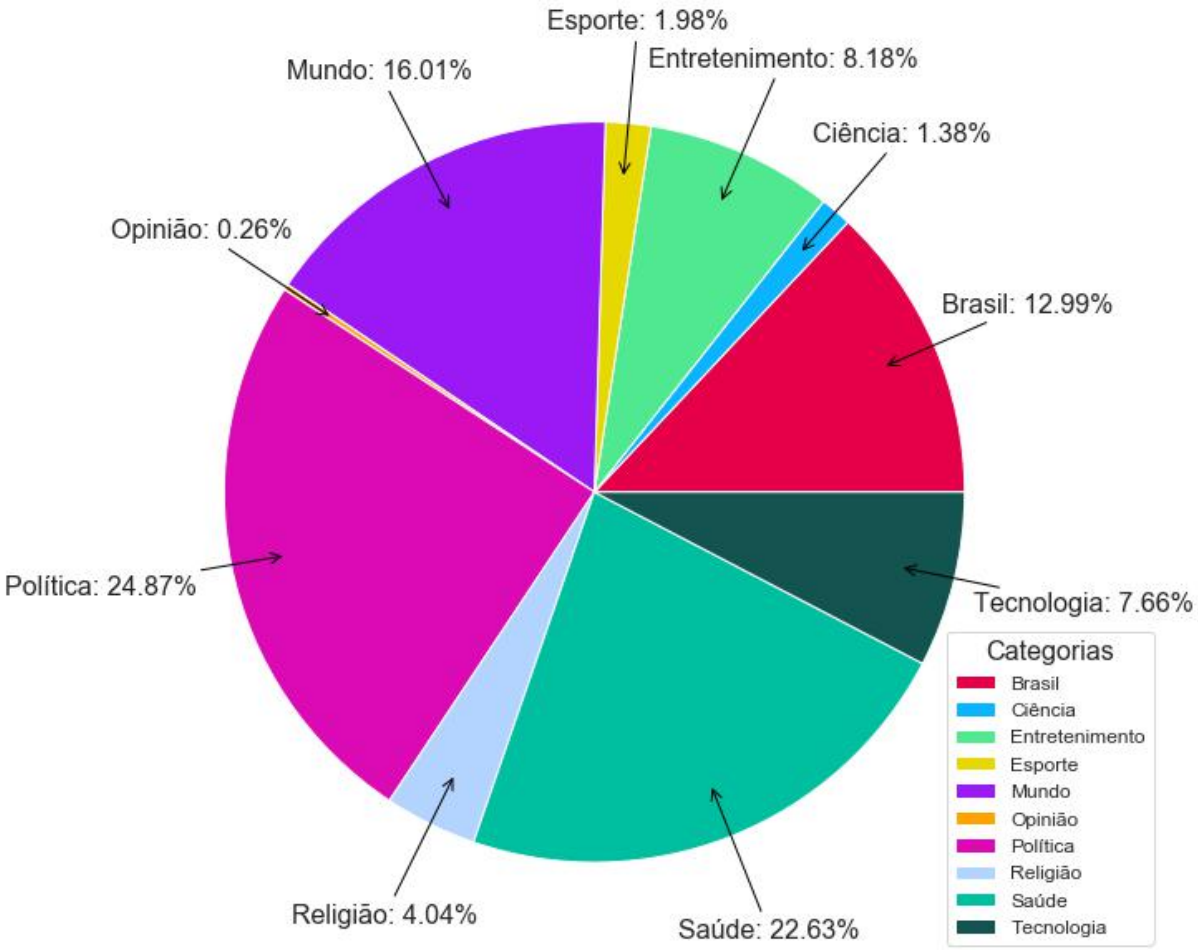
Gráfico 33: Quantidade de notícias falsas por categoria ao longo do ano de 2019



Fonte: Elaborado pelo autor (2023).

Em 2020, é possível perceber uma grande mudança no cenário de abordagem das notícias falsas, isto é, a categoria relativa a saúde passa a estar mais presente junto com a categoria relativa a assuntos políticos, conforme pode ser visto no Gráfico 34. Observando as principais notícias desse período, constata-se que muitas notícias tratam assuntos referentes a pandemia de COVID-19, mais especificamente abordando assuntos relativos a possíveis efeitos da doença, e possíveis curas para a doença.

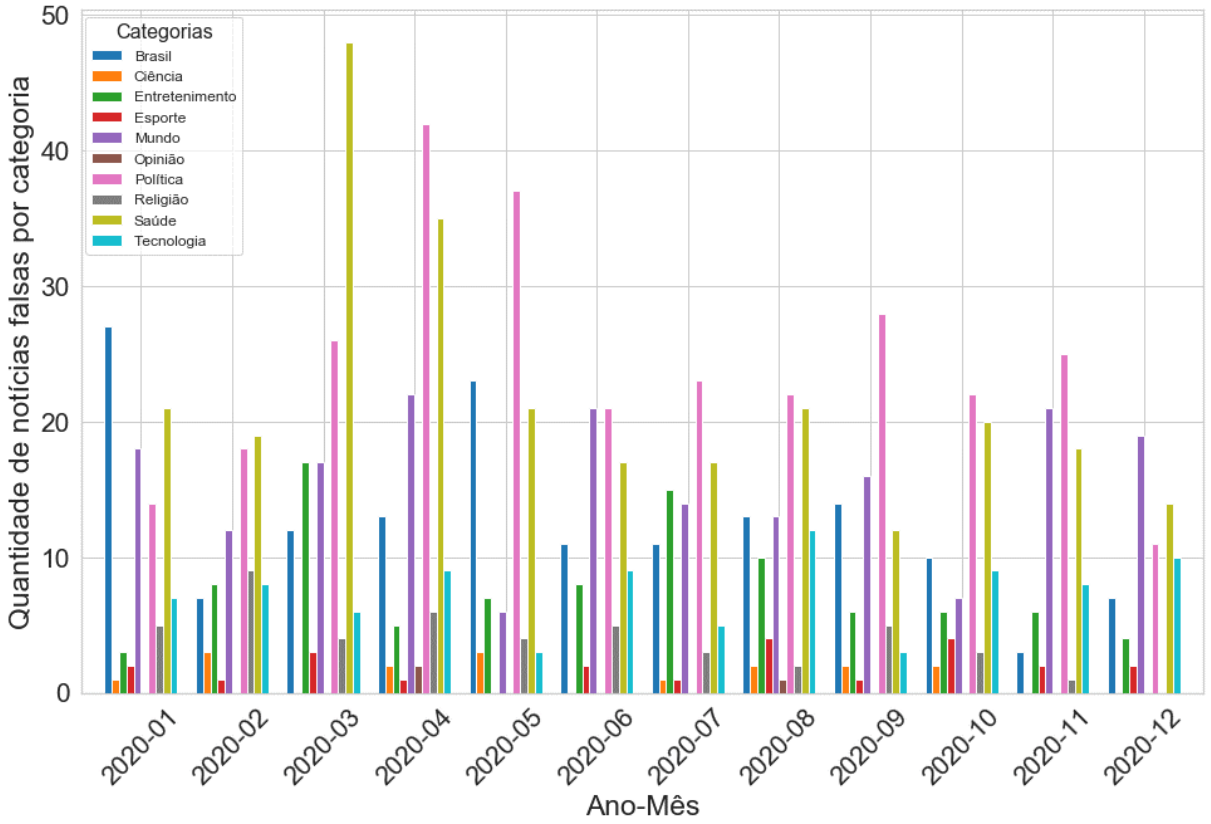
Gráfico 34: Proporção de cada categoria no ano de 2020



Fonte: Elaborado pelo autor (2023).

Observando a caracterização de 2020 com relação aos meses apontados no Gráfico 35, obtém-se maior entendimento, ou seja, temos que esse ano inicia tendo o tema sobre assuntos nacionais (Brasil) como a categoria mais frequente, e nos meses de fevereiro e março o tema sobre saúde torna-se a categoria mais frequente. Após isso, do mês de abril até o mês de novembro, a categoria mais frequente refere-se a política. Por fim, o mês de dezembro tem como categoria mais frequente, o tema Mundo, que se refere a notícias internacionais, em geral. Esta evolução ao longo dos meses pode ter relação com a pandemia de COVID-19, visto que nos meses de fevereiro e março, o assunto mais comentado era saúde, e após o surgimento da pandemia, nos meses subsequentes, o tema sobre política começou a ganhar mais espaço nas notícias falsas, muito provavelmente devido às movimentações políticas decorrentes de ações em relação à pandemia.

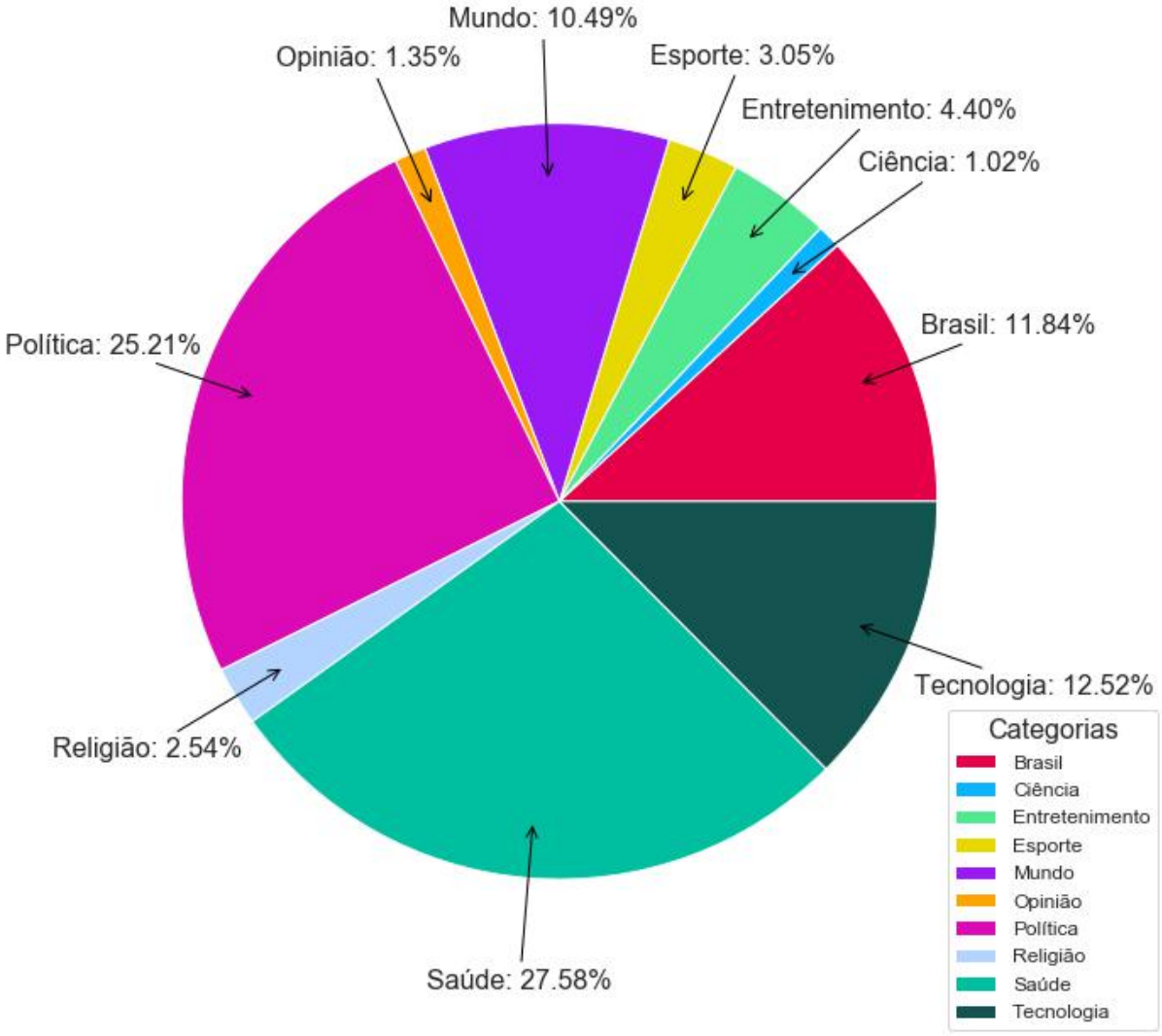
Gráfico 35: Quantidade de notícias falsas por categoria ao longo do ano de 2020



Fonte: Elaborado pelo autor (2023).

Considerando o ano de 2021, temos que as notícias falsas relativas ao tema saúde se consolida como tema mais tratado, seguido por política, conforme pode ser visto no Gráfico 36. A principal diferença pode ser percebida ao considerar que as principais notícias desse período abordam tanto assuntos relacionados com medidas políticas de combate a pandemia, quanto assuntos relativos às vacinas desenvolvidas.

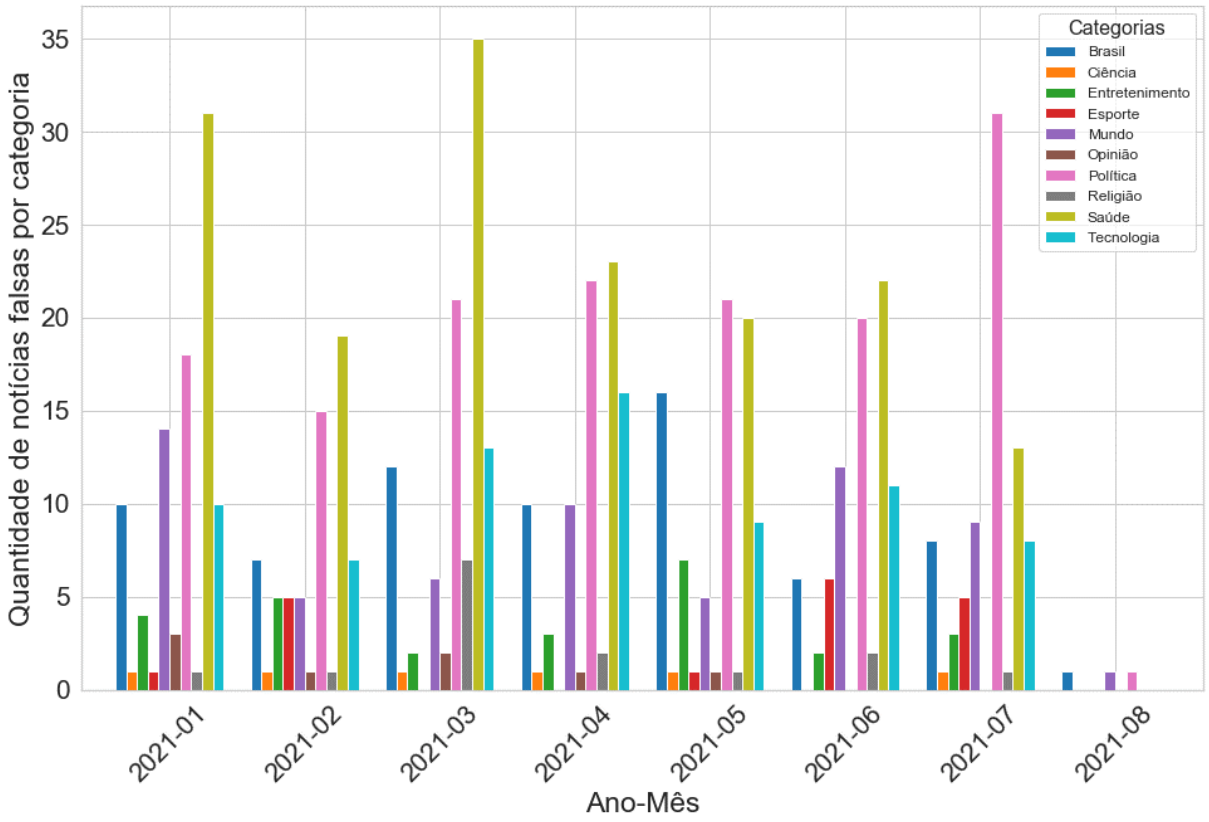
Gráfico 36: Proporção de cada categoria no ano de 2021



Fonte: Elaborado pelo autor (2023).

Analisando o ano de 2021 no Gráfico 37, temos que o *corpus* possui notícias até o mês de agosto, sendo que agosto possui apenas 3 notícias no *dataset*, muito provavelmente devido ao fato da construção do *dataset* ter sido realizada com coleta de notícias que encerram em agosto de 2021. Em geral, no ano de 2021 quase todos os meses tem como categoria principal, o tema sobre saúde, com exceção dos meses de maio e julho, que possuem como categoria mais frequente, o tema de política, fato também relacionado com os acontecimentos da pandemia de COVID-19.

Gráfico 37: Quantidade de notícias falsas por categoria ao longo do ano de 2021



Fonte: Elaborado pelo autor (2023).

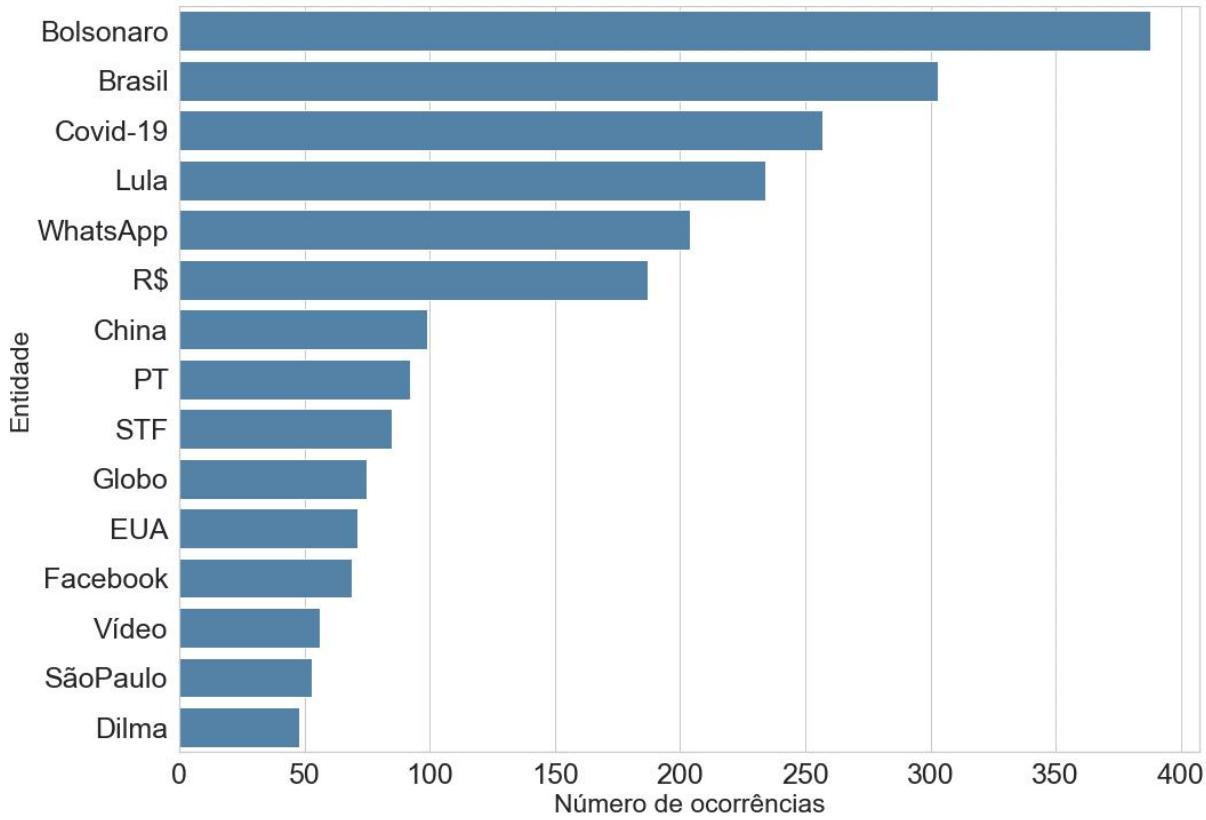
Portanto, ao considerar a análise relativa as categorias realizadas no *corpus*, é possível perceber a grande relação entre acontecimentos históricos e sua relação com as categorias mais frequentes entre as notícias falsas. Por conseguinte, é importante destacar como acontecimentos históricos impactantes implicam na alta construção de notícias falsas, como pode ser visto nessa análise ao considerar o período de surgimento da pandemia de COVID-19 e sua relação com o aumento do número de notícias falsas, bem como o aumento de notícias da categoria relacionada à saúde. Além disso, é possível perceber que outro acontecimento histórico que destaca essa implicação pode ser notado ao considerar o período de eleições de 2018. Portanto, percebe-se no contexto geral, que a maior parte das notícias presentes no *corpus* está associada a assuntos nacionais e política, sendo que as notícias falsas relacionadas com o tema sobre saúde tomaram grande proporção a partir do momento de deflagração da pandemia de COVID-19.

5.3 ANÁLISE TEMPORAL DAS ENTIDADES PRESENTES NAS NOTÍCIAS

Nesta seção é realizada uma análise temporal relativa às entidades nomeadas presentes nas notícias com o objetivo de compreender as principais entidades que fazem parte do assunto das notícias presentes no *dataset* ao longo do tempo. Consequentemente, essa análise pode corroborar com as inferências realizadas na seção anterior, onde foram tratadas a distribuição da frequência de categorias ao longo do *dataset*.

Considerando o *corpus* completo, há um total de 9.472 ocorrências de entidades ao longo do texto das notícias, sendo 3.746 entidades nomeadas distintas, isto é, boa parte das 3.746 entidades distintas é referenciada mais de uma vez ao longo das notícias estudadas. Dentre as entidades abordadas em todas as notícias do *corpus*, aproximadamente 20,2% estão relacionadas com pessoas, 22,9% estão relacionadas com tipos de entidades variadas (eventos, nacionalidades, produtos, etc), 12,6% estão relacionadas com organizações e 44,1% estão relacionadas com localidades. Ainda considerando todo o *corpus*, temos que as principais entidades nomeadas citadas são: “Bolsonaro” com 388 ocorrências, “Brasil” com 303 ocorrências, “Covid-19” com 257 ocorrências, “Lula” com 234 ocorrências e “WhatsApp” com 204 ocorrências. Dessa forma, é possível compreender como as figuras do cenário político são citadas de maneira recorrente em notícias falsas. Além disso, a entidade “WhatsApp” foi identificada como uma das mais frequentes e, de forma específica, tem a maior parte de suas citações sendo feitas como o veículo de disseminação das notícias falsas, fato este que é constatado ao ler as principais notícias que contém sua presença. A distribuição geral das entidades nomeadas ao longo de todo o *corpus* pode ser observada no Gráfico 38.

Gráfico 38: Distribuição geral das entidades nomeadas ao longo de todo o corpus



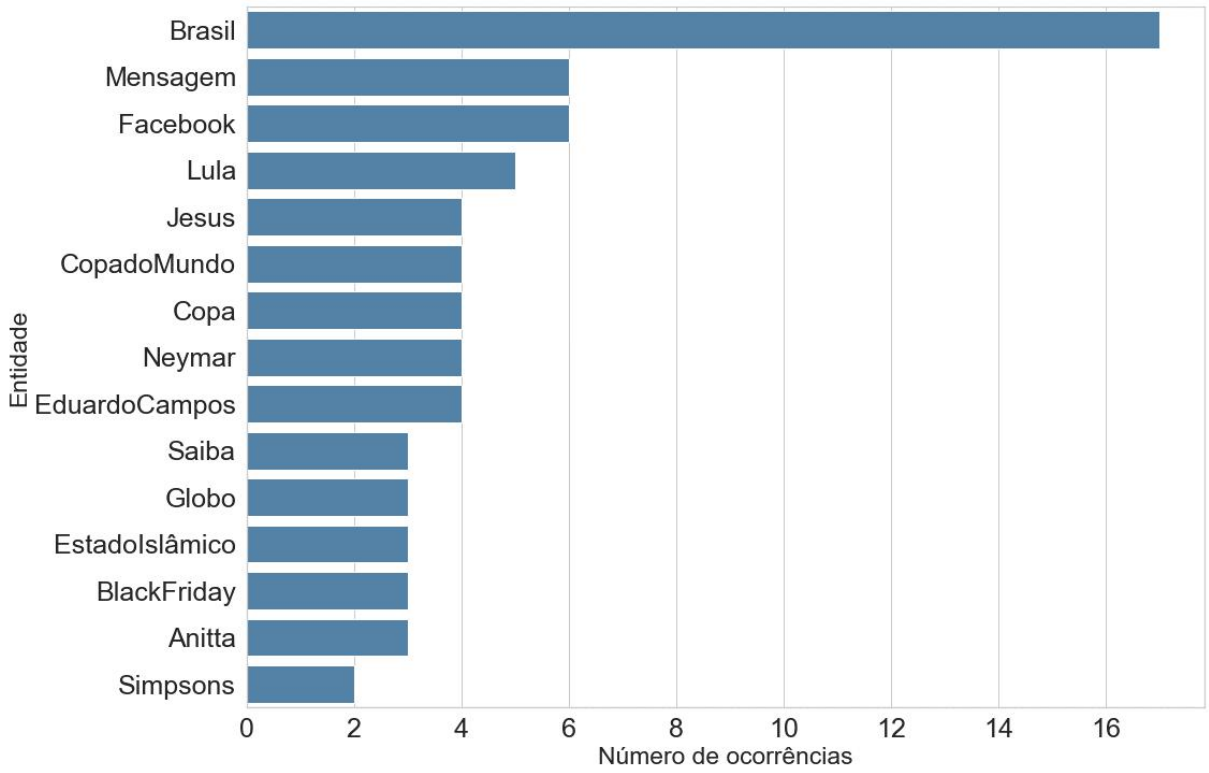
Fonte: Elaborado pelo autor (2023).

Como já citado anteriormente, os anos de 2013, 2014 e 2015 foram analisados em conjunto, isto é, esses anos foram armazenados em único *dataframe* para análise, devido ao número reduzido de notícias no triênio. Considerando o triênio 2013-2015, temos que a entidade nomeada mais frequente refere-se ao termo “Brasil” com 17 ocorrências, ou seja, ao considerar a análise de distribuição de categorias realizadas na seção anterior, percebemos que a categoria mais frequente levando em conta a média no triênio 2013-2015 refere-se às notícias nacionais, em outras palavras, tem como categoria o tema sobre assuntos nacionais (Brasil). Com essa análise de entidades pode-se corroborar esse fato, visto que a entidade mais frequente nas notícias desse período de tempo refere-se ao Brasil também. Em seguida, aparecem outras entidades menos frequentes, tais como: Copa do Mundo e Eduardo Campos, que estão associadas a acontecimentos históricos referidos nos anos do triênio, incluindo tanto a Copa do Mundo FIFA de 2014[27], bem como a eleição presidencial

[27] Disponível em: https://www.fifa.com/tournaments/mens/worldcup/2014brazil

ocorrida no ano de 2014. No Gráfico 39, é possível observar a distribuição geral das entidades nomeadas no referido triênio.

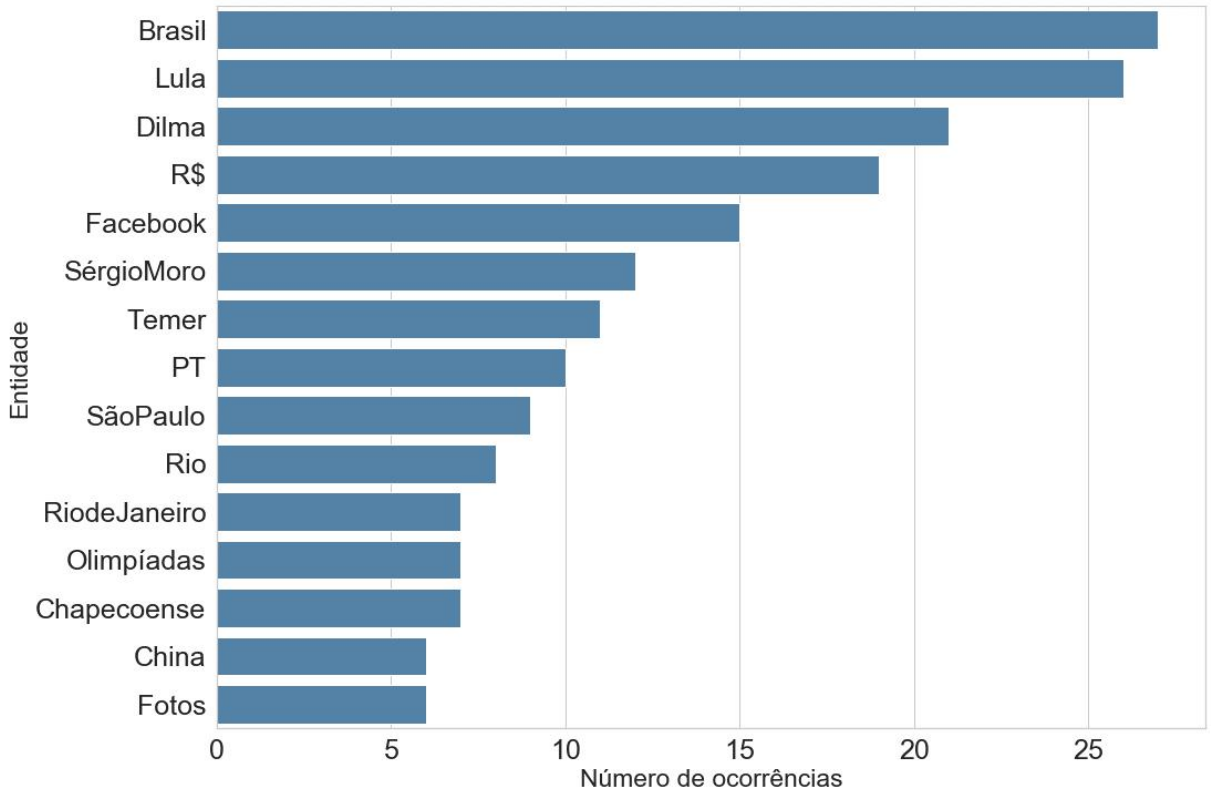
Gráfico 39: Distribuição geral das entidades nomeadas ao longo do triênio (2013-2015)



Fonte: Elaborado pelo autor (2023).

No ano de 2016 no *dataset*, as 3 entidades nomeadas mais citadas nos textos das notícias são: “Brasil” com 27 ocorrências, “Lula” com 26 ocorrências e “Dilma” com 21 ocorrências, sendo que essas entidades de certa forma corroboram os resultados referentes ao ano de 2016 encontrados na seção anterior, pois se considerarmos a análise relativa às categorias realizada anteriormente, temos que as categorias mais recorrentes nas notícias estão associadas tanto ao tema sobre assuntos nacionais (Brasil), quanto ao tema política. O principal acontecimento ocorrido em 2016 na categoria sobre política refere-se ao processo de *impeachment* instaurado. No Gráfico 40, temos a distribuição geral das principais entidades nomeadas referentes ao ano de 2016.

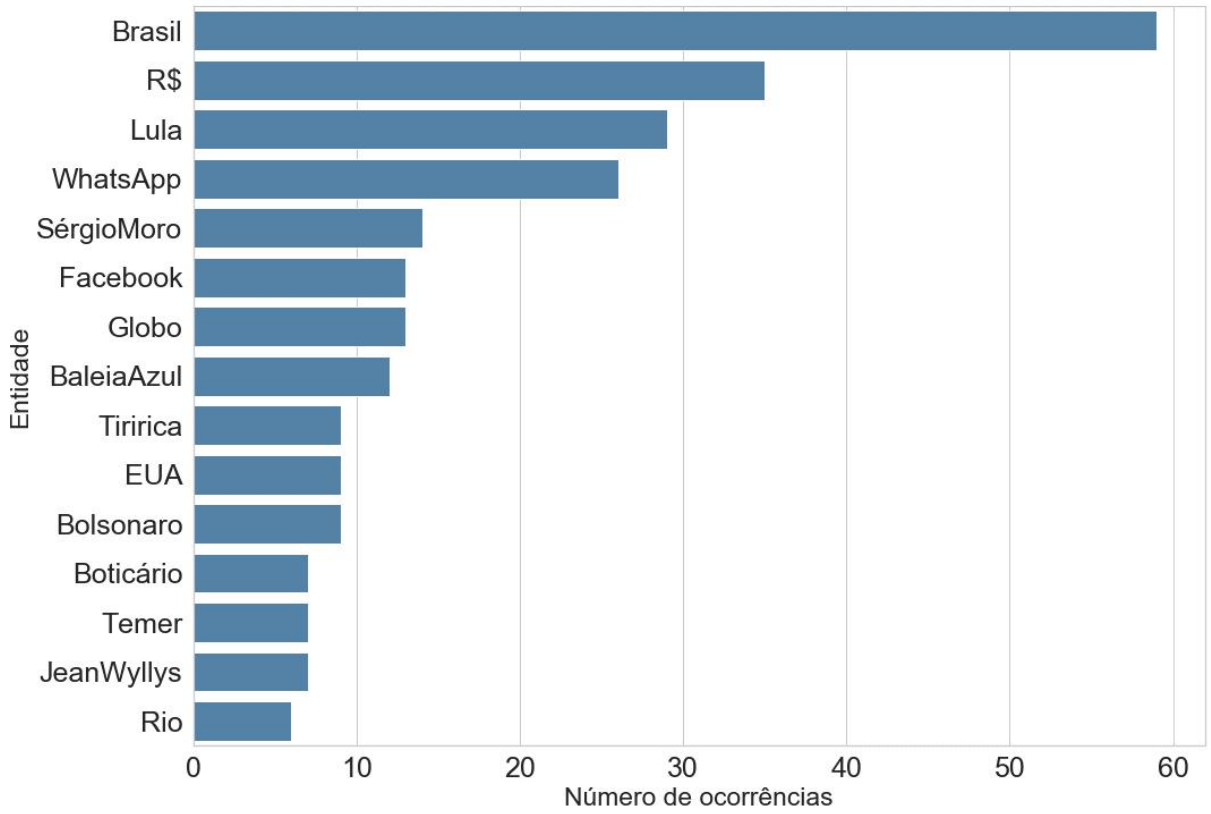
Gráfico 40: Distribuição geral das entidades nomeadas ao longo do ano de 2016



Fonte: Elaborado pelo autor (2023).

Observando as entidades nomeadas do ano de 2017 presentes nas notícias do *dataset* no Gráfico 41, temos que as entidades mais recorrentes são: “Brasil” com 59 ocorrências, “R$” (referente a valores monetários) com 35 ocorrências e “Lula” com 29 ocorrências. O fato dessas entidades serem as mais recorrentes, pode ser explicado se considerarmos a análise de categorias realizada na seção anterior, onde foi constatado que as categorias de notícias mais recorrentes no ano de 2017 foram relacionadas ao tema Brasil e ao tema política que, por sua vez, abriga a maior parte das ocorrências da entidade relacionada aos valores monetários, especificamente ao considerar notícias falsas sobre política associadas à corrupção.

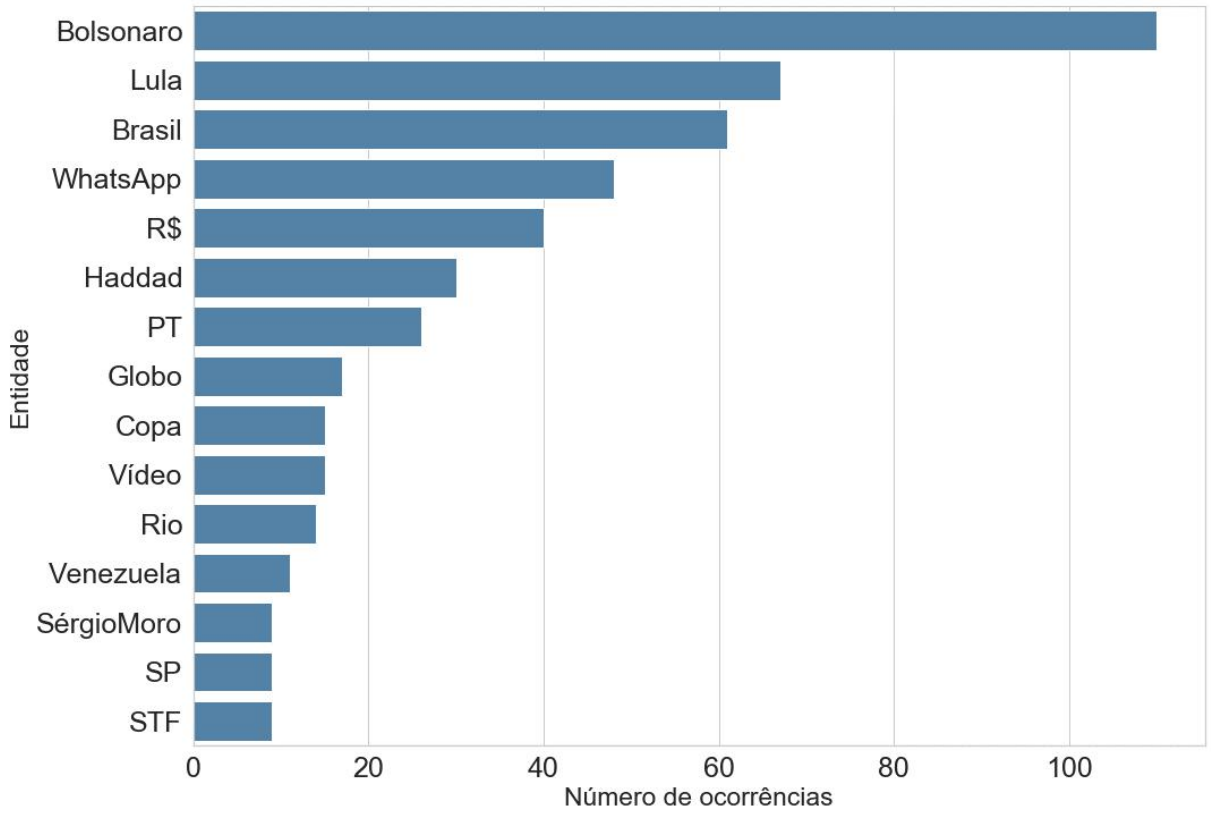
Gráfico 41: Distribuição geral das entidades nomeadas ao longo do ano de 2017



Fonte: Elaborado pelo autor (2023).

Se considerarmos a distribuição de entidades nomeadas das notícias do ano de 2018 presentes no *dataset*, representada no Gráfico 42, podemos verificar que as entidades mais recorrentes são: “Bolsonaro” com 110 ocorrências, “Lula” com 67 ocorrências e “Brasil” com 61 ocorrências. Comparando novamente a ocorrência dessas entidades com a análise relativa às categorias de notícias verifica-se que elas possuem intrínseca relação, ou seja, ao ler muitas das notícias falsas que tratavam sobre assuntos políticos foi identificada a presença do nome dos políticos Bolsonaro e Lula.

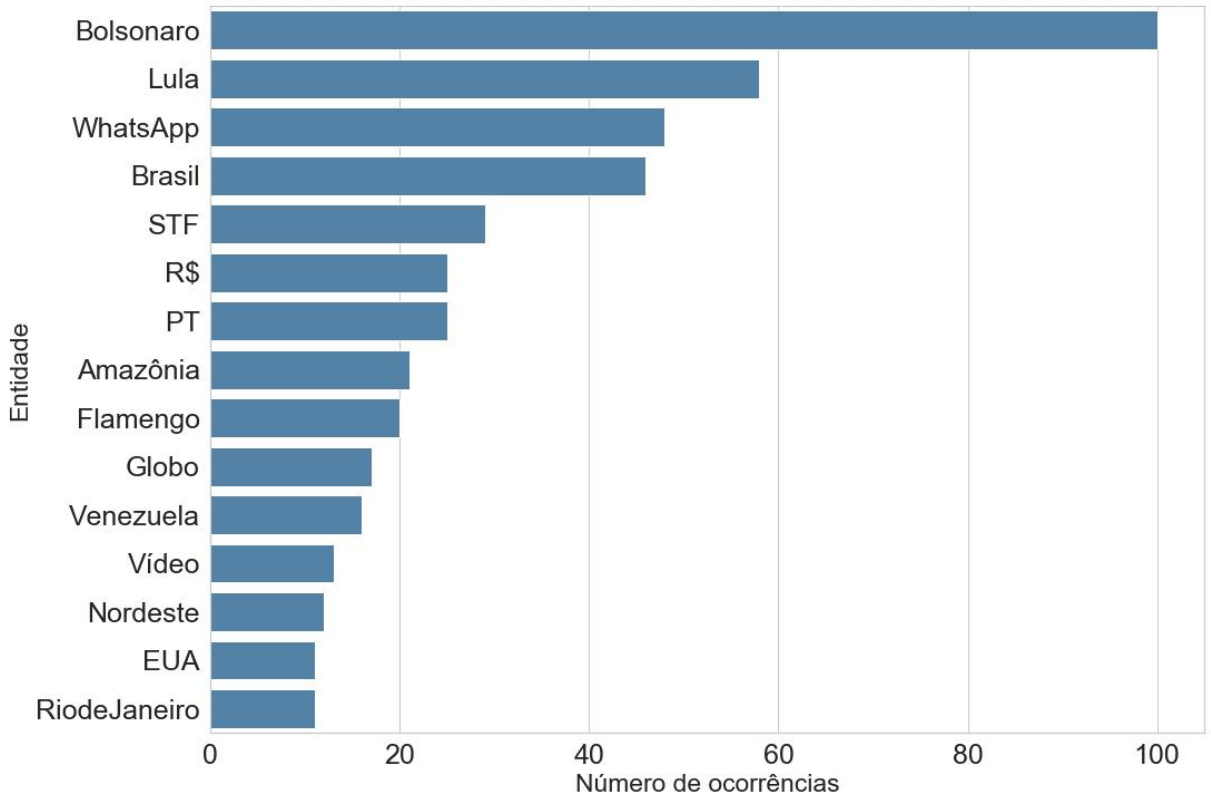
Gráfico 42: Distribuição geral das entidades nomeadas ao longo do ano de 2018



Fonte: Elaborado pelo autor (2023).

No ano de 2019 no *dataset*, temos que a distribuição de entidades nomeadas ilustrada no Gráfico 43, se caracteriza tendo as 3 principais entidades como: “Bolsonaro” com 100 ocorrências, “Lula” com 58 ocorrências e “Whatsapp” com 48 ocorrências. Sendo o ano de 2019, um ano com número majoritário de notícias atreladas a categoria de notícias relacionadas à política, como foi constatado na seção anterior. Além disso, é possível explicar a presença da entidade nomeada “Whatsapp” ao ler as principais notícias do referido período, ou seja, em boa parte das notícias ocorre a citação do termo “Whatsapp” com o intuito de apontar o meio de veiculação da notícia falsa.

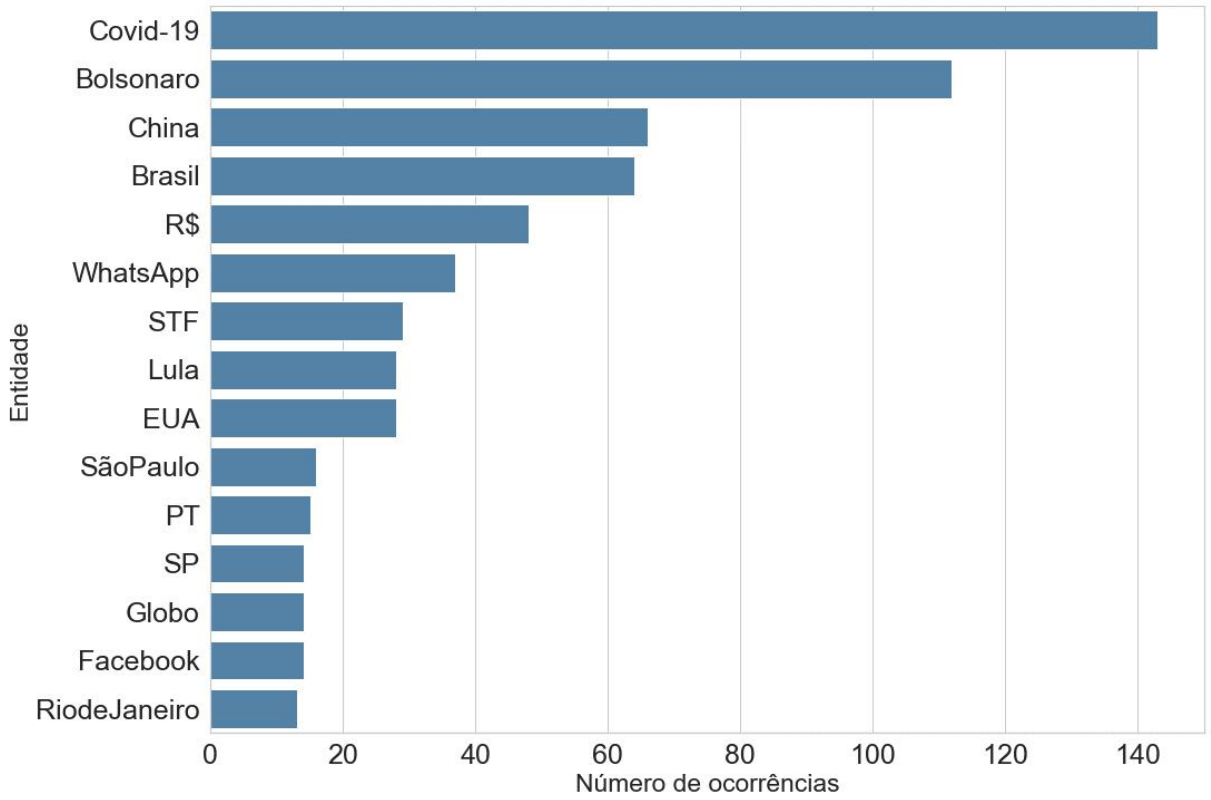
Gráfico 43: Distribuição geral das entidades nomeadas ao longo do ano de 2019



Fonte: Elaborado pelo autor (2023).

Ao analisarmos o *dataset* no ano de 2020 no Gráfico 44, temos que as 3 principais entidades referenciadas nas notícias falsas são: “Covid-19” com 143 ocorrências, “Bolsonaro” com 112 ocorrências e “China” com 66 ocorrências. Observando a análise por categoria realizada anteriormente, temos que as categorias mais recorrentes nas notícias falsas do ano de 2020 presentes no *dataset* referem-se ao tema saúde, e também ao tema política. Em síntese, podemos corroborar as inferências realizadas na seção anterior, atribuindo o fato da COVID-19 ser um fator para a alta de notícias falsas relacionadas ao tema de saúde, considerando que as entidades supracitadas foram as mais frequentes.

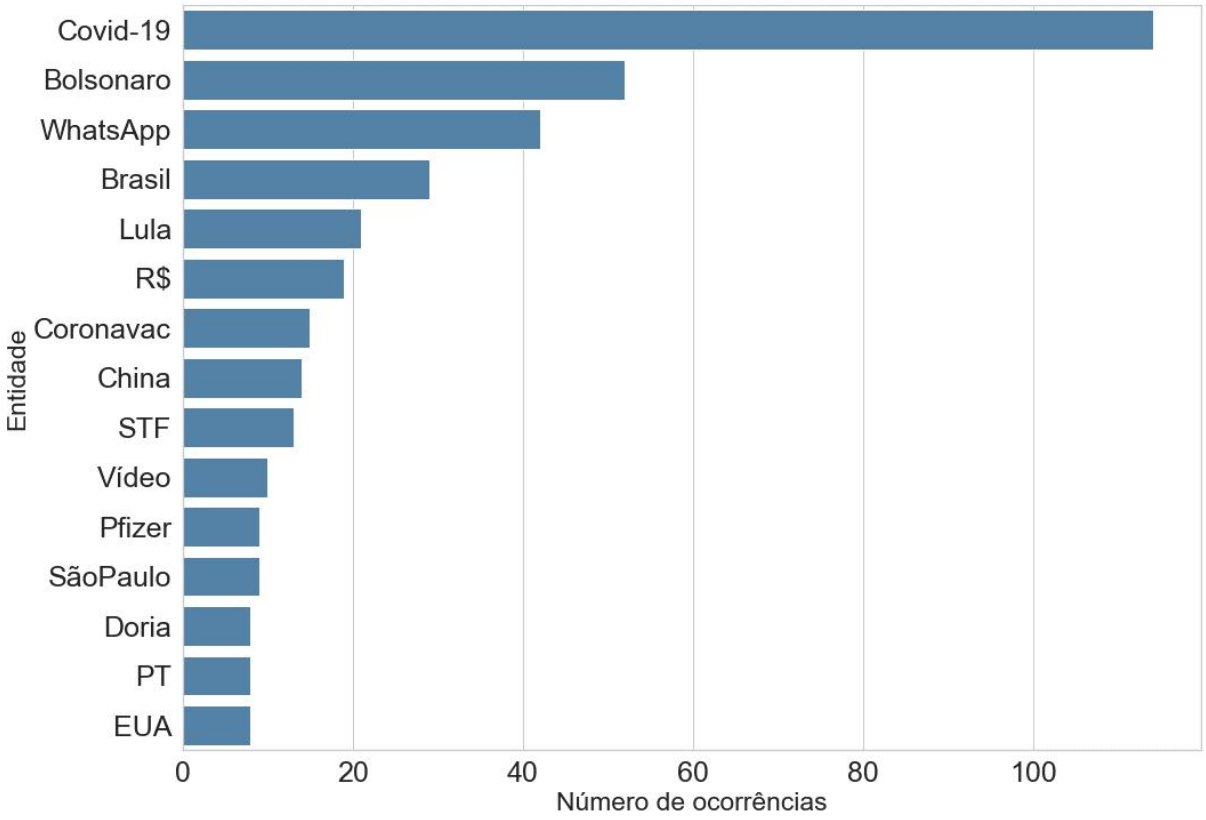
Gráfico 44: Distribuição geral das entidades nomeadas ao longo do ano de 2020



Fonte: Elaborado pelo autor (2023).

Por fim, considerando o ano de 2021 do *dataset* apresentado no Gráfico 45, temos que as principais entidades nomeadas são: “Covid-19” com 114 ocorrências, “Bolsonaro” com 52 ocorrências e “Whatsapp” com 42 ocorrências. Estas entidades corroboram as inferências realizadas na seção de análise de categorias, onde se atribui o fato do assunto saúde e o assunto política serem os mais frequentes ao longo do ano de 2021 devido à ocorrência da pandemia de COVID-19, ou seja, observando as principais notícias sobre os dois temas é possível perceber que muitas notícias falsas têm relação com medidas políticas de combate a pandemia, o que explica a grande quantidade de notícias atreladas a esses temas. Além disso, é importante mencionar que o fato do termo “Whatsapp” ser uma das entidades mais frequentes, novamente tem relação com o meio de disseminação de notícias falsas desse período, visto que em uma quantidade considerável das notícias analisadas observou-se que o meio de disseminação das mesmas ocorreu por meio do *Whatsapp*.

Gráfico 45: Distribuição geral das entidades nomeadas ao longo do ano de 2021



Fonte: Elaborado pelo autor (2023).

Portanto, foi possível compreender nessa seção a importância da análise temporal de entidades nomeadas na compreensão de padrões de assuntos abordados em notícias falsas, visto que é possível identificar quais são os entes mencionados, bem como o aspecto de abordagem dos mesmos nos conteúdos de desinformação disseminados.

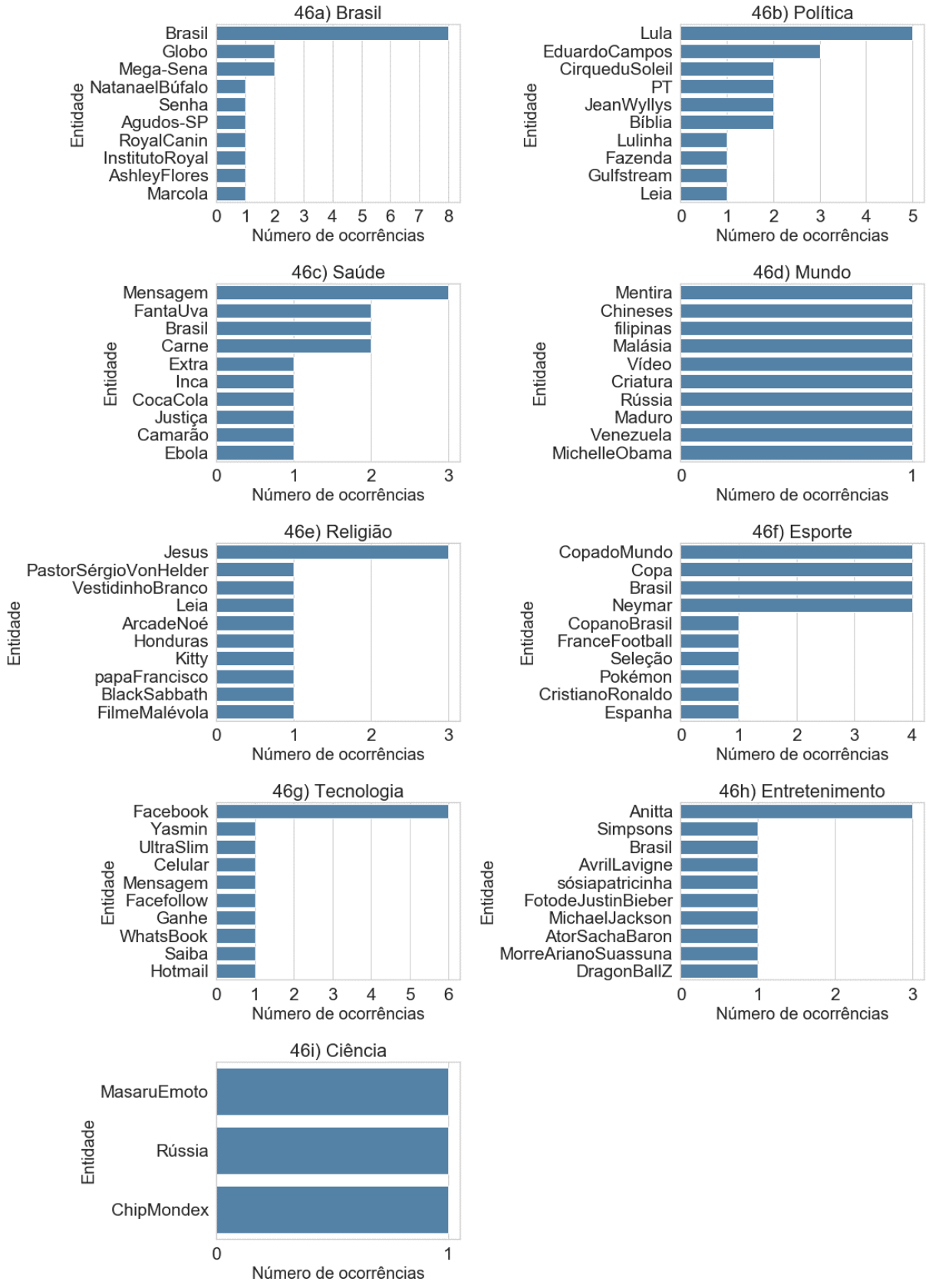
5.4 ANÁLISE TEMPORAL DAS ENTIDADES POR CATEGORIA

Nesta seção foi realizada uma análise temporal relativa às entidades nomeadas presentes em cada categoria de notícia separadamente com o objetivo de compreender as principais entidades de cada categoria de notícias falsas presentes no *corpus* ao longo do tempo. Portanto, o principal motivo dessa análise consiste em corroborar todas as inferências realizadas anteriormente e realizar a agregação das análises já realizadas com o intuito de confirmar os resultados já alcançados.

Ao considerar o triênio 2013-2015, foi analisado na seção 5.2 que as categorias mais frequentes estão relacionadas a assuntos nacionais, política e saúde. Ao observar no Gráfico 46, temos que na categoria sobre assuntos nacionais, a entidade

com maior número de menções nas notícias refere-se ao “Brasil”, seguido da entidade “Globo”. Em geral, ao observar a categoria política, “Lula” e “Eduardo Campos” figuram como as entidades mais citadas, sendo “Lula” citado em notícias falsas relativas à corrupção, e “Eduardo Campos” em notícias falsas disseminadas após o acidente fatal ocorrido com o candidato à presidência no ano de 2014. Adicionalmente, na categoria de notícias relacionadas à saúde existem referências à entidades, tais como: “Ebola”, “Inca”, “CocaCola”, etc. Porém, é importante ressaltar que em outras categorias existem entidades que se destacam por sua grande quantidade de ocorrências, tais como: entidade “Facebook” na categoria de tecnologia, entidade “CopadoMundo” na categoria de esporte, entidade “Jesus” na categoria religião. Ao observar essas entidades que se destacam, constata-se que o termo “Facebook”, em geral, aparece de maneira frequente como o meio de propagação das informações falsas. Além disso, temos “CopadoMundo” como entidade recorrente devido a Copa do Mundo FIFA de 2014. Portanto, as entidades mais recorrentes no triênio analisadas na seção 5.3, tem relação com as entidades analisadas por categoria na presente seção. Além disso, ao observar as principais entidades do triênio separadas por categoria, corrobora-se as inferências realizadas sobre os principais assuntos tratados.

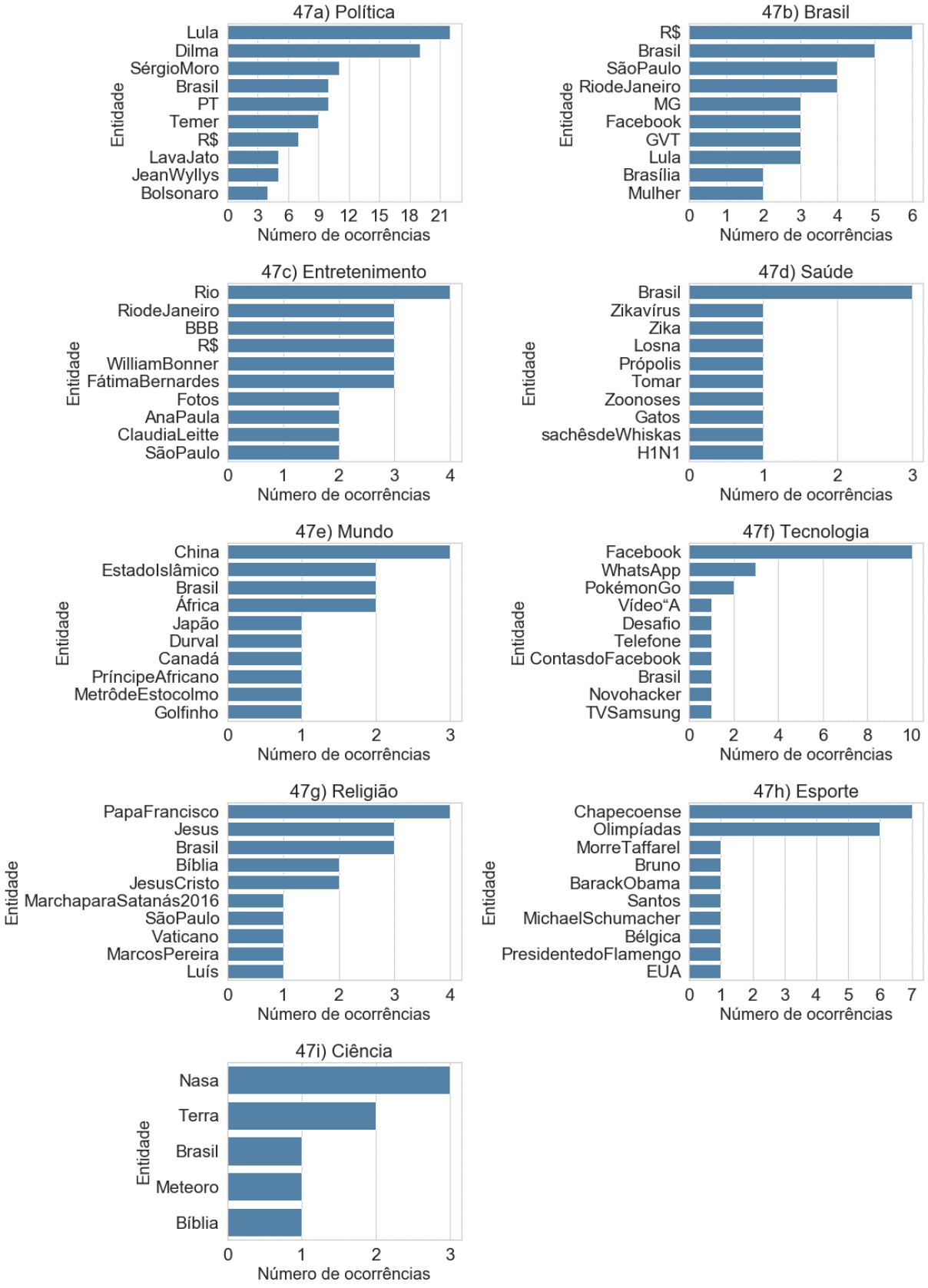
Gráfico 46: Distribuição de entidades por categoria nos anos (2013-2015). 46a) Brasil. 46b) Política. 46c) Saúde. 46d) Mundo. 46e) Religião. 46f) Esporte. 46g) Tecnologia. 46h) Entretenimento. 46i) Ciência.



Fonte: Elaborado pelo autor (2023).

Na seção 5.2 foi analisado que as categorias mais recorrentes no ano de 2016 estão relacionadas com política, assuntos nacionais e entretenimento. Ao observar a distribuição de entidades por categoria no Gráfico 47, constata-se que na categoria política existem três entidades muito recorrentes, isto é, as entidades “Lula”, “Dilma” e “SérgioMoro”. Nesse caso, ao verificar as principais notícias dessa categoria que mencionam as entidades supracitadas, os principais assuntos estão relacionados com *Impeachment*, corrupção e Operação Lava Jato. Portanto, o grande número de referências à essas entidades é explicável, visto que os assuntos inferidos anteriormente têm relação com as referidas entidades. Paralelamente, considerando as notícias mais importantes da categoria sobre assuntos nacionais, temos destaque para as entidades “Brasil” com 5 ocorrências, “SãoPaulo” com 4 ocorrências e “RiodeJaneiro” com 4 ocorrências, não havendo nenhum assunto predominante em relação a essas entidades. Por fim, a categoria de notícias sobre entretenimento traz muitas citações à entidade “Rio” com 4 ocorrências e a entidade “RiodeJaneiro” com 3 ocorrências, sendo a maior parte dessas ocorrências relacionadas à falsas notícias sobre a localidade do falecimento de famosos. Por conseguinte, é possível observar que a categoria mais importante no ano de 2016 está relacionada ao tema sobre política, e suas entidades supracitadas evidenciam os principais processos políticos em decorrência nesse período.

Gráfico 47: Distribuição de entidades por categoria no ano de 2016. 47a) Política. 47b) Brasil. 47c) Entretenimento. 47d) Saúde. 47e) Mundo. 47f) Tecnologia. 47g) Religião. 47h) Esporte. 47i) Ciência.

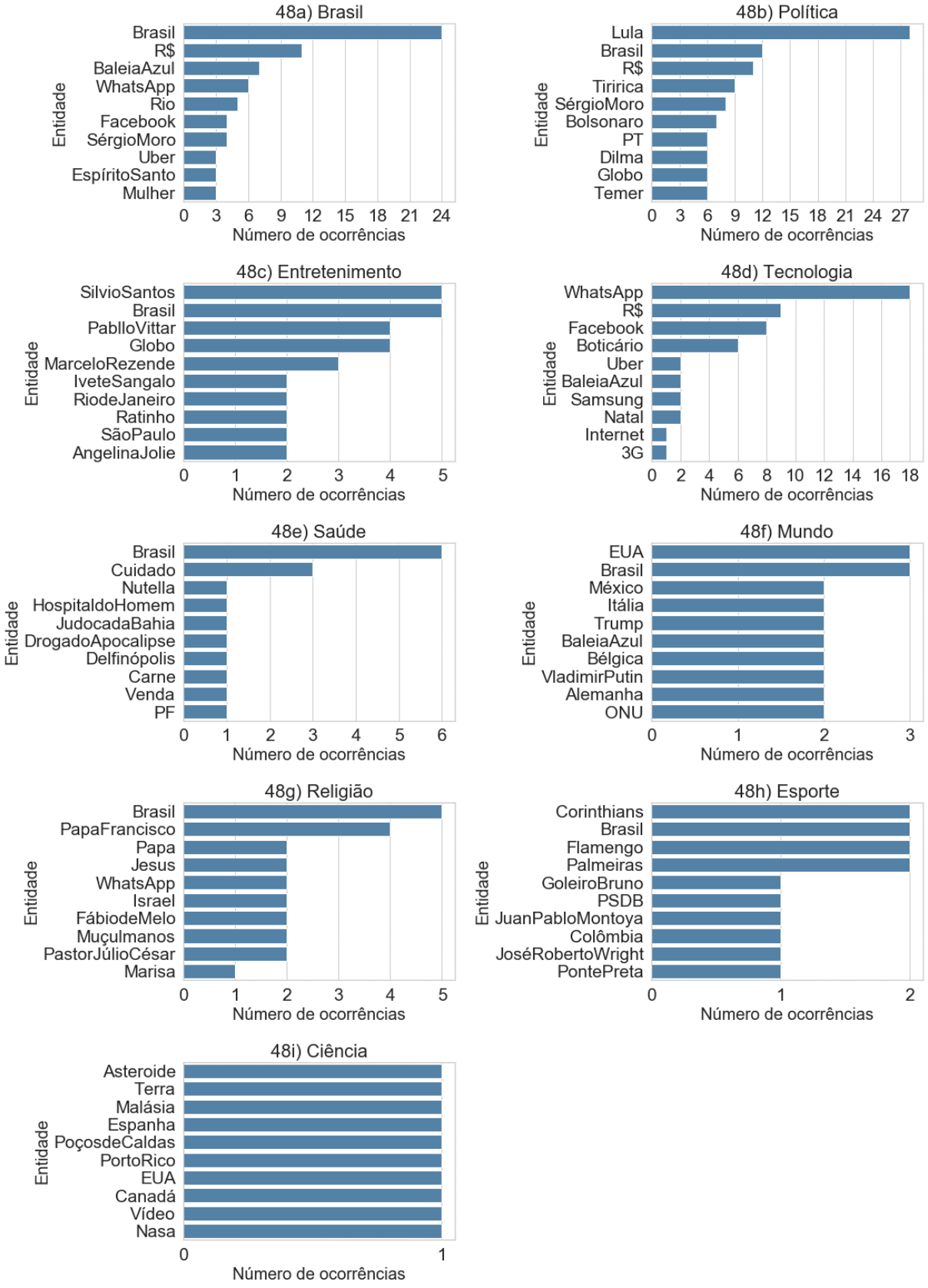


Fonte: Elaborado pelo autor (2023).

Foi analisado na seção 5.2 que as categorias mais recorrentes em 2017 foram relativas a assuntos nacionais e política, sendo que a temática sobre assuntos nacionais foi predominante. Observando a distribuição de entidades por categoria no Gráfico 48, constatou-se que na categoria sobre assuntos nacionais há entidades com maior destaque, entre elas, temos: “Brasil” com 24 ocorrências, “R$” com 11 ocorrências e “BaleiaAzul” com 7 ocorrências. Observando as principais notícias dessa categoria percebe-se que a entidade “Brasil” ocorre em notícias falsas que tratam sobre boatos variados, não havendo um assunto mais específico. Por outro lado, a entidade “R$” tem um número considerável de suas ocorrências ligadas a notícias falsas que tratam sobre assuntos ligados à corrupção. Além disso, a entidade “BaleiaAzul” tem relação com notícias sobre o suposto fenômeno[28] surgido na rede social russa, VK, ligado ao aumento de suicídios de adolescentes. Observando a categoria sobre política constata-se que as principais entidades envolvidas são: “Lula” com 28 ocorrências, “Brasil” com 12 ocorrências e “R$” com 11 ocorrências. Dentre as principais notícias sobre política temos: boatos variados sobre a entidade “Lula”, não havendo um assunto predominante nesses boatos; notícias falsas sobre corrupção com grande quantidade de referências a entidade “R$”; boatos diversos envolvendo a entidade “Brasil”. Portanto, novamente a entidade mais citada no ano entre as duas categorias de notícias com maior quantidade de notícias falsas refere-se a uma entidade política, o que evidencia o grande número de notícias falsas relacionadas ao tema.

[28] Disponível em: https://www.bbc.com/portuguese/internacional-39753889

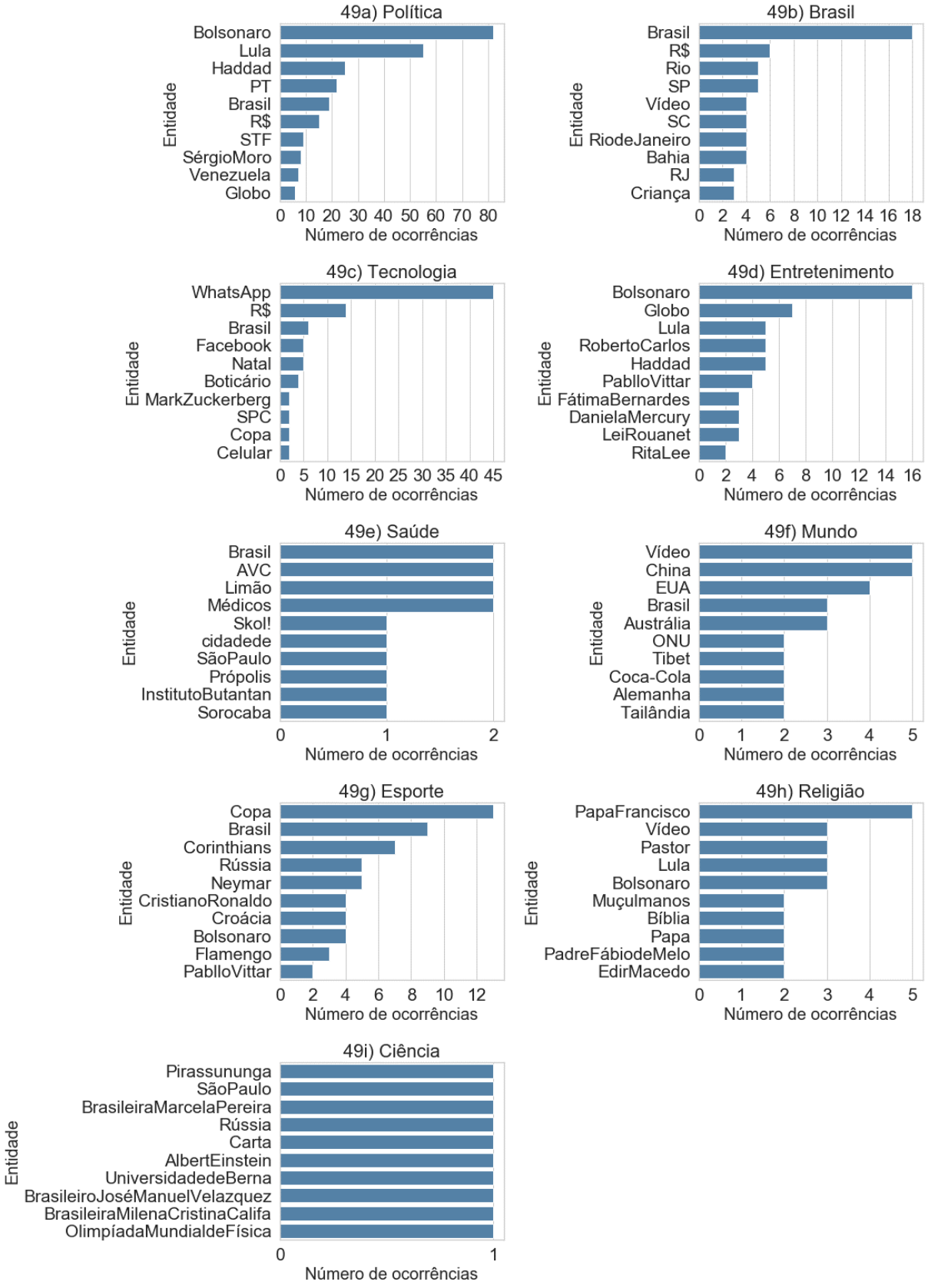
Gráfico 48: Distribuição de entidades por categoria no ano de 2017. 48a) Brasil. 48b) Política. 48c) Entretenimento. 48d) Tecnologia. 48e) Saúde. 48f) Mundo. 48g) Religião. 48h) Esporte. 48i) Ciência.



Fonte: Elaborado pelo autor (2023).

No ano de 2018, as categorias mais recorrentes novamente foram assuntos nacionais e política, com a particularidade da mudança da predominância, ou seja, nesse ano a política voltou a assumir a liderança em quantidade de notícias falsas disseminadas, conforme foi analisado na seção 5.2. Considerando a distribuição de entidades por categoria no Gráfico 49, é possível identificar que na categoria política as principais entidades com maior destaque são: “Bolsonaro” com 82 ocorrências, “Lula” com 55 ocorrências e “Haddad” com 25 ocorrências. Nesse contexto, ao observar as principais notícias sobre política nesse período, temos que a maior parte dessas notícias tratam sobre boatos sobre essas três entidades supracitadas, onde a maior parte dos boatos são acusações difamatórias. Dessa forma, compreende-se a disseminação de boatos difamatórios nesse período, como um fato que está atrelado principalmente a disputa presidencial ocorrida no ano de 2018, principalmente ao considerar as duas entidades que disputaram a eleição, isto é, “Bolsonaro” e “Haddad”. Além disso, ao observar as notícias sobre assuntos nacionais, é possível identificar que as entidades com maior destaque são: “Brasil” com 18 ocorrências, “R$” com 6 ocorrências, “Rio” e “SP”, ambos com 5 ocorrências. Nesse caso, novamente temos que a entidade “Brasil” é citada em notícias falsas com assuntos variados, não havendo nenhum assunto predominante ao qual seja citada. Além disso, ao observar as principais notícias foi constatado novamente que a entidade “R$” tem suas citações amplamente associadas com notícias falsas sobre corrupção que citam valores monetários. Por fim, “Rio” e “SP” aparecem em notícias falsas variadas, em outras palavras, não há um assunto predominante em relação as citações dessas duas entidades.

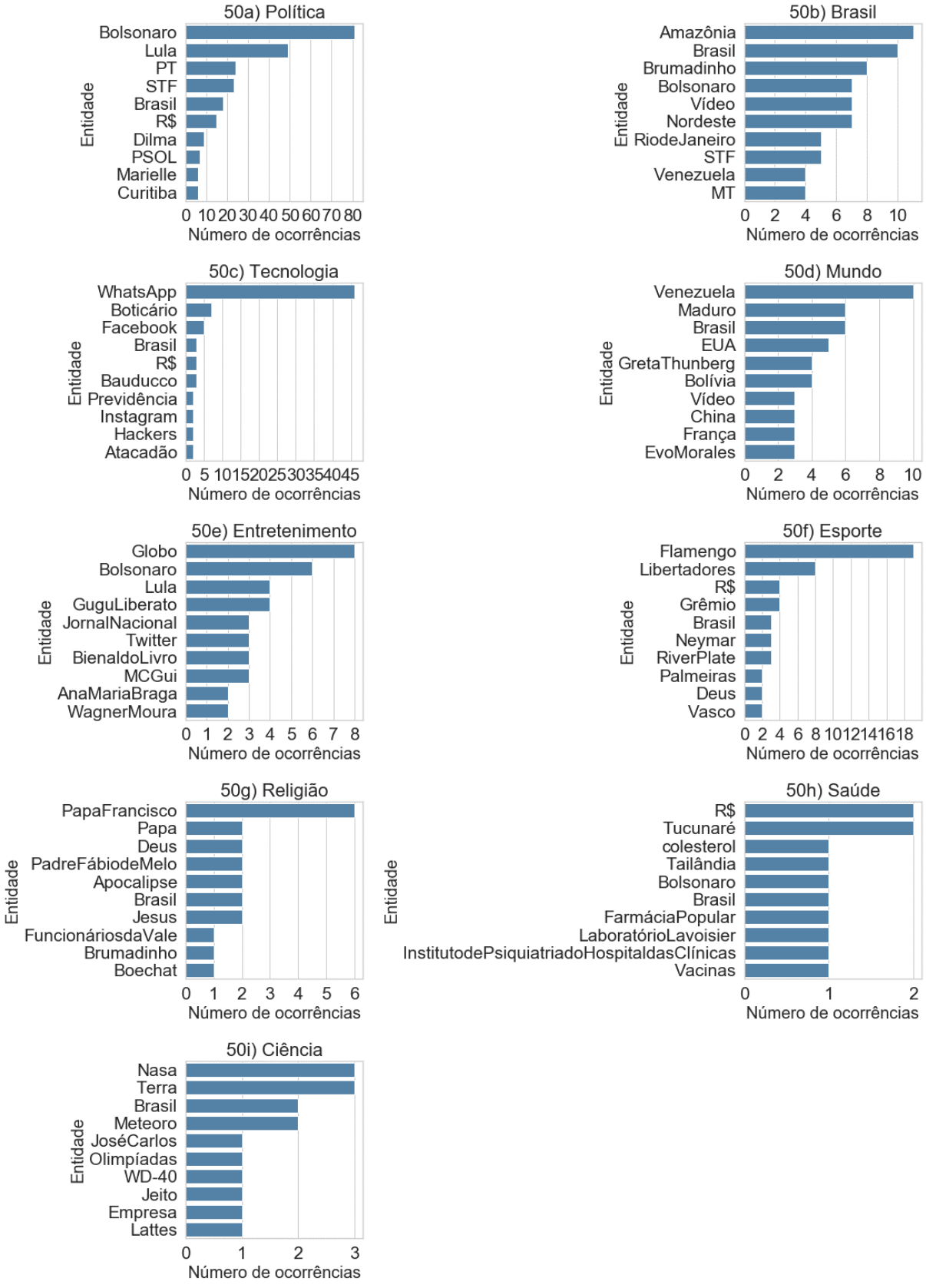
Gráfico 49: Distribuição de entidades por categoria no ano de 2018. 49a) Política. 49b) Brasil. 49c) Tecnologia. 49d) Entretenimento. 49e) Saúde. 49f) Mundo. 49g) Esporte. 49h) Religião. 49i) Ciência.



Fonte: Elaborado pelo autor (2023).

Ao considerar o ano de 2019, novamente as categorias predominantes estão relacionadas a política e assuntos nacionais, conforme foi visto na seção 5.2. Ao observar a distribuição de entidades por categoria no Gráfico 50, é possível constatar que nas notícias falsas relacionadas à política, tem-se que as principais entidades são: “Bolsonaro” com 81 ocorrências, “Lula” com 49 ocorrências, “PT” com 24 ocorrências e “STF” com 23 ocorrências. Em geral, a maior parte das notícias que fizeram referências à essas entidades tratavam de assuntos variados, tanto envolvendo boatos difamatórios, quanto envolvendo questões políticas em decorrência no período de 2019. Por outro lado, nas notícias falsas relacionadas a assuntos nacionais, temos a grande quantidade de ocorrências de entidades, tais como: “Amazônia” com 11 ocorrências, “Brasil” com 10 ocorrências e “Brumadinho” com 8 ocorrências. Nesse caso, a única entidade com assunto predominante se refere a “Brumadinho”, cuja ocorrência está exclusivamente atrelada à notícias falsas relacionadas ao rompimento de barragem em Brumadinho, onde a maior parte das notícias falsas estão ligadas a vídeos e fotos falsas sobre o ocorrido.

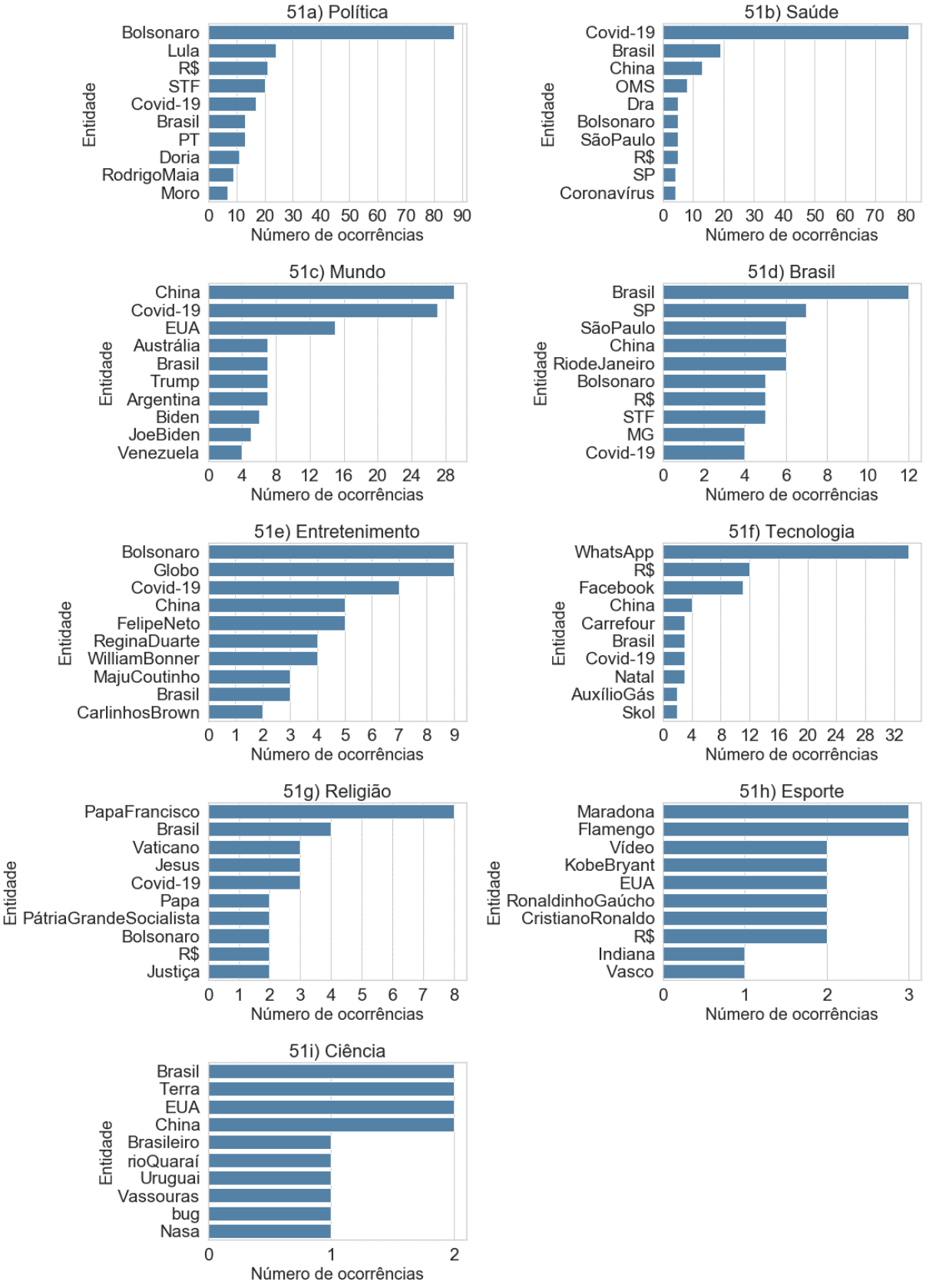
Gráfico 50: Distribuição de entidades por categoria no ano de 2019. 50a) Política. 50b) Brasil. 50c) Tecnologia. 50d) Mundo. 50e) Entretenimento. 50f) Esporte. 50g) Religião. 50h) Saúde. 50i) Ciência.



Fonte: Elaborado pelo autor (2023).

Observando o ano de 2020, as categorias predominantes estão relacionadas com política e saúde, conforme foi observado na seção 5.2. Ao observar a distribuição de entidades por categoria no Gráfico 51, percebe-se que as entidades citadas de forma mais recorrente nas notícias ligadas à política são: “Bolsonaro” com 87 ocorrências, “Lula” com 24 ocorrências, “R$” com 21 ocorrências, “STF” com 20 ocorrências e “Covid-19” com 17 ocorrências. Dessa forma, constata-se novamente que os nomes dos dois representantes dos polos ideológicos referentes ao tema sobre política surgem como os mais citados, além disso a permanência da entidade “R$” novamente tem suas citações atreladas à notícias falsas relacionadas com supostos casos de corrupção e, também é possível perceber a presença da entidade “Covid-19” que, em boa parte das notícias falsas, tem relação com falsas informações sobre medidas políticas adotadas no período de disseminação do vírus na pandemia de COVID-19. Analisando as entidades mais recorrentes nas notícias relacionadas com o tema sobre saúde, temos: “Covid-19” com 81 ocorrências, “Brasil” com 19 ocorrências e “China” com 13 ocorrências. Nesse caso, temos a entidade “Covid-19” liderando com predominância o ranking das entidades mais citadas, principalmente devido ao fato do surgimento da pandemia de COVID-19, que teve sua deflagração em grande escala no início de 2020. Além disso, temos a entidade “Brasil” sendo citada principalmente em notícias falsas relacionadas a movimentações políticas adotadas no país durante o início da pandemia. Por fim, a entidade “China” tem grande parte de suas citações realizadas em notícias com boatos sobre a origem do vírus SARS-CoV-2.

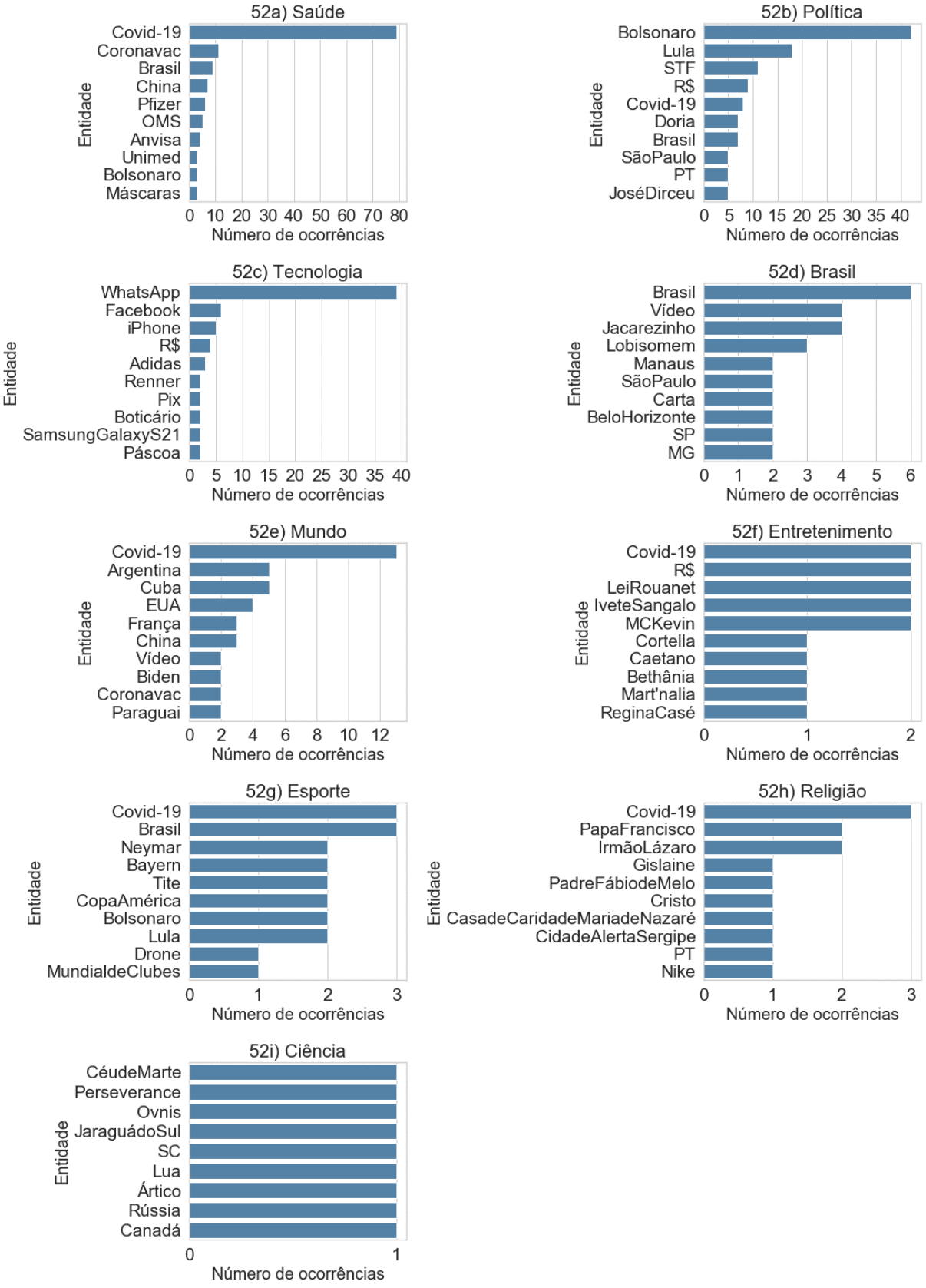
Gráfico 51: Distribuição de entidades por categoria no ano de 2020. 51a) Política. 51b) Saúde. 51c) Mundo. 51d) Brasil. 51e) Entretenimento. 51f) Tecnologia. 51g) Religião. 51h) Esporte. 51i) Ciência.



Fonte: Elaborado pelo autor (2023).

Analisando o ano de 2021 no Gráfico 52, foi possível observar que as notícias falsas continuaram a tratar de maneira mais recorrente sobre saúde e política e, além disso, o tema saúde se consolidou como o mais predominante entre os temas, conforme foi visto na seção 5.2. Com isso, ao observar as principais entidades citadas nas notícias sobre saúde, foi identificado que as principais entidades são: “Covid-19” com 79 ocorrências, “Coronavac” com 11 ocorrências e “Brasil” com 9 ocorrências. Desse modo, ao observar esse ano, é importante relembrar o fato de que as notícias foram coletadas até agosto na construção do *corpus*. Esse fato, por sua vez, demonstra que mesmo sem a coleta de todos os meses do ano, a entidade “Covid-19” continua tendo número de citações similares ao ano anterior, fato esse que traz uma magnitude da importância do tema no referido ano. Além disso, é importante ressaltar o aparecimento da entidade “Coronavac”, isto é, a vacina de combate a COVID-19. Nesse caso, temos que a entidade “Coronavac” é majoritariamente referenciada em notícias falsas relatando falsa ineficácia da vacina ou falsos efeitos adversos. Observando as entidades mais recorrentes nas notícias relacionadas com o tema sobre política, temos: “Bolsonaro” com 42 ocorrências, “Lula” com 18 ocorrências, “STF” com 11 ocorrências, “R$” com 9 ocorrências e “Covid-19” com 8 ocorrências. Portanto, ao analisar as notícias desse período, temos que as entidades mencionadas em notícias falsas estão majoritariamente relacionadas com o assunto sobre a pandemia de COVID-19. Dentre as notícias, ainda ocorre a predominância de boatos sobre medidas políticas adotadas no combate à pandemia de COVID-19.

Gráfico 52: Distribuição de entidades por categoria no ano de 2021. 52a) Saúde. 52b) Política. 52c) Tecnologia. 52d) Brasil. 52e) Mundo. 52f) Entretenimento. 52g) Esporte. 52h) Religião. 52i) Ciência.



Fonte: Elaborado pelo autor (2023).

6 DICIONÁRIO DE TÓPICOS VIA MODELAGEM DE TÓPICOS

O objetivo principal deste capítulo consiste em detalhar as etapas realizadas para a construção do dicionário de tópicos referentes às notícias do *corpus* estudado, com o uso de técnicas de modelagem de tópicos. Desse modo, neste capítulo serão descritas todas as técnicas utilizadas, bem como será detalhada a metodologia utilizada na construção do dicionário de tópicos.

Para a construção do dicionário foram realizadas modelagens de tópicos distintas fazendo uso de duas técnicas de modelagens de tópicos, isto é, *Latent Dirichlet Allocation* (LDA) e *Latent Semantic Analysis* (LSA). Nesse caso, o principal objetivo foi realizar uma abordagem com o uso da técnica LDA, e uma abordagem exclusiva com o uso da técnica LSA, com o objetivo de realizar uma comparação de resultados. Além disso, é importante ressaltar que foram gerados três dicionários de tópicos para cada uma das duas abordagens realizadas, totalizando seis dicionários de tópicos distintos. Portanto, para cada uma das abordagens existe um dicionário que caracteriza os tópicos do *corpus* por período de tempo (ano de publicação da notícia), existe outro dicionário que descreve os tópicos do *corpus* por categoria (assunto da notícia publicada), e por fim existe um último dicionário que identifica os tópicos do *corpus* utilizando todas as amostras existentes no *corpus*, ou seja, realizando uma modelagem geral das notícias falsas presentes no conjunto de dados.

Portanto, a organização deste capítulo abrange três seções, ou seja, a seção 6.1 traz a descrição do processo de seleção e preparação dos dados utilizados nas modelagens. Na seção 6.2 é explicada a abordagem via LDA com otimização via ajuste de hiperparâmetros, bem como os resultados obtidos. Na seção 6.3 é descrita a abordagem que utiliza o algoritmo LSA com otimização via ajuste de hiperparâmetros. Concluindo o capítulo, na seção 6.4 foi realizada uma breve discussão sobre os resultados alcançados de forma comparativa, finalizando a discussão. Os códigos referentes ao processo de criação do dicionário de tópicos via modelagem de tópicos, com as duas abordagens (LDA e LSA) e o dicionário de tópicos podem ser vistos no Apêndice E.

6.1 SELEÇÃO E PREPARAÇÃO DOS DADOS

O principal objetivo da etapa de seleção e preparação dos dados para este trabalho consistiu em separar o *corpus* em conjuntos de dados individualizados para construção dos dicionários de tópicos, e além disso, reduzir ao máximo as informações irrelevantes contidas nos textos das notícias utilizadas na modelagem de tópicos, tais como: *stopwords*, *links*, *hashtags*, números, entre outros. Portanto, é importante ressaltar que essa etapa é comum para as duas abordagens, isto é, os dados resultantes dessa etapa foram utilizados como entrada nas duas abordagens descritas nas seções a seguir.

De forma a definir os conjuntos de dados individualizados que seriam considerados na modelagem, foram realizadas inicialmente as divisões do *corpus* por período de tempo. Nesse caso, as divisões foram realizadas considerando o triênio 2013-2015 como um único conjunto, principalmente devido ao fato desse período possuir uma quantidade menor de notícias em relação aos outros anos considerados individualmente, conforme foi explicado na seção 5.1. Após a separação do triênio, os anos posteriores, isto é, os anos de 2016 a 2021, foram considerados em conjuntos individuais. Após a separação temporal, foi realizada a separação categórica, ou seja, o *corpus* foi separado em conjuntos de dados menores de acordo com a categoria de notícia (tema do assunto da notícia). Por conseguinte, foram consideradas as seguintes categorias para a separação dos dados: política, assuntos nacionais, assuntos internacionais, religião, ciência, entretenimento, esporte, saúde e tecnologia. Portanto, cada uma das categorias supracitadas foi separada em um conjunto de dados individual, com exceção das categorias lista e opinião, que não foram utilizadas na modelagem, visto que somadas as duas categorias havia uma quantidade muito pequena de notícias que, por sua vez, não trariam informações relevantes para a caracterização do *corpus* via modelagem de tópicos. Finalizando a seleção dos dados, também foi realizada a separação de todo o *corpus* para a modelagem geral.

Além da seleção e separação dos dados, também é importante ressaltar o pré-processamento realizado nos documentos. Nesse sentido, temos que os documentos utilizados para a modelagem de tópicos foram armazenados na coluna "*message_norm_treatment_ssw2*" que, por sua vez, foi descrita na seção 4.3. Nesse caso, a coluna referida possui todo o tratamento realizado no *corpus* e descrito na seção 4.3 e, além disso, possui a remoção de *stopwords* mais avançadas. Após isto,

os documentos foram *tokenizados*, isto é, os mesmos foram particionados em listas de termos, denominados *tokens*, a partir das *strings* que representavam os textos por completo. Nesta etapa, acentuações também foram removidas. Para o processo de *tokenização*, a biblioteca para modelagem de tópicos *gensim* foi utilizada, mais especificamente sua função *simple_preprocess* (ŘEHŮŘEK; SOJKA, 2010). Após isso, foi realizado o tratamento em relação aos bigramas e trigramas presentes nos dados. Para isso, foram utilizados recursos da biblioteca *gensim*. Por fim, foi realizado o processo de lematização nos dados finais. O processo de lematização foi utilizado visto que traz benefícios para a modelagem por meio do método LDA conforme pode ser visto em (MAY; COTTERELL; DURME, 2019), isto é, o estudo mostra que a lematização traz benefícios relativos à modelagem de tópicos em linguagens morfologicamente ricas. Além disso, ao considerar o método LSA, temos que para sua execução, processar um *corpus* lematizado seria computacionalmente mais eficiente conforme pode ser visto em (ZIPITRIA; ARRUARTE; ELORRIAGA, 2006).

6.2 ABORDAGEM VIA LDA OTIMIZADA COM AJUSTE DE HIPERPARÂMETROS

Esta seção traz tanto a descrição dos principais conceitos e técnicas utilizadas na modelagem de tópicos desenvolvida com a otimização dos hiperparâmetros referentes a modelagem, quanto as análises e resultados alcançados. Nesse sentido, a subseção 6.2.1 traz a definição dos principais conceitos referentes ao modelo utilizado, bem como a métrica utilizada como parâmetro para otimização do modelo. Na subseção 6.2.2 é descrita a abordagem construída no presente trabalho, bem como todo o processo utilizado na modelagem de tópicos. A subseção 6.2.3 descreve a análise e os resultados alcançados na modelagem de tópicos geral, isto é, considerando todo o *corpus* de estudo. Na subseção 6.2.4 é descrita a análise e resultados alcançados na modelagem temporal. Por fim, a subseção 6.2.5 descreve a análise e resultados da modelagem de tópicos categórica, em outras palavras, considerando os assuntos das notícias do *corpus* de estudo.

6.2.1 Conceitos Fundamentais

Para introduzir o processo de modelagem de tópicos proposto e realizado no presente trabalho, é importante descrever alguns conceitos fundamentais que foram utilizados no processo. Inicialmente é importante destacar como o algoritmo *Latent*

Dirichlet Allocation e seus parâmetros e hiperparâmetros funcionam. Desse modo, é importante ressaltar que segundo (BOYD-GRABER et al., 2017) o processo generativo do LDA pode ser resumido em três etapas, entre elas: **geração dos tópicos**, **alocação de tópicos para documentos** e **geração de palavras**.

- Para **geração de tópicos**, existe a necessidade de inicialmente definir o número de tópicos distintos no *corpus*. Nesse sentido, existe um hiperparâmetro no modelo denominado K que, por sua vez, define o número de tópicos. Após definir K, é importante saber que cada um dos K tópicos é definido por uma distribuição Dirichlet representada por $\varphi_k \sim Dir(\beta)$. Nesse caso, β denomina-se parâmetro de concentração que, por sua vez, também é um hiperparâmetro no modelo. Desse modo, a distribuição φ_k que define o tópico contém pesos (probabilidades) designadas a cada um dos termos do vocabulário;
- Para a etapa de **alocação de tópicos** é necessário determinar os tópicos que compõem o documento do *corpus* estudo. Dessa forma, esse processo busca determinar a distribuição de tópicos para os documentos. Para isso, cada um dos documentos possui uma distribuição sobre todos os tópicos, com o intuito de caracterizar o conteúdo do mesmo. Essa distribuição, por sua vez, também é uma distribuição de Dirichlet definida como $\theta_d \sim Dir(\alpha)$, sua função é definir as atribuições dos tópicos entre os documentos do *corpus*. Portanto, é importante destacar o parâmetro α, que também é um hiperparâmetro do modelo, responsável por controlar a esparsidade da distribuição;
- Por fim, a etapa de **geração de palavras** trata cada documento d do *corpus* como um conjunto definido por N_d palavras. Com isso, para cada palavra n do documento em questão, atribui-se um tópico $z_{d,n} \sim Discrete(\theta_d)$. Nesse caso, o tópico escolhe a palavra para compor o documento. Finalmente, escolhe-se a palavra $w_{d,n} \sim \varphi_{z_{d,n}}$, por meio da atribuição de tópico.

Portanto, é possível identificar que a abordagem generativa é útil para a partir de documentos, obter os tópicos que descrevem os temas presentes nos mesmos. Esse processo, por sua vez, realiza a inferência dos tópicos a partir do *corpus* estudado, identificando os tópicos que melhor definem os documentos do *corpus*.

Portanto, a partir da hipótese generativa o LDA trabalha realizando a inferência dos tópicos que formam o *corpus* analisado em questão.

Além dos parâmetros e hiperparâmetros relativos à construção da modelagem de tópicos via *Latent Dirichlet Allocation*, é importante tratar sobre o conceito relativo à métrica de coerência para avaliação da modelagem. Nesse sentido, temos que a coerência consiste em um valor que indica quantitativamente a ocorrência mútua entre os termos associados ao tópico estudado que, por sua vez, se traduz em um valor que indica o pertencimento dos termos a um mesmo tema. Em (RÖDER; BOTH; HINNEBURG, 2015), é possível constatar que a métrica de coerência tem correlação com resultados advindos de observações humanas, isto é, ao considerar o julgamento humano para determinar a qualidade dos tópicos, temos que a coerência indica resultados aproximadamente semelhantes.

Portanto, é importante destacar que tópicos onde os termos coexistem com grande frequência possuem valores de coerência elevados, enquanto tópicos que possuem termos com baixo grau de coexistência, possuem valores de coerência baixos. Frequentemente, essa métrica é calculada utilizando os termos mais relevantes do tópico estudado, e utiliza uma métrica de confirmação para cada combinação de pares de palavras associados ao tópico. Finalmente, os cálculos para cada par são utilizados para o cálculo do valor final da coerência do tópico estudado. A métrica de confirmação calculada para cada par tem diferentes variações, entre elas temos a variação denominada C_V (RÖDER; BOTH; HINNEBURG, 2015) que foi escolhida para a avaliação de qualidade dos modelos construídos no presente trabalho.

Especificamente, o valor de coerência C_V é o resultado de uma combinação descoberta por um estudo sistemático do espaço de configuração das medidas de coerência. Esta medida (C_V) combina a medida indireta do cosseno com a *Normalized Pointwise Mutual Information* (NPMI) e a janela deslizante booleana.

No presente trabalho, foi realizada a implementação da pipeline de coerência descrita acima por meio da biblioteca *gensim* que, por sua vez, foi empregada para o cálculo da métrica. A principal classe utilizada *CoherenceModel* presente na biblioteca *gensim* consolidou a implementação. Além disso, temos que essa classe possui diferentes métricas de confirmação além do C_V. Para cálculo da coerência para os modelos treinados, foi necessário utilizar os tópicos gerados pelo modelo, o conjunto de documentos de validação criado especificamente para este passo, o dicionário de

palavras do vocabulário e também o número de palavras a serem consideradas como mais relevantes dentro do tópico.

6.2.2 Modelagem de Tópicos e Otimização de Hiperparâmetros

Na abordagem descrita nesta seção foi utilizada a técnica *Latent Dirichlet Allocation* (LDA), onde toda a implementação do algoritmo foi realizada via Python com o uso da biblioteca *gensim*. Utilizando os dados resultantes do processo descrito na seção 6.1, foi criado o dicionário *id2word* com o uso da classe *Dictionary* da biblioteca *gensim* que, por sua vez, mapeia o identificador de uma palavra com seu respectivo *token*, e além disso foi criada a entrada do *corpus* para a utilização na modelagem.

Após isso, foi construído o modelo-base utilizado para experimentação inicial. Nesse sentido, o primeiro fator a ser considerado no treinamento foi o número K de tópicos, visto que esse é o principal hiperparâmetro para o treinamento do modelo de tópicos. Para isso, foram analisadas as principais características conhecidas do *dataset*, isto é, número de categorias (11), quantidade temporal (anos) presentes no *dataset* (9) e quantidade de notícias presentes no *dataset* (5201). Nesse sentido, ao considerar essas informações optou-se por considerar $K = 10$, como uma primeira estimativa para a quantidade de tópicos. O principal motivo consiste no fato de que tanto a divisão categórica, quanto a divisão temporal se aproxima deste valor, sendo esse K, um valor próximo dessas duas métricas. Além disso, outros hiperparâmetros também foram definidos no treinamento do modelo, entre eles:

- **random_state** (Valor = 100) - Usado para definir a semente para o gerador aleatório para que seja possível garantir que os resultados obtidos possam ser reproduzidos.
- **chunksize** (Valor = 100) - É o número de documentos a ser considerado de uma só vez na modelagem de tópicos. É importante ressaltar que esse hiperparâmetro afeta diretamente o consumo de memória.
- **passes** (Valor = 10) – Define quantas vezes o algoritmo deve ser executado por todo o *corpus*.
- **per_word_topics** (Valor = True) – Assumindo o valor True, temos que esse hiperparâmetro permite a extração dos tópicos mais prováveis de acordo com uma palavra. Dessa forma, o processo de treinamento é

definido de forma que cada palavra seja atribuída a um tópico. Caso contrário, palavras que não são indicativas serão omitidas.

Desse modo, para cada um dos três dicionários de tópicos resultantes, temos que os scripts de treinamento possuem código semelhante. Por conseguinte, os modelos são treinados com o conjunto de dados específico de acordo com a separação detalhada na seção 6.1, e após isso, o valor da métrica de coerência de tópicos é calculado a partir dos dados de validação. Adicionalmente, ao finalizar cada treinamento são obtidos os tópicos apresentados por listas de palavras, onde cada lista representa um tópico. Consequentemente, os dez termos mais relevantes de cada tópico são considerados nesta etapa. Além disso, temos que as probabilidades dos termos dentro de cada tópico também são extraídas neste momento.

Após realizar o cálculo da coerência para o modelo treinado, e após obter a modelagem representada pelas listas de palavras, é possível obter uma primeira visão sobre a qualidade da modelagem. Consequentemente, após isso é possível realizar a otimização dos hiperparâmetros com o objetivo de aumentar a qualidade da modelagem. Nesse sentido, os hiperparâmetros considerados para a otimização do modelo são:

- Número de tópicos (K) – Determina a quantidade de tópicos obtidos na modelagem.
- Hiperparâmetro alfa do modelo de Dirichlet (α) – Representa a densidade documento-tópico.
- Hiperparâmetro beta do modelo de Dirichlet (β) – Equivale a densidade palavra-tópico.

Dessa forma, a ideia principal é realizar o ajuste desses hiperparâmetros performando testes em sequência, isto é, considerando um hiperparâmetro por vez. Para isso, os outros hiperparâmetros são mantidos constantes, e com isso são executados os testes para o hiperparâmetro que estiver sendo testado e, além disso, os testes são feitos considerando dois conjuntos de validação do *corpus*. Um dos conjuntos de validação consiste em 75% dos dados do *corpus* utilizado, e o outro consiste em 100% dos dados do *corpus* utilizado.

A métrica utilizada para comparar a performance dos hiperparâmetros testados consiste no valor de coerência (C_V**)**. Desse modo, foram criadas duas funções para a otimização dos modelos de acordo com o ajuste dos hiperparâmetros, ou seja, uma função auxiliar para realizar o cálculo da métrica de coerência e a função principal responsável por realizar os testes para cada hiperparâmetro por meio de loop. A

função auxiliar foi criada para calcular os valores de coerência que, por sua vez, recebe como hiperparâmetros: o *corpus*, o dicionário *id2word*, o valor K de tópicos, o hiperparâmetro α e o hiperparâmetro β. Portanto, a função trabalha criando o modelo LDA utilizando a função externa *LdaMulticore* da biblioteca *gensim*, com o uso das entradas supracitadas e, por fim, a função utiliza a classe *CoherenceModel* para calcular os valores de coerência, utilizando como entrada o modelo LDA criado anteriormente.

A função principal possui quatro laços de repetição, isto é: O primeiro laço percorre o *corpus* utilizado para modelagem; O segundo laço itera sobre os valores de K que, por sua vez, podem ser intervalos passados como entrada da função; O terceiro laço é responsável pelo valor de α; Por fim, temos que o quarto laço é responsável pelo valor de β. Portanto, a cada repetição dos laços da função principal, é executada a função auxiliar para calcular os valores de coerência considerando a configuração de todos os hiperparâmetros envolvidos e, por consequência, esses valores são armazenados em um *Dataframe* que, posteriormente é convertido para o formato *Comma-separated values* (CSV).

Com isso, os dados gerados e armazenados no arquivo em formato CSV são usados para a otimização dos modelos de acordo com as configurações dos hiperparâmetros. De forma mais específica, inicialmente é realizada uma estimativa para o número ótimo de tópicos. Para isso, são fixados os hiperparâmetros α e β, ou seja, são atribuídos valores fixos para os valores desses hiperparâmetros de forma a identificar a melhor quantidade de tópicos K para esses valores fixados. Como valor fixado para os hiperparâmetros α e β foram considerados os resultados encontrados em (WALLACH; MIMNO; MCCALLUM, 2009), isto é, segundo esse trabalho ao realizar simulações com o método *Markov Chain Monte Carlo* (MCMC) foi possível constatar que a utilização de uma distribuição prior de Dirichlet assimétrica e hierárquica na distribuição documento-tópico (α), e uma distribuição prior de Dirichlet simétrica na distribuição tópico-palavra (β) resulta em modelos significativamente melhores, tanto considerando a performance, quanto considerando a probabilidade de reter documentos, e também na qualidade dos tópicos inferidos. Portanto, os valores fixados de α e β são: assimétrico e simétrico, respectivamente.

Após a obtenção do valor ótimo de K para os hiperparâmetros α e β fixados, é então realizada a filtragem dos valores de otimização para o valor de K ótimo

encontrado, de forma a obter dessa vez os valores de α e β ótimos para o K ótimo, que não necessariamente são os fixados na etapa anterior. Finalmente, após essa etapa, temos que o modelo é treinado com todos os hiperparâmetros encontrados.

É importante destacar que mesmo com a otimização dos modelos via ajuste de hiperparâmetros, também foi realizada uma análise manual adicional dos tópicos de forma a observar a presença de problemas, tais como: *junk topics*, palavras com baixa probabilidade, repetição de tópicos, entre outros. Nos casos onde foram identificados esses problemas, optou-se por realizar a diminuição unitária iterativa manual do intervalo de tópicos a ser considerado na otimização dos modelos via ajuste de hiperparâmetros, de forma a realizar nova otimização dos modelos até a resolução do problema, isto é, obtenção da modelagem de tópicos adequada. Em geral, a solução via diminuição do intervalo de tópicos pode ser justificada ao considerar (GAN J, 2021), onde é constatado que grandes números de tópicos no modelo de teste implicam em duplicação de tópicos, e consequentemente em problemas semelhantes aos citados anteriormente. Além disso, também é constatado no trabalho que ao considerar um conjunto de dados de amostragem menor que o *corpus* completo, é provável que em alguma amostragem, a modelagem no conjunto de dados de treino resulte em tópicos que não correspondem ao conjunto de dados de teste. Portanto, a diminuição do número de tópicos atua como uma ferramenta para a prevenção desses problemas. Como exemplo, se após otimização for obtido $K = n, n \in [4, 20]$ tópicos como número ótimo de tópicos para a modelagem, e com esse valor de tópicos a modelagem apresentar os problemas supracitados, então a ideia consiste em diminuir o intervalo de $[4, 20]$ para $[4, n-1], n-1 \in [4, 20]$. Dessa forma, evita-se a persistência dos problemas supracitados, visto que com um intervalo menor de tópicos existe menor probabilidade de obter tópicos com baixa qualidade.

6.2.3 Dicionário de Tópicos Geral

Esta subseção traz a descrição da aplicação da abordagem já descrita e os resultados obtidos, de forma geral. Desse modo, a descrição será realizada de forma a abordar a modelagem de tópicos para todos os documentos presentes no *corpus* de estudo. Ao considerar todos os documentos do *corpus* é possível ter uma quantidade mais completa de dados para execução do modelo, o que pode se traduzir em uma modelagem com tópicos mais definidos. Entretanto, como os assuntos de cada notícia

podem variar muito, podem haver tópicos que agregam termos relativos a diferentes assuntos de diferentes categorias de notícia, principalmente devido ao grande número de assuntos diferentes presentes no *corpus*.

Inicialmente foi realizada a modelagem do *corpus* geral utilizando o número de tópicos fixo com $K = 10$, sem ajuste dos hiperparâmetros. Além disso, também foi calculada a coerência para essa modelagem inicial. O principal intuito desse experimento é obter o valor de coerência para comparação com o modelo gerado após a otimização via ajuste de hiperparâmetros com o objetivo de verificar se houve melhora com o método de trabalho utilizado. Como o objetivo da modelagem é apresentar os resultados otimizados, os experimentos iniciais encontram-se no Apêndice E.

Levando em conta o processo detalhado na subseção 6.2.2, foi realizada a modelagem para o *corpus* completo, isto é, considerando todos os documentos (notícias falsas) para a modelagem. Dessa forma, os valores de coerência considerando: conjunto de validação, K no intervalo [4, 20], α e β podem ser vistos no Apêndice F. Utilizando, os valores fixados para α e β, foi encontrado o valor de tópicos ótimo como $K = 5$. Após isso, realizando a otimização dos modelos ajustando os hiperparâmetros de forma automatizada e também por meio da análise manual foram encontrados os seguintes valores para os hiperparâmetros de treino:

- $K = 5$
- α = 0.31
- β = 0.01

Do ponto de vista quantitativo, ao considerar a primeira modelagem sem a realização da otimização foi obtido 0,5467 como valor de coerência. Consequentemente, após a otimização foi obtido 0,6136 como valor de coerência para o modelo treinado. Portanto, constata-se um percentual de melhora de aproximadamente 12,23%, o que demonstra os resultados positivos da otimização proposta e realizada no presente trabalho.

Após a definição da quantidade de tópicos, a etapa seguinte correspondeu ao treinamento e execução do algoritmo LDA com os hiperparâmetros ótimos encontrados, onde foi possível verificar os tópicos gerados com o respectivo valor de probabilidade atribuído a cada termo, como pode ser visto no Quadro 3.

Quadro 3 - Modelagem de tópicos geral do corpus via LDA

Tópico (0 até K-1)
'0.164*"causa" + 0.126*"ganhar" + 0.090*"anunciar" + 0.062*"lancar" + ' '0.055*"acabar" + 0.054*"sair" + 0.050*"mundo" + 0.040*"prova" + ' '0.027*"governador" + 0.026*"livro"'
'0.282*"mostrar" + 0.144*"foto" + 0.130*"vacino" + 0.067*"morrer" + ' '0.045*"Video" + 0.039*"vazar" + 0.029*"facebook" + 0.027*"deixar" + ' '0.023*"prender" + 0.022*"encontrar"'
'0.130*"presidente" + 0.107*"grande" + 0.070*"ficar" + 0.070*"mulher" + ' '0.061*"carro" + 0.049*"aparecer" + 0.049*"homem" + 0.046*"cidade" + ' '0.039*"eleicoes" + 0.032*"brasileiro"'
'0.832*"novo" + 0.027*"falso" + 0.023*"pessoa" + 0.019*"pedir" + 0.011*"dia" ' '+ 0.010*"apontar" + 0.008*"virus" + 0.006*"site" + 0.005*"telefone" + ' '0.005*"enganar"'
'0.294*"video" + 0.089*"policia" + 0.074*"matar" + 0.064*"globo" + ' '0.052*"retirar" + 0.044*"tomar" + 0.043*"aprovar" + 0.027*"pagar" + ' '0.026*"ano" + 0.025*"chinês"'

Fonte: Elaborado pelo autor (2023).

Analisando qualitativamente cada tópico, é possível obter uma visão geral dos principais assuntos tratados nas notícias falsas presentes no *corpus*. A seguir são analisados os tópicos T_0 a T_4.

- T_0: 0.164 * **causa** + 0.126 * **ganhar** + 0.090 * **anunciar** + 0.062 * **lancar** + 0.055 * **acabar** + 0.054 * **sair** + 0.050 * **mundo** + 0.040 * **prova** + 0.027 * **governador** + 0.026 * **livro** - esse tópico é caracterizado pelas palavras mais importantes, entre elas: “causa”, “ganhar”, “anunciar” e “lancar”, o que remete as notícias falsas presentes no *corpus* que tratam de eventos e produtos que geralmente causam grande interesse público, tais como: promoções, sorteios, concursos, com o objetivo de obter benefícios financeiros, enganando as pessoas. Em geral, as principais notícias falsas presentes no *corpus* tratando sobre esse assunto estão relacionadas com a categoria sobre tecnologia. Portanto, é possível afirmar que esse tópico reflete notícias falsas associadas à categoria sobre tecnologia presente no *corpus*.
- T_1: 0.282 * **mostrar** + 0.144 * **foto** + 0.130 * **vacino** + 0.067 * **morrer** + 0.045 * **video** + 0.039 * **vazar** + 0.029 * **facebook** + 0.027 * **deixar** + 0.023 * **prender** + 0.022 * **encontrar** - esse tópico é caracterizado por termos, tais como: “mostrar”, “foto”, “vacino” (versão lematizada de vacina) e “morrer”. Em geral, as principais notícias falsas do *corpus* utilizando esses termos trazem como assunto principal, boatos sobre a eficácia das vacinas contra doenças, incluindo a COVID-19, com o objetivo de semear a desconfiança nas vacinas e espalhar informações errôneas sobre os casos fatais da doença. Além disso, foi possível encontrar notícias falsas nesse período que associam de maneira falsa a

aplicação da vacina à possíveis casos de óbito. Portanto, de maneira geral esse tópico tem uma associação direta com notícias pertencentes a categoria de notícias relativa à saúde.

• T_2: 0.130 * **presidente** + 0.107 * **grande** + 0.070 * **ficar** + 0.070 * **mulher** + 0.061 * **carro** + 0.049 * **aparecer** + 0.049 * **homem** + 0.046 * **cidade** + 0.039 * **eleicoes** + 0.032 * **brasileiro** – esse tópico é caracterizado pelas palavras “presidente”, “grande”, “mulher” e “carro”. Porém, é importante destacar os seguintes termos: “presidente”, “eleicoes” e “brasileiro” que, em geral, ocorrem nas principais notícias falsas sobre política do *corpus*, e trazem boatos envolvendo figuras políticas envolvidas na disputa presidencial ocorrida em 2018, frequentemente com o objetivo de prejudicar suas reputações e influenciar as eleições. Portanto, temos nesse tópico termos que trazem associações à categoria de notícias ligadas à política.

• T_3: 0.832 * **novo** + 0.027 * **falso** + 0.023 * **pessoa** + 0.019 * **pedir** + 0.011 * **dia** + 0.010 * **apontar** + 0.008 * **virus** + 0.006 * **site** + 0.005 * **telefone** + 0.005 * **enganar** – esse tópico é caracterizado por termos, tais como: "novo", “falso”, “enganar”, “vírus”, “site” e “telefone”. À vista disso, é possível afirmar que esse é um tópico com assunto bem definido, visto que está relacionado com notícias falsas sobre novas tecnologias, golpes cibernéticos, fraudes bancárias, explorando a falta de conhecimento das pessoas em relação a assuntos novos e tendências. Como exemplo, podemos citar notícias falsas presentes no *corpus*, tais como: “*SIM Swap Fraud é um novo golpe em que bandidos pedem que você pressione a tecla 1 no celular*” e “*Hackers fazem ligação do seu próprio número de celular para clonar o seu telefone*”, onde no texto dessas notícias é possível encontrar os termos supracitados. Portanto, novamente temos um tópico associado a categoria de notícias relacionadas à tecnologia, visto que a maior parte das notícias desse tipo pertencem ao tema sobre tecnologia no *corpus*.

• T_4: 0.294 * **video** + 0.089 * **policia** + 0.074 * **matar** + 0.064 * **globo** + 0.052 * **retirar** + 0.044 * **tomar** + 0.043 * **aprovar** + 0.027 * **pagar** + 0.026 * **ano** + 0.025 * **chinês** – esse tópico tem como termos principais, as palavras-chave: “vídeo”, “polícia” e “matar”. Em geral, boa parte desses termos no *corpus* estão associados a falsas notícias sobre crimes e violência. Além disso, existe a presença de termos como: “globo” e “chines”, que se enquadram em outros tipos de notícia, tais como: falsas notícias criadas utilizando veículos de comunicação

conhecidos e notícias falsas sobre a pandemia de COVID-19. Portanto, ao considerar os principais termos do tópico, temos associação com notícias sobre crimes que, em sua maior parte pertencem a categoria sobre assuntos nacionais. Por outro lado, temos os termos isolados supracitados que, por sua vez, se associam tanto a notícias sobre política, quanto a boatos sobre saúde. Portanto, temos que esse tópico tem relação a três grandes categorias: assuntos nacionais, política e saúde.

Portanto, de maneira geral, temos que a modelagem para o *corpus* completo obteve seu valor máximo de coerência com 5 tópicos. A maior parte dos tópicos possui uma categoria de notícia dominante entre os assuntos a que os termos presentes remetem, com exceção do tópico T_4 que, por sua vez, possui relação com 3 categorias distintas. Adicionalmente, é importante mencionar o fato de que tópicos como T_1 e T_2 destacam as inferências realizadas na análise exploratória realizada no capítulo 5 onde constatou-se que assuntos como as eleições presidenciais e a pandemia de COVID-19 foram destaques em uma grande quantidade de notícias falsas disseminadas.

6.2.4 Dicionário de Tópicos por Período de Tempo

Esta subseção traz a descrição da aplicação da abordagem já descrita e os resultados obtidos, de forma temporal. Logo, a descrição será realizada de forma a abordar de maneira cronológica a modelagem de tópicos dos períodos de tempo abrangidos no *corpus* de estudo. Adicionalmente, esta subseção realiza uma análise quantitativa dos dados por meio do estudo da métrica de coerência para os hiperparâmetros de treino da modelagem, assim como traz uma análise qualitativa dos resultados alcançados por meio da descrição e interpretação dos tópicos gerados para cada período temporal presente no *corpus*.

Considerando o processo descrito na subseção 6.2.2, e realizando tanto a otimização dos hiperparâmetros, quanto a análise manual dos tópicos foram encontradas as coerências máximas para a modelagem de tópicos de cada ano presente no *corpus*. Em vista disso, foi construída uma tabela comparativa com o intuito de realizar uma comparação entre o valor de coerência obtido com o experimento inicial com 10 tópicos fixos que foi realizada para todos os períodos de tempo, com o valor de coerência obtido após a otimização dos hiperparâmetros. O

principal objetivo consiste em demonstrar o percentual de melhora em relação ao experimento inicial. Dessa forma, é possível demonstrar a efetividade da abordagem na modelagem de tópicos. Na Tabela 1, é possível visualizar a comparação. No Apêndice F, é possível observar os valores de coerência para todas as combinações possíveis dos hiperparâmetros usados na otimização da abordagem LDA realizada temporalmente.

Tabela 1 - Evolução da métrica de coerência para as modelagens temporais via LDA após otimização

Ano	Coerência (Experimento inicial)	Coerência (Após otimização dos hiperparâmetros)	Percentual de ganho
2013-2015	0,4199	0,5511	31,24%
2016	0,3595	0,4952	37,74%
2017	0,4445	0,4802	8,03%
2018	0,4144	0,5172	24,80%
2019	0,3777	0,4735	25,36%
2020	0,4280	0,4929	15,16%
2021	0,3735	0,5108	36,76%

Fonte: Elaborado pelo autor (2023).

Na Tabela 1, é possível observar que a média de ganho após otimização dos hiperparâmetros entre as modelagens é de 25,58%. Alguns anos tendem a demonstrar maiores ganhos principalmente devido ao experimento inicial ter $K = 10$ fixo para todos os anos, consequentemente alguns anos terão número de tópicos ótimo próximo desse valor, e outros não.

Do ponto de vista qualitativo, temos no Quadro 4, a modelagem realizada para os anos presentes no *corpus*. Para cada tópico identificado (com as *Top* 10 palavras do tópico), é apresentado o ano em que se enquadra o mesmo. Nesse caso, temos que cada linha corresponde a um tópico (começando no tópico 0 até o tópico $K - 1$), onde K é o número de tópicos ótimo para o modelo). Os termos de cada tópico são separados por "+" e tem suas probabilidades associadas.

Quadro 4 – Modelagem de tópicos temporal via LDA

Ano	Tópico (0 até K-1)
2013-2015	'0.044*"carne" + 0.026*"humano" + 0.016*"encontrar" + 0.015*"restaurante" + ' '0.011*"usar" + 0.010*"conhecer" + 0.010*"fabricar" + 0.010*"enchimento" + ' '0.008*"acabar" + 0.007*"acordo"'
	0.017*"estudante" + 0.013*"falso" + 0.013*"tirar" + 0.010*"cama" + ' '0.010*"verdade" + 0.009*"pobre" + 0.009*"escandalo" + 0.009*"diabo" + ' '0.008*"profissional" + 0.008*"foco
	0.017*"brasilia" + 0.016*"loterico" + 0.014*"falar" + 0.013*"precisar" + ' '0.012*"atencao" + 0.012*"dia" + 0.011*"mega" + 0.011*"premio" + ' '0.011*"seno" + 0.009*"evitar
	0.020*"errar" + 0.018*"informacao" + 0.017*"apontar" + 0.014*"mercado" + '

	'0.014*"aceitar" + 0.013*"crise" + 0.013*"sociedade" + 0.013*"negro" + ' '0.013*"economico" + 0.013*"imobiliario
	0.018*"ano" + 0.015*"justica" + 0.014*"verdadeiro" + 0.012*"pai" + ' '0.012*"soldado" + 0.012*"politicar" + 0.012*"guerra" + 0.012*"banir" + ' '0.012*"allahu" + 0.012*"cristianismo
	0.008*"episodio" + 0.006*"atentado" + 0.006*"circular" + 0.005*"otico" + ' '0.005*"ilusao" + 0.004*"interromper" + 0.003*"vida" + 0.003*"prefeitura" + ' '0.003*"perceber" + 0.003*"chocar
	0.031*"desastre" + 0.021*"natural" + 0.020*"decreto" + 0.013*"crime" + ' '0.012*"culpa" + 0.012*"movimento" + 0.012*"passar" + 0.011*"grande" + ' '0.011*"tragediar" + 0.011*"assinar
	0.007*"comprar" + 0.005*"poder" + 0.005*"sinal" + 0.005*"fazenda" + ' '0.005*"substancia" + 0.005*"utilizar" + 0.005*"matar" + 0.005*"alto" + ' '0.004*"falar" + 0.004*"surgir
	0.032*"fazenda" + 0.018*"animal" + 0.015*"vaca" + 0.010*"possivel" + ' '0.008*"apenas" + 0.008*"convocar" + 0.007*"movimento" + 0.007*"agrar" + ' '0.007*"pericia" + 0.007*"reforma
	0.027*"estadual" + 0.017*"nasser" + 0.017*"associacoes" + 0.017*"condenar" ' '+ 0.017*"youssef" + 0.017*"desviar" + 0.017*"deputado" + 0.015*"juiz" + ' '0.012*"publico" + 0.011*"chamar
	0.020*"catolico" + 0.018*"processar" + 0.012*"igreja" + 0.009*"manifestar" ' '+ 0.009*"parada" + 0.006*"movimento" + 0.006*"cristofobia" + ' '0.006*"organizacao" + 0.004*"casca" + 0.003*"deixar
	0.027*"guerra" + 0.019*"mundial" + 0.015*"grupo" + 0.015*"feira" + ' '0.015*"lancar" + 0.014*"nome" + 0.014*"terrorista" + 0.012*"unir" + ' '0.012*"nacoes" + 0.009*"conhecer
	0.010*"querer" + 0.008*"tomar" + 0.008*"atleta" + 0.007*"gesto" + ' '0.007*"proibir" + 0.006*"nome" + 0.006*"presidente" + 0.006*"falar" + ' '0.006*"saber" + 0.006*"pessoa
	0.024*"levar" + 0.018*"ficar" + 0.016*"tiro" + 0.011*"querer" + ' '0.011*"direito" + 0.011*"chegar" + 0.011*"chamar" + 0.010*"mulher" + ' '0.010*"video" + 0.010*"ocorrer
2016	'0.002*"produto" + 0.002*"adesivo" + 0.001*"indicar" + 0.001*"etiqueta" + ' '0.001*"numero" + 0.001*"banana" + 0.001*"fruta" + 0.001*"modo" + ' '0.001*"exemplo" + 0.001*"modificar
	0.005*"governo" + 0.005*"presidente" + 0.004*"brasileiro" + 0.003*"pai" + ' '0.003*"temer" + 0.003*"nacional" + 0.003*"novo" + 0.003*"programa" + ' '0.003*"publicar" + 0.002*"federal
	0.003*"frango" + 0.002*"carne" + 0.001*"deslizamento" + 0.001*"adicionar" + ' '0.001*"livro" + 0.001*"arsenico" + 0.001*"ilha" + 0.001*"comer" + ' '0.001*"nitro" + 0.001*"olho
	0.007*"pessoa" + 0.006*"ficar" + 0.005*"passar" + 0.004*"gente" + ' '0.004*"saber" + 0.004*"hoje" + 0.004*"falar" + 0.004*"casa" + ' '0.004*"querer" + 0.004*"chegar
	0.001*"igreja" + 0.001*"visita" + 0.001*"cabra" + 0.001*"congregacao" + ' '0.001*"papa" + 0.001*"denominacao" + 0.000*"agredir" + 0.000*"crista" + ' '0.000*"dello" + 0.000*"outubro
	0.002*"bafometro" + 0.001*"vinagre" + 0.001*"corpo" + 0.001*"queimar" + ' '0.001*"gordura" + 0.001*"governador" + 0.001*"faixa" + 0.001*"alcool" + ' '0.001*"cetona" + 0.001*"cetonico
	0.005*"arroz" + 0.002*"plastico" + 0.002*"teste" + 0.001*"coloquar" + ' '0.001*"fogo" + 0.001*"quantidade" + 0.001*"colher" + 0.001*"mofo" + ' '0.001*"truque" + 0.001*"artificial
	0.003*"beber" + 0.002*"dia" + 0.002*"doenca" + 0.002*"amarelo" + ' '0.002*"motoqueiro" + 0.002*"agua" + 0.002*"virus" + 0.002*"americano" + ' '0.002*"atleta" + 0.002*"farol
2017	0.005*"ano" + 0.004*"passar" + 0.004*"pessoa" + 0.004*"saber" + ' '0.004*"mensagem" + 0.004*"falar" + 0.004*"amigo" + 0.004*"pedir" + ' '0.004*"chegar" + 0.004*"homem

	0.004*"falar" + 0.004*"site" + 0.003*"achar" + 0.003*"hoje" + 0.003*"casa" + ' '0.003*"pessoa" + 0.003*"globo" + 0.003*"saber" + 0.003*"virar" + ' '0.003*"pedir
	0.003*"ano" + 0.003*"idoso" + 0.002*"vaga" + 0.002*"show" + 0.002*"premio" + ' '0.002*"liberar" + 0.002*"cacau" + 0.002*"passagem" + 0.002*"plastico" + ' '0.001*"gritar
	0.002*"vender" + 0.002*"julgamento" + 0.002*"imposto" + ' '0.001*"manifestante" + 0.001*"propagandar" + 0.001*"multa" + ' '0.001*"politico" + 0.001*"embraer" + 0.001*"bono" + 0.001*"criticar
	0.003*"recarga" + 0.003*"Gratis" + 0.003*"projeto" + 0.002*"ganhar" + ' '0.002*"general" + 0.002*"criar" + 0.002*"tabu" + 0.002*"print_noticia" + ' '0.002*"realizar" + 0.002*"feliciano
	0.003*"precisamos_santo" + 0.002*"novo" + 0.002*"cientista" + ' '0.002*"cobrar" + 0.002*"comunismo" + 0.002*"tecido" + 0.001*"creditar" + ' '0.001*"tornar" + 0.001*"direita" + 0.001*"inventar
	0.002*"fruta" + 0.001*"comer" + 0.001*"encher" + 0.001*"vegetal" + ' '0.001*"tiririca" + 0.001*"jejum" + 0.001*"cadastro" + 0.001*"dar" + ' '0.001*"prolongar" + 0.001*"print_noticia
2018	0.006*"receber" + 0.005*"bolsonaro" + 0.004*"presidente" + 0.004*"ano" + ' '0.004*"ganhar" + 0.003*"governo" + 0.003*"federal" + 0.003*"site" + ' '0.003*"acabar" + 0.003*"beneficio
	0.008*"bolsonaro" + 0.005*"brasileiro" + 0.004*"russo" + 0.004*"apoio" + ' '0.003*"presidente" + 0.003*"maduro" + 0.002*"natal" + 0.002*"imigrante" + ' '0.002*"espaco" + 0.002*"indulto
	0.006*"mulher" + 0.005*"video" + 0.005*"federal" + 0.004*"vizinho" + ' '0.004*"senador" + 0.004*"mostrar" + 0.003*"oferecer" + 0.003*"caso" + ' '0.003*"engravidar" + 0.003*"condenar
	0.006*"bolsonaro" + 0.005*"olhar" + 0.005*"ficar" + 0.005*"saber" + ' '0.004*"pessoa" + 0.004*"querer" + 0.004*"passar" + 0.004*"vida" + ' '0.004*"novo" + 0.003*"ano
	0.003*"mostrar" + 0.003*"foto" + 0.002*"preco" + 0.002*"banana" + ' '0.002*"justificar" + 0.001*"paulo_guede" + 0.001*"atestado" + 0.001*"jovem" + ' '0.001*"idoso" + 0.001*"produzir
2019	0.007*"presidente" + 0.007*"bolsonaro" + 0.004*"brasileiro" + ' '0.003*"amazonio" + 0.003*"publicar" + 0.003*"livro" + 0.002*"americano" + ' '0.002*"grande" + 0.002*"terra" + 0.002*"pai
	0.001*"vagabundo" + 0.001*"forjar" + 0.001*"disparo" + 0.001*"fundo" + ' '0.001*"ataque" + 0.001*"soldado" + 0.001*"prefeito" + 0.001*"lado" + ' '0.001*"confederar" + 0.001*"amazonico
	0.006*"pessoa" + 0.006*"falar" + 0.004*"ficar" + 0.004*"ano" + ' '0.004*"querer" + 0.004*"presidente" + 0.004*"hoje" + 0.004*"novo" + ' '0.003*"pedir" + 0.003*"saber
	0.003*"vaga" + 0.002*"aluno" + 0.002*"ministro" + 0.002*"curso" + ' '0.001*"plurinacional" + 0.001*"limpeza" + 0.001*"prova" + 0.001*"decretar" + ' '0.001*"estudante" + 0.001*"juiz
	0.008*"site" + 0.007*"abaixo" + 0.005*"ganhar" + 0.005*"compartilhar" + ' '0.004*"receber" + 0.002*"limite" + 0.002*"compartilhe" + 0.002*"responder" + ' '0.002*"retirar" + 0.002*"amigo
	0.004*"bolsonaro" + 0.004*"dinheiro" + 0.002*"nordeste" + 0.002*"oleo" + ' '0.002*"crime" + 0.002*"governador" + 0.002*"esquerdo" + 0.002*"bandido" + ' '0.002*"governo" + 0.002*"natal
2020	0.004*"bolsonaro" + 0.003*"urna" + 0.002*"votar" + 0.002*"dono" + ' '0.002*"ministro" + 0.002*"presidente" + 0.002*"chines" + 0.002*"voto" + ' '0.002*"acaso" + 0.002*"comando
	0.006*"site" + 0.005*"presente" + 0.005*"ganhar" + 0.005*"receber" + ' '0.005*"amigo" + 0.004*"valer" + 0.004*"novo" + 0.004*"cadastro" + ' '0.004*"grupo" + 0.004*"natal
	0.002*"candidato" + 0.002*"testa" + 0.002*"advogado" + 0.002*"perdao" + ' '0.002*"cliente" + 0.001*"apontar" + 0.001*"chopp" + 0.001*"traicao" + '

	'0.001*"pagamento" + 0.001*"Exposicao
	0.007*"pessoa" + 0.005*"saber" + 0.005*"vacino" + 0.005*"covid" + ' '0.005*"falar" + 0.004*"tomar" + 0.004*"bolsonaro" + 0.004*"ficar" + ' '0.004*"ano" + 0.004*"virus
	0.002*"espelho" + 0.001*"foto" + 0.001*"resort" + 0.001*"maravilhoso" + ' '0.001*"agulha" + 0.001*"praga" + 0.001*"maikelly" + 0.001*"cicatriz" + ' '0.001*"trecho" + 0.001*"quadro
	0.002*"noel" + 0.001*"chegada" + 0.001*"mafia" + 0.001*"renunciar" + ' '0.001*"cargo" + 0.001*"lir" + 0.001*"papai" + 0.001*"helicoptero" + ' '0.001*"derby" + 0.001*"quartel
2021	0.001*"policial" + 0.001*"argentino" + 0.001*"empresario" + 0.001*"fila" + ' '0.001*"populacao" + 0.001*"argentina" + 0.001*"pais" + 0.001*"destacar" + ' '0.001*"promocao" + 0.001*"ganhar
	0.012*"bolsonaro" + 0.009*"presidente" + 0.005*"ministro" + ' '0.004*"brasileiro" + 0.004*"querer" + 0.004*"mostrar" + 0.003*"ficar" + ' '0.003*"povo" + 0.003*"saber" + 0.003*"acabar
	0.001*"mascara" + 0.001*"cair" + 0.001*"predio" + 0.001*"motocicleta" + ' '0.001*"falso" + 0.001*"cura" + 0.001*"derramar" + 0.001*"cafe" + ' '0.001*"ver" + 0.001*"tamanho
	0.003*"motoqueiro" + 0.003*"senador" + 0.003*"presidente" + 0.001*"assumir" ' '+ 0.001*"hein" + 0.001*"tecnico" + 0.001*"presidio" + 0.001*"gotinha" + ' '0.001*"categoria" + 0.001*"doente
	0.001*"empresario" + 0.001*"trabalho" + 0.001*"digital" + 0.001*"prefeito" ' '+ 0.001*"atividade" + 0.001*"direito_constitucional" + 0.001*"seguir" + ' '0.001*"produto" + 0.001*"informar" + 0.001*"mail
	0.006*"presente" + 0.005*"hoje" + 0.004*"vacino" + 0.003*"ganhar" + ' '0.003*"tomar" + 0.003*"parabens" + 0.003*"abaixo" + 0.003*"premio" + ' '0.003*"conhecer" + 0.003*"dose
	0.004*"agua" + 0.003*"covid" + 0.003*"frio" + 0.002*"temperatura" + ' '0.002*"causa" + 0.002*"chegar" + 0.001*"forte" + 0.001*"virus" + ' '0.001*"alertar" + 0.001*"extremo
	0.009*"pessoa" + 0.007*"covid" + 0.004*"querer" + 0.004*"ano" + ' '0.004*"saber" + 0.004*"olhar" + 0.004*"hospital" + 0.004*"receber" + ' '0.004*"tomar" + 0.004*"novo

Fonte: Elaborado pelo autor (2023).

A seguir é feita a descrição e interpretação dos tópicos gerados nas modelagens realizadas para cada período de tempo. A ideia central consiste em identificar os principais termos de cada tópico, realizando uma busca entre as notícias falsas do *corpus* de forma a identificar as principais notícias de cada período de tempo que utilizam os mesmos termos. Consequentemente, é possível identificar o assunto tratado por cada tópico. Além disso, são citados os títulos de notícias falsas do *corpus* que usam os mesmos termos e abordam o assunto tratado, como uma maneira de exemplificar a descrição e interpretação realizada.

Analisando detalhadamente é possível perceber que a modelagem do triênio (2013-2015) possui uma modelagem com 14 tópicos em sua versão otimizada. O tópico 0 está relacionado aos boatos sobre restaurantes estarem usando ingredientes questionáveis em seus alimentos. Os termos mais frequentes neste tópico são “carne”, “humano” e “restaurante”. Embora não haja evidências reais para sustentar

essa afirmação, a ideia de que restaurantes usam ingredientes questionáveis é algo que já foi abordado em outras matérias jornalísticas e documentários. Como exemplo de notícia falsa presente no *corpus* que trata sobre esse assunto, é possível identificar: "*Carne humana é encontrada em fábrica do McDonald's no EUA*".

O tópico 1 tem como termos mais frequentes: "estudante", "falso" e "tirar". Em geral, esse tópico faz associação com notícias falsas do *corpus*, tais como: "*Candidato do Enem finge que se atrasa, tenta pular portão e mídia cai no boato*" e "*História falsa fala do experimento socialista de Adrian Rogers*". O tópico 2 traz entre os termos mais frequentes: "brasilia", "loterico" e "mega". Portanto, está relacionado com boatos sobre prêmios da loteria, como a Mega-Sena. Em geral, faz parte de um conjunto de notícias falsas divulgadas sobre um boato sobre uma suposta lotérica de uma figura pública sortear dois prêmios da Mega-Sena. Como exemplo, existem notícias no *corpus*, tais como: "*Lotérica de Brasília pagou dois prêmios da Mega-Sena em 60 dias*", "*Prêmio de R$ 205 milhões da Mega Sena saiu na lotérica do doleiro Youssef*", "*Doleiro Alberto Youssef é dono de lotérica em Brasília*", "*Agência do bilhete premiado da Mega-Sena de R$ 200 milhões não existe?*", "*Ex-deputado Nasser Youssef Nasr é dono da lotérica da Mega-Sena*".

O Tópico 3 está relacionado com problemas econômicos e imobiliários. Os termos mais frequentes neste tópico são: "crise", "imobiliario" e "economico". A notícia falsa "*Crise econômica prejudicou mercado imobiliário*" é um exemplo de desinformação presente no *corpus* que pode explicar esse tópico. O tópico 4 faz menção a termos relacionados com notícias falsas que versam sobre conflitos religiosos. Os termos mais frequentes neste tópico são: "politicar", "cristianismo" e "allahu". A notícia "*Mesquita Brasil ameaça o Cristianismo e o Ateísmo*" é um exemplo de boato presente no *corpus* que cita esses termos. No tópico 5 há a ocorrência de termos, tais como: "episodio", "atentado", entre outros termos sem muita correlação. A notícia falsa "*Dragon Ball Z foi interrompido no dia 11 de setembro*" traz termos semelhantes. Porém, o tópico não possui um assunto predominante.

O tópico 6 traz como assunto principal, um desastre natural. Os termos mais frequentes neste tópico são: "desastre", "crime" e "tragediar". Em geral, termos semelhantes aparecem em notícias falsas relativas ao Rompimento da barragem em Mariana no Estado de Minas Gerais[29], entre as notícias: "*Dilma isenta Samarco por*

[29] Disponível em: https://www.mpf.mp.br/grandes-casos/caso-samarco/o-desastre

tragédia em Mariana com decreto", "*Lama tóxica de Mariana atinge a Bahia*", "*Grupo terrorista Vale/Samarco assume atentado em Mariana*". O tópico 7 não traz um assunto específico. Os termos mais frequentes neste tópico são: "comprar", "substância" e "matar". O tópico 8 traz como principais termos: "fazenda", "animal", "vaca", "movimento", "agraria", "reforma", o que remete a notícias falsas sobre o MST ter matado vacas em fazendas. O tópico 9 novamente traz termos que remetem à figura de Nasser Youssef para criação de notícias falsas.

O tópico 10, por sua vez, está relacionado com notícias falsas relativas à manifestações e questões religiosas. Os termos mais frequentes neste tópico são: "catolico", "cristofobia" e "igreja". No tópico 11 há a presença de termos, tais como: "guerra", "mundial", "grupo", "terrorista", que remetem a notícias falsas, tais como: "*Terceira Guerra Mundial está próxima de acontecer*", "*Brasil vai construir embaixada para terroristas na Palestina*". Portanto, podemos afirmar que o tópico reflete notícias da categoria sobre assuntos internacionais. O tópico 12 traz assuntos relacionados a categoria de notícias relativa ao esporte. Os termos mais frequentes neste tópico são: "atleta", "proibir" e "gesto". Por fim, o tópico 13 tem relação com assuntos ligados a violência. Os termos mais frequentes neste tópico são: "direito", "tiro" e "mulher". A notícia falsa "*Motoqueiro fantasma prende e mata bandidos em Teresina*" emprega a maior parte dos termos desse tópico em sua construção.

A modelagem relativa ao ano de 2016 possui um total de 8 tópicos. O tópico 0 traz como assunto, produtos e etiquetas, tendo relação com à indústria alimentícia e de bens de consumo em geral. Por exemplo, termos como: "produto", "etiqueta", "numero" e "modificar" estão relacionados à rotulagem de alimentos. A notícia falsa "*Não coma frutas com a etiqueta número 8; são transgênicas*" presente no *corpus*, é um exemplo de boato que aborda essa temática. O tópico 1 está relacionado ao governo brasileiro e a políticas públicas. Por conseguinte, termos como: "governo", "presidente", "brasileiro", "temer", "nacional" e "federal" indicam que as notícias falsas se concentram em questões políticas. A notícia falsa: "*Obama diz que não reconhece governo Temer*" é um exemplo de boato contido no *corpus*.

O tópico 2 tem seu entendimento dificultado, principalmente devido ao fato de que o conjunto de palavras não parece formar um padrão claro. Os termos "frango" e "carne" sugerem boatos relacionados à indústria alimentícia, enquanto "deslizamento" sugere algo relacionado a desastres naturais. Como exemplo de notícias falsas presentes no *corpus* que tratam dos dois assuntos, é possível observar: "*75% dos*

frangos têm uma substância cancerígena, o arsênico” e “*Nasa confirma catástrofe no Brasil e faz alerta de tsunami*”, que aponta uma possível erupção de vulcão acarretar em deslizamentos de terra. O tópico 3 traz termos que remetem as pessoas e suas interações diárias. Nesse caso, termos como: “pessoa”, “ficar”, “passar”, “gente”, “saber” e “casa” sugerem que as notícias falsas se concentram nas vidas cotidianas das pessoas, visto que boa parte das palavras se concentra em verbos que indicam ações cotidianas. Entretanto, não há ligação clara com nenhum assunto específico.

O tópico 4 é bastante difícil de interpretar, pois os termos são muito diversos e não parecem formar um padrão claro. O termo “igreja” sugere algo relacionado à religião ou espiritualidade, enquanto “agredir” sugere notícias ligadas à violência. Além disso, outros termos como: “cabra”, “papa” e “outubro” reforçam boatos ligados à categoria de notícias sobre religião. O tópico 5 tem relação com boatos sobre assuntos nacionais. Nesse caso, termos como: “bafometro”, “vinagre”, “corpo”, “queimar”, “gordura” e “doença” estão atrelados a boatos, tal como: “*Beber vinagre altera resultado do teste do bafômetro*”. O tópico 6 traz termos, tais como: “arroz”, “plastico”, “colher” e “mofo” e estão ligados a um boato que diz: “*Arroz de plástico da China é vendido no Brasil*”. Portanto, temos que o tópico está ligado à categoria sobre assuntos internacionais que, por sua vez, é a categoria que contém esse boato. Por fim, o tópico 7 parece estar relacionado a diferentes temas, incluindo doenças, esportes e trânsito. Nesse caso, termos como: “doença”, “virus”, “atleta”, “farol” e “motoqueiro” sugerem uma variedade de assuntos. Portanto, não há um assunto predominante.

O ano de 2017 possui em sua modelagem um total de 7 tópicos. O tópico 0 traz termos, tais como: “ano”, “mensagem”, “falar” e “amigo”. Em geral, não há um assunto específico. Mas é importante destacar que os termos indicam a forma de disseminação das informações falsas que estão constantemente presentes em notícias falsas, onde é descrito por onde foi encaminhada e qual a relação entre as pessoas envolvidas na disseminação da notícia falsa. O tópico 1 têm termos como “site”, “hoje” e “globo”, sugerindo um tema relacionado às notícias e mídia. Em geral, não há um assunto predominante.

O tópico 2 inclui tanto termos como: “idoso”, “vaga” e “passagem”, que estão relacionadas com notícias falsas sobre desconto em passagens aéreas para idosos, quanto termos como: “cacau”, “show”, “premio”, que estão relacionados ao seguinte boato: “*Cacau Show está presenteando com Kit Natal em site oficial*”. Portanto, não

há um assunto predominante. O tópico 3 parece estar relacionado à política e justiça, com termos, tais como: “julgamento”, “imposto” e “manifestante”. Em geral, algumas das notícias falsas que estão relacionadas com esse tópico são: “*Decreto aumenta alíquota do Imposto de Renda de 27,5% para 35%*”, “*Receita Federal pede em carta para você atualizar dados no info2010.x10.mx*” e “*Toda cobrança de taxa de incêndio é inconstitucional*”. O tópico 4 inclui termos como: “projeto” e “ganhar”, e tem relação com o seguinte boato: “*Deputados vão ganhar R$ 6 mi para votar reforma da Previdência*”. Portanto, temos que o tópico trata sobre assuntos políticos.

O tópico 5 inclui termos como “precisamos_santo” e “novo”, sugerindo um tema relacionado à religião. Nesse período há a ocorrência de notícias falsas sobre religião. A notícia “*Papa Francisco disse que precisamos de santos de calça jeans*” é um exemplo de boato do *corpus* que utiliza os termos supracitados. Por fim, o tópico 6 inclui termos como: “fruta”, “comer” e “vegetal”, que por sua vez foram ostensivamente utilizados em um falso boato sobre uma entrevista polêmica de um médico sobre dicas de saúde.

A modelagem relativa ao ano de 2018 tem um total de 5 tópicos. O tópico 0 tem relação ao tema dos benefícios sociais recebidos pelos cidadãos brasileiros durante o governo do presidente Bolsonaro em 2018. Os principais termos que compõem este tópico são: “receber”, “bolsonaro”, “presidente”, “ganhar” e “beneficio”. Em geral, há a presença no *corpus* de notícias falsas concentradas em criticar ou difamar o governo por meio de boatos. O tópico 1 traz termos, tais como: “bolsonaro”, “brasileiro”, “russo”, “apoio” e “presidente”. Entre as principais notícias desse período que envolvem esses termos é possível citar: “*Padre Marcelo Rossi apoia Bolsonaro e grava áudio contra comunistas*”, “*Avião russo Sukhoi 30 faz acrobacias com dança no solo em vídeo*”, entre outras.

O tópico 2 traz termos como: “federal”, “senador”, “oferecer”. Em geral, grande parte dos boatos com esses termos estão associados com notícias sobre política. Entre as notícias que usam esses termos, podem ser citadas: “*Voto parcial: quem votar só para presidente e branco nos outros candidatos terá voto anulado*” e “*Petistas distribuem capim para nordestinos em carreata falsa de Bolsonaro*”. O tópico 3 traz termos relacionados com comentários difamatórios falsos sobre o político Bolsonaro. Os principais termos que compõem este tópico são: “bolsonaro”, “olhar”, “saber”, “ficar” e “pessoa”. É possível que essas notícias falsas tenham sido criadas para gerar polêmica, bem como manipular a opinião pública. O estudo (VILMER et al., 2018)

mostra de forma mais detalhada o processo de manipulação de opinião gerado por conteúdos de desinformação.

O tópico 4 está relacionado com notícias falsas sem grande relevância ou importância. Os principais termos que compõem esse tópico são: “mostrar”, “foto”, “preco”, “banana” e “justificar”. Essas notícias falsas provavelmente refletem variados assuntos, não havendo nenhuma temática predominante. Entretanto, é importante destacar o termo “foto” que muitas vezes é usado em notícias falsas com o intuito de embasar falsas informações e acontecimentos. O estudo de (SHEN et al., 2019) mostra um experimento online para entender como as pessoas avaliam a credibilidade de imagens que são compartilhadas nas plataformas online, segundo esse estudo os participantes foram aleatoriamente designados para observar fontes de notícias com imagens falsas associadas e avaliaram a credibilidade das imagens com base em vários recursos. O estudo descobriu que as habilidades dos participantes na Internet, a experiência de edição de fotos e o uso de mídia social foram preditores significativos da avaliação da credibilidade das imagens, enquanto a maioria das dicas sociais e heurísticas de credibilidade online (por exemplo, confiabilidade da fonte, confiabilidade do intermediário, etc) não tiveram impacto significativo. A atitude dos espectadores em relação a uma questão retratada por imagens falsas também influenciou positivamente sua avaliação de credibilidade.

O ano de 2019 traz em sua modelagem um total de 6 tópicos. O tópico 0 está relacionado ao presidente Bolsonaro e à Amazônia. Os termos mais relevantes são: “presidente”, “bolsonaro”, “brasileiro”, “amazonico” e “publicar”. Em 2019, houve uma grande preocupação com o desmatamento da Floresta Amazônica e a postura do governo brasileiro em relação a isso. O presidente Bolsonaro enfrentou críticas internacionais pelo que muitos consideraram como políticas prejudiciais ao meio ambiente. O tópico 1 possui interpretação difícil, pois não parece ter um tema específico. Alguns dos termos incluem: “forjar”, “ataque” e “soldado”. De forma mais específica, temos que os dois primeiros termos estão presentes nos seguintes boatos: “*Ataque ao Porta dos Fundos foi forjado por Fábio Porchat*”, “*Pai de Fábio Porchat, Fábio Porchat Assis, é acusado de desviar recursos da Lei Rouanet*”. O último termo aparece em notícias, tais como: “*Obama falou que é graças aos soldados que temos liberdade de religião, imprensa, etc*”.

O tópico 2 está relacionado com pessoas e suas opiniões. Os termos mais importantes são: “pessoa”, “falar”, “querer”, “presidente” e “novo”. Em geral, boa parte

das notícias falsas do *corpus* que estão associadas com esse tópico trazem discussões políticas e sociais que ocorreram em 2019. Durante esse ano, houve muitas discussões sobre as políticas do governo federal e o papel das pessoas na mudança política. O tópico 3 está claramente associado com questões educacionais. Os termos mais relevantes são: “vaga”, “aluno”, “ministro”, “curso” e “prova”. Em 2019, houve muitas discussões sobre as políticas educacionais do governo federal, incluindo o corte de verbas para universidades públicas e a reforma do ensino médio. O tópico 4 tem relação com golpes virtuais. Os termos mais importantes são: “site”, “abaixo”, “ganhar”, “compartilhar” e “receber”. Em geral, eventos como promoções em mídias sociais ou campanhas de marketing virais que ocorreram em 2019 com o intuito de realizar golpes nos participantes. Finalmente, o tópico 5 está relacionado com questões políticas, incluindo corrupção e crimes. Os termos mais importantes são: “bolsonaro”, “dinheiro”, “nordeste”, “crime” e “governador”.

A modelagem relativa ao ano de 2020 tem um total de 6 tópicos. O tópico 0 está associado à política brasileira. Os termos mais importantes incluem: “Bolsonaro”, “urna”, “votar”, “ministro” e “presidente”. Em 2020, houve muita controvérsia em torno do sistema eleitoral brasileiro, com muitas teorias sendo propagadas sobre supostas fraudes nas urnas eletrônicas. O tópico 1 traz novamente golpes e fraudes virtuais por meio de sorteios. Os termos mais importantes incluem: “site”, “presente”, “ganhar”, “receber” e “amigo”. Em 2020, com muitas pessoas passando mais tempo em casa e fazendo compras online, houve um aumento significativo em golpes virtuais[30]. O tópico 2 traz assuntos sobre política. Os termos mais importantes incluem: “candidato”, “advogado”, “traição”, “perdao” e “pagamento”. Entretanto, as notícias falsas que incluem termos semelhantes tratam de variados assuntos políticos, não havendo nenhuma predominância temática.

O tópico 3 traz como assunto predominante boatos relacionados à pandemia de COVID-19. Os termos mais importantes incluem: “pessoa”, “vacina”, “covid”, “virus” e “tomar”. Em 2020, a pandemia de COVID-19 afetou enormemente todo o mundo, e houve muita discussão sobre a eficácia das vacinas e medidas de prevenção. Portanto, a disseminação de boatos sobre essas questões foi constante. O tópico 4 inclui termos que não estão claramente relacionados uns com os outros. Os termos mais importantes incluem: “espelho”, “foto”, “maravilhoso”, “quadro” e “resort”.

[30] Disponível em: https://g1.globo.com/rj/rio-de-janeiro/noticia/2021/01/28/numero-de-golpes-pela-internet-quase-triplicou-em-2020-aponta-isp.ghtml

Portanto, não há predominância temática. Por fim, o tópico 5 traz entre os termos mais importantes: “noel”, “mafia”, “renunciar”, “cargo” e “helicoptero”. Os termos “renunciar” e “cargo” estão envolvidos em notícias falsas, tais como: “*Papa Bento XVI foi exilado na Alemanha após ser expulso do Vaticano*”, “*Mineradora norueguesa se instalou na Amazônia em um acordo com Lula*”. Por outro lado, os termos “noel”, “helicoptero” estão ligados ao seguinte boato: “*Papai Noel fica pendurado em helicóptero durante chegada, mostra vídeo*”.

O ano de 2021 tem em sua modelagem um total de 8 tópicos. O tópico 0 tem relação com notícias falsas sobre a Argentina e envolve termos como: “policial”, “argentino”, “empresario” e “fila”. Portanto, temos que este tópico está associado a notícias falsas sobre questões políticas e econômicas na Argentina em 2021. A notícia falsa: “*Dono da Volpato Engates faz relato sobre miséria na Argentina e exalta Bolsonaro*” inclui os termos: “argentino”, “empresario” e “fila”. O tópico 1 está claramente relacionado com o presidente brasileiro Jair Bolsonaro, com termos como: “bolsonaro”, “presidente” e “ministro”. Em geral, os termos estão associados a notícias falsas sobre as políticas do governo brasileiro em 2021.

O tópico 2 apresenta termos bastante peculiares, como: “mascara”, “cair”, “predio” e “motocicleta”. O termo “motocicleta” está constantemente associados a boatos que envolvem o nome de Jair Bolsonaro, incluindo notícias falsas, tais como: “*Motociata de Bolsonaro teve 1.324.523 motos e bateu recorde no Guinness Book*”. Por outro lado, o termo “mascara” está constantemente associado a notícias falsas sobre o uso de máscara durante a pandemia de COVID-19. O tópico 3 apresenta um termo em específico que demonstra a desinformação sobre a COVID-19, o termo “gotinha”. O termo “gotinha” se faz presente no texto de uma notícia falsa com o seguinte título: “*Creolina cura a COVID-19 e mata o coronavírus, diz médico Roberto Klaus*”. Portanto, dessa maneira é possível constatar as falsas informações disseminadas no período de pandemia.

O tópico 4 apresenta termos como “empresario”, “trabalho” e “prefeito”. Em geral, existem no *corpus* notícias falsas que incluem cobertura sobre prefeitos envolvidos em escândalos empresariais. O tópico 5 novamente traz como assunto, a pandemia de COVID-19, com termos como: “vacino” (versão lematizada de “vacina”) e “covid”. Novamente, é importante ressaltar que o *corpus* possui notícias sobre a distribuição de vacinas contra a COVID-19 em 2021, juntamente com histórias falsas sobre efeitos adversos. O tópico 6 apresenta termos como: “agua”, “covid” e “frio”.

Algumas notícias falsas do *corpus* que envolvem os dois primeiros termos são: “*Nebulização caseira com bicarbonato de sódio e água oxigenada previne e cura Covid-19*” e “*Respirador caseiro feito com garrafa pet é eficaz contra Covid-19*”. Finalmente, o tópico 7 parece novamente trazer assuntos relacionados com a pandemia de COVID-19, com termos como: “covid”, “hospital” e “tomar”. Em geral, notícias falsas, tais como: “*Médicos do Hospital Beneficente Portuguesa, em Manaus (AM) fazem desabafo sobre pandemia*” e “*Roberto Klaus, médico virologista do Albert Einstein, diz que Coronavac não imuniza, vacina muda DNA e ivermectina é eficaz*”, são exemplos de boatos que incluem os termos capturados pela modelagem.

6.2.5 Dicionário de Tópicos por Categoria

Esta subseção traz a descrição da aplicação da abordagem já descrita e os resultados obtidos, de forma categórica. Consequentemente, a descrição será realizada de forma a abordar de maneira temática a modelagem de tópicos das categorias abrangidas no *corpus* de estudo.

Levando em conta o processo descrito na subseção 6.2.2, e realizando a otimização dos hiperparâmetros, e também a análise manual dos tópicos foram encontradas as coerências máximas para a modelagem de tópicos de cada categoria presente no *corpus*. À vista disso, foi construída uma tabela comparativa com o intuito de realizar uma comparação entre o valor de coerência obtido com o experimento inicial com 10 tópicos fixos que foi realizada para todos as categorias do *corpus*, com o valor de coerência obtido após a otimização dos hiperparâmetros. O principal objetivo consiste em demonstrar o percentual de melhora em relação ao experimento inicial. Dessa forma, é possível demonstrar a efetividade da abordagem na modelagem de tópicos. Na Tabela 2, é possível visualizar a comparação. Além disso, no Apêndice F é possível observar os valores de coerência para todas as combinações possíveis dos hiperparâmetros usados na otimização da abordagem LDA realizada categoricamente.

Tabela 2 - Evolução da métrica de coerência para as modelagens categóricas via LDA após otimização

Categoria	Coerência (Experimento inicial)	Coerência (Após otimização dos hiperparâmetros)	Percentual de ganho
Política	0,4642	0,6615	42,50%
Brasil	0,4016	0,6033	50,22%
Saúde	0,3275	0,4963	51,54%
Entretenimento	0,4305	0,6090	41,46%
Tecnologia	0,4288	0,5481	27,82%
Ciência	0,4517	0,5096	12,81%
Esporte	0,4147	0,5800	39,86%
Mundo	0,3230	0,5115	58,35%
Religião	0,3715	0,5738	54,45%

Fonte: Elaborado pelo autor (2023).

É possível observar que a média de ganho após otimização dos hiperparâmetros entre as modelagens é de 42,11%. Detalhadamente, categorias específicas demonstraram maiores ganhos principalmente devido ao experimento inicial ter $K = 10$ fixo para todas as categorias estudadas, consequentemente algumas das categorias tiveram em sua modelagem, número de tópicos ótimo distante desse valor, e outras não. Outra observação importante refere-se fato de que o percentual de ganho nas modelagens realizadas entre as categorias do *corpus* foi maior do que nas modelagens realizadas entre os períodos de tempo do *corpus*. Portanto, essa observação demonstra a influência do conjunto de documentos utilizado em relação a métrica de coerência estudada.

Qualitativamente, temos no Quadro 5, a modelagem realizada para as categorias presentes no *corpus*. Para cada tópico identificado (com as *Top* 10 palavras do tópico), é apresentado a categoria em que se enquadra o mesmo. Nesse caso, temos que cada linha corresponde a um tópico (começando no tópico 0 até o tópico $K - 1$), onde K é o número de tópicos ótimo para o modelo). Os termos de cada tópico são separados por "+" e tem suas probabilidades associadas.

Quadro 5 – Modelagem de tópicos categórica via LDA

Categoria	Tópicos
Política	0.126*"presidente" + 0.103*"criar" + 0.076*"moro" + 0.072*"acabar" + ' '0.058*"governador" + 0.048*"governo" + 0.038*"temer" + 0.038*"sair" + ' '0.032*"federal" + 0.031*"chorar
	0.444*"bolsonaro" + 0.136*"foto" + 0.066*"grande" + 0.038*"globo" + ' '0.033*"prender" + 0.030*"pagar" + 0.026*"causa" + 0.025*"fraude" + ' '0.023*"fotor" + 0.022*"empresa
	0.106*"vazar" + 0.102*"pedir" + 0.091*"militar" + 0.073*"familia" + ' '0.059*"intervencao" + 0.058*"general" + 0.052*"urna" + 0.048*"matar" + ' '0.044*"eleicoes" + 0.039*"policia

	0.330*"mostrar" + 0.153*"video" + 0.066*"ministro" + 0.048*"decreto" + ' '0.045*"deixar" + 0.040*"expulsar" + 0.035*"brasileiro" + 0.031*"aprovar" + ' '0.022*"deputado" + 0.020*"eleger
Brasil	0.120*"ano" + 0.086*"menina" + 0.070*"policia" + 0.052*"greve" + ' '0.043*"matar" + 0.041*"escola" + 0.039*"causa" + 0.037*"chinês" + ' '0.031*"desaparecer" + 0.026*"bala
	0.393*"mostrar" + 0.238*"video" + 0.046*"morrer" + 0.027*"matar" + ' '0.027*"bater" + 0.023*"cobrar" + 0.023*"procurar" + 0.020*"governo" + ' '0.016*"comemorar" + 0.015*"possivel
	0.330*"dia" + 0.070*"motorista" + 0.049*"multa" + 0.039*"vender" + ' '0.037*"tomar" + 0.030*"site" + 0.029*"transito" + 0.025*"conta" + ' '0.023*"globo" + 0.021*"aplicar
	0.266*"foto" + 0.180*"forte" + 0.053*"carro" + 0.047*"bandido" + ' '0.045*"mulher" + 0.037*"assalto" + 0.035*"pedir" + 0.032*"menino" + ' '0.029*"prender" + 0.022*"colocar
Saúde	0.173*"vacina" + 0.094*"pessoa" + 0.063*"alertar" + 0.057*"tratamento" + ' '0.039*"poder" + 0.038*"febre" + 0.037*"agulha" + 0.037*"funcionar" + ' '0.032*"existir" + 0.029*"amarelo
	0.140*"cura" + 0.106*"coronavirus" + 0.092*"falso" + 0.063*"medico" + ' '0.057*"matar" + 0.043*"alho" + 0.041*"populacao" + 0.040*"verdade" + ' '0.032*"agua" + 0.028*"reduzir
	0.219*"vacino" + 0.069*"causar" + 0.068*"anunciar" + 0.059*"video" + ' '0.042*"novo" + 0.035*"vitamina" + 0.031*"fabricar" + 0.029*"imunidade" + ' '0.029*"natural" + 0.024*"produto
	0.223*"mostrar" + 0.124*"causa" + 0.069*"foto" + 0.061*"virus" + ' '0.060*"cancer" + 0.057*"mulher" + 0.035*"Video" + 0.035*"morrer" + ' '0.023*"coca" + 0.023*"cola
Entretenimento	0.167*"morrer" + 0.082*"hoje" + 0.069*"vida" + 0.026*"cantor" + ' '0.025*"nacional" + 0.021*"jogar" + 0.020*"Video" + 0.019*"acidente" + ' '0.019*"carro" + 0.017*"jornal
	0.074*"cantar" + 0.060*"neta" + 0.057*"flagrar" + 0.056*"musica" + ' '0.035*"novo" + 0.029*"ganhar" + 0.028*"homem" + 0.026*"video" + ' '0.025*"homenagem" + 0.023*"pedir
	0.089*"bolsonaro" + 0.068*"politico" + 0.059*"lancar" + 0.053*"presidente" ' '+ 0.044*"camiseta" + 0.042*"posar" + 0.038*"apoio" + 0.030*"apoiar" + ' '0.029*"globo" + 0.028*"mulher
	0.094*"pessoa" + 0.090*"receber" + 0.054*"globo" + 0.046*"mostrar" + ' '0.037*"video" + 0.031*"demitir" + 0.029*"garoto" + 0.029*"vazar" + ' '0.028*"programa" + 0.019*"apresentar
	0.146*"morte" + 0.091*"flagrar" + 0.084*"policia" + 0.056*"ano" + ' '0.035*"foto" + 0.032*"circular" + 0.030*"filha" + 0.022*"candidato" + ' '0.019*"baile" + 0.019*"funk
Tecnologia	0.042*"dar" + 0.028*"site" + 0.012*"boticario" + 0.009*"natal" + ' '0.007*"passagem" + 0.005*"perfume" + 0.005*"Gratis" + 0.005*"gratis" + ' '0.005*"gratuito" + 0.005*"compartilhar
	0.006*"dar" + 0.004*"boticario" + 0.004*"ganhar" + 0.004*"compartilhar" + ' '0.004*"site" + 0.004*"virus" + 0.004*"mulher" + 0.004*"natal" + ' '0.004*"celular" + 0.004*"liberar
	0.007*"ganhar" + 0.005*"foto" + 0.005*"celular" + 0.004*"pedir" + ' '0.004*"mostrar" + 0.004*"mensagem" + 0.004*"video" + 0.004*"virus" + ' '0.004*"liberar" + 0.004*"online
	0.019*"facebook" + 0.009*"ganhar" + 0.008*"bloquear" + 0.008*"post" + ' '0.007*"compartilhar" + 0.006*"foto" + 0.006*"social" + 0.005*"mostrar" + ' '0.005*"reeleger" + 0.005*"bolsonaro"
	0.006*"celular" + 0.005*"mostrar" + 0.005*"credito" + 0.004*"ganhar" + ' '0.004*"whatsapp" + 0.004*"pedir" + 0.004*"video" + 0.004*"receber" + ' '0.004*"carro" + 0.004*"mensagem
	'0.005*"ganhar" + 0.004*"celular" + 0.004*"pedir" + 0.004*"liberar" + ' '0.004*"virus" + 0.004*"foto" + 0.004*"dar" + 0.004*"mensagem" + '

	'0.004*"online" + 0.004*"whatsapp
Ciência	0.022*"mostrar" + 0.022*"video" + 0.022*"foto" + 0.022*"causa" + ' '0.022*"otico" + 0.022*"falso" + 0.022*"ilusao" + 0.022*"pessoa" + ' '0.022*"cientista" + 0.012*"espaco
	0.031*"gigante" + 0.031*"ficar" + 0.031*"lugar" + 0.031*"russia" + ' '0.031*"interior" + 0.031*"nuclear" + 0.031*"meteorito" + 0.031*"concurso" + ' '0.031*"brasileiro" + 0.031*"pirassununguo
	0.051*"maquina" + 0.051*"quantico" + 0.051*"bolinha" + 0.051*"fisico" + ' '0.051*"cor" + 0.051*"separar" + 0.051*"causa" + 0.006*"mostrar" + ' '0.006*"mundo" + 0.006*"setembro
	0.009*"mundo" + 0.009*"mostrar" + 0.009*"setembro" + 0.009*"cientista" + ' '0.009*"video" + 0.009*"gigante" + 0.009*"foto" + 0.009*"causa" + ' '0.009*"pessoa" + 0.009*"criar
	0.051*"reproduzir" + 0.051*"crescer" + 0.051*"berinjela" + 0.051*"dentro" + ' '0.051*"escorpioes" + 0.051*"video" + 0.051*"mostrar" + 0.006*"mundo" + ' '0.006*"cientista" + 0.006*"setembro
	0.047*"cafe" + 0.047*"agua" + 0.047*"copo" + 0.047*"teste" + ' '0.047*"definir" + 0.047*"impuro" + 0.047*"puro" + 0.046*"video" + ' '0.046*"mostrar" + 0.006*"mundo
	0.009*"ilusao" + 0.009*"setembro" + 0.009*"pessoa" + 0.009*"ae ids" + ' '0.009*"invisivel" + 0.009*"cientista" + 0.009*"criar" + 0.009*"despovoar" + ' '0.009*"foto" + 0.009*"mostrar
Esporte	0.016*"time" + 0.012*"morrer" + 0.012*"premio" + 0.012*"grande" + ' '0.012*"colar" + 0.012*"flamengo" + 0.011*"goleiro" + 0.010*"selecao" + ' '0.009*"mostrar" + 0.009*"brasileirao
	0.020*"sopa" + 0.020*"sushi" + 0.020*"olimpiada" + 0.020*"causa" + ' '0.020*"expulso" + 0.012*"jogador" + 0.011*"copa" + 0.010*"perder" + ' '0.008*"vaza" + 0.008*"piscinar
	0.008*"temer" + 0.008*"diver" + 0.008*"perdoar" + 0.008*"receita" + ' '0.008*"federal" + 0.008*"prender" + 0.008*"matar" + 0.008*"humilhar" + ' '0.008*"neymar" + 0.008*"encontrar
	'0.005*"confirmar" + 0.005*"colar" + 0.005*"cigano" + 0.005*"briga" + ' '0.005*"beijar" + 0.005*"flagrar" + 0.005*"justica" + 0.005*"leiloar" + ' '0.005*"anunciar" + 0.005*"casar
Mundo	0.037*"mostrar" + 0.023*"video" + 0.017*"foto" + 0.006*"homem" + ' '0.006*"mulher" + 0.006*"carro" + 0.005*"encontrar" + 0.005*"presidente" + ' '0.005*"pessoa" + 0.005*"matar
	0.009*"prender" + 0.008*"bolsonaro" + 0.008*"biden" + 0.007*"senar" + ' '0.005*"terrorista" + 0.005*"amazonio" + 0.005*"cuidar" + 0.005*"bombardear" ' '+ 0.005*"ameacou" + 0.005*"apresentar
	0.010*"comunista" + 0.008*"trump" + 0.008*"voto" + 0.007*"guerra" + ' '0.007*"ficar" + 0.007*"armar" + 0.007*"acao" + 0.007*"declarar" + ' '0.007*"entrar" + 0.007*"socialista
	0.008*"capar" + 0.008*"dizer" + 0.008*"destacar" + 0.008*"time" + ' '0.007*"pai" + 0.006*"prejudicar" + 0.006*"ajudar" + 0.006*"argentina" + ' '0.006*"quarentena" + 0.006*"prometer
	0.008*"vazar" + 0.008*"genetico" + 0.008*"fruto" + 0.008*"Mail" + ' '0.008*"coronavirus" + 0.002*"ministro" + 0.002*"homem" + 0.002*"matar" + ' '0.002*"pessoa" + 0.002*"presidente
Religião	0.025*"mostrar" + 0.022*"papar" + 0.021*"foto" + 0.019*"morrer" + ' '0.012*"pastor" + 0.012*"novo" + 0.012*"igreja" + 0.011*"demonio" + ' '0.011*"flagrar" + 0.011*"rosto
	0.009*"livro" + 0.009*"seculo" + 0.009*"sir" + 0.009*"isaio" + ' '0.009*"guerra" + 0.009*"previu" + 0.008*"corpo" + 0.008*"globo" + ' '0.008*"novela" + 0.008*"boicote
	0.009*"criticar" + 0.009*"social" + 0.009*"movimento" + ' '0.009*"esquizofrenia" + 0.005*"vinda" + 0.005*"simular" + 0.005*"destroir" ' '+ 0.005*"estatua" + 0.005*"proibir" + 0.005*"temer
	0.005*"simular" + 0.005*"vinda" + 0.005*"estatua" + 0.005*"destroir" + '

	'0.005*"corpo" + 0.005*"querer" + 0.005*"temer" + 0.005*"proibir" + ' '0.005*"livro" + 0.005*"abrir

Fonte: Elaborado pelo autor (2023).

Analisando cada modelagem é possível perceber que a categoria de notícias falsas sobre política tem 4 tópicos em sua modelagem. O tópico 0 está relacionado ao governo federal e ao governo estadual. Ele é caracterizado pelos seguintes termos: “presidente”, “governador”, “temer” e “federal”. Outros termos incluem: “moro”, “acabar” e “sair”. Em geral, muitas notícias falsas sobre política de períodos diferentes incluem esses termos. Entre os assuntos que citam os termos associados a este tópico estão: o *impeachment* da ex-presidente Dilma Rousseff em 2016, a prisão do ex-presidente Lula em 2018, as eleições presidenciais de 2018 que levaram Jair Bolsonaro ao poder, e a pandemia de COVID-19 no Brasil, que gerou muitas discussões sobre a resposta do governo federal e estadual à crise.

O tópico 1 está relacionado ao presidente Jair Bolsonaro e sua administração. O tópico inclui o termo “bolsonaro”, mas também inclui “foto”, “globo”, “fraude” e “prender”. Em geral, existem muitas notícias falsas sobre política que incluem os termos citados. Exemplos de eventos associados a este tópico no período estudado incluem a eleição presidencial de 2018, quando Bolsonaro foi eleito, as controvérsias em torno de sua resposta à pandemia de COVID-19 e as tensões entre o governo federal e a Rede Globo.

O tópico 2 aborda questões militares e de segurança. É caracterizado pelos termos “militar”, “intervencao” e “general”, mas também inclui “vazar”, “eleicoes” e “policia”. As principais notícias falsas do *corpus* que citam esses termos estão associadas com a crise de segurança pública no Rio de Janeiro em 2018, as eleições presidenciais de 2018, que geraram muita discussão sobre a integridade do processo eleitoral e as tensões entre o governo federal e estadual. Por fim, o tópico 3 está relacionado a vídeos e imagens que supostamente mostram evidências de irregularidades e escândalos. É caracterizado pelos seguintes termos: “mostrar”, “video”, “ministro” e “decreto”. Também inclui “expulsar”, “aprovar” e “eleger”. Em geral, existe a presença de notícias falsas no *corpus* que buscam influenciar a opinião pública sobre questões políticas e candidatos.

A categoria de notícias relacionada a assuntos nacionais possui 4 tópicos em sua modelagem. O tópico 0 está relacionado a notícias falsas sobre crimes violentos, com alguns termos que indicam homicídios, tais como: “matar” e “bala”. Alguns dos

outros termos importantes incluem: “ano”, “menina”, “policia” e “greve”. O tópico 1 está relacionado a vídeos e imagens chocantes, compartilhados nas redes sociais, muitas vezes sem contexto ou precisão. Os termos mais importantes incluem: “mostrar”, “vídeo”, “morrer” e “bater”. Em geral, vídeos e imagens são constantemente usados para o embasamento de notícias falsas no *corpus*.

O tópico 2 está relacionado a multas de trânsito e informações falsas, e conta com termos, tais como: “multa”, “motorista” e “transito”. Como exemplo, a notícia falsa com título “*Novas multas do Detran a partir de hoje*” traz informações falsas sobre novos valores para as multas. Por fim, o tópico 3 novamente traz a temática de crimes violentos, incluindo assaltos, sequestros e outros tipos de violência. Os termos mais importantes incluem: “foto”, “forte”, “carro” e “bandido”. Em geral, notícias falsas desse tipo estão concentradas na categoria sobre assuntos nacionais.

Na modelagem associada a categoria de notícias sobre saúde há 4 tópicos. O tópico 0 traz como assunto principal, o tratamento de doenças e sintomas. Há também menções a termos, tais como: “alerta”, “tratamento”, “febre” e “agulha”. Em geral, as principais notícias contendo esses termos no *corpus* estão associadas a boatos sobre a pandemia de COVID-19. O tópico 1 está relacionado a tratamentos falsos para doenças, com destaque para a palavra "cura". Também há menções aos seguintes termos: “coronavirus”, “matar”, “alho” e “reduzir”. Entre os acontecimentos relacionados a esse tópico, podemos citar as diversas *fake news* relacionadas a falsos métodos de tratamento para o coronavírus que circularam nas redes sociais durante a pandemia.

O tópico 2 está relacionado à vacinação e produtos naturais, com destaque para a palavra “vacino” (versão lematizada de “vacina”). Também há menções aos seguintes termos: “anunciar”, “vitamina”, “imunidade” e “produto”. Um acontecimento que pode ser relacionado a este tópico é a polêmica envolvendo a eficácia das vacinas[31] contra a COVID-19 e a defesa do uso de produtos naturais como tratamento alternativo[32].

Por fim, o tópico 3 parece estar relacionado com doenças e imagens impactantes, com destaque para as palavras "mostrar" e "causa". Entre os termos mais relevantes também estão: “cancer”, “mulher”, “virus” e “morrer”. Em geral, partes

[31] Disponível em: https://informe.ensp.fiocruz.br/noticias/51261
[32] Disponível em: http://www.unifap.br/sociedade-de-farmacognosia-alerta-sobre-falsos-tratamentos-naturais-para-covid-19/

das notícias falsas disseminadas sobre sáude trazem imagens e vídeos para o seu falso embasamento, o que explica a ocorrência dos termos “mostrar”, “foto” e “video”. Por fim, os termos “coca”, “cola”, “cancer” remetem a boatos, tais como: “*Coca-Cola com abacaxi é um veneno e muitas pessoas morreram por isso*”.

A modelagem relacionada as notícias da categoria sobre entretenimento possui 5 tópicos. O tópico 0 está relacionado com informações falsas com eventos trágicos envolvendo mortes em acidentes, com a palavra "morrer" sendo a mais importante neste assunto. Palavras como “acidente”, “carro” e “nacional” também são bastante relevantes neste contexto. O tópico 1 está muito relacionado com música e artistas do mundo do entretenimento, com as palavras “cantar” e “neta” sendo as mais importantes. Em geral, a uma série de notícias com mesmo padrão utilizando esses dois termos associadas a cantores famosos, entre elas é possível citar: “*Neta de vocalista do Narazeth (Dan McCafferty) canta Love Hurts em homenagem ao avô*”, “*Neta de Elvis Presley canta em homenagem aos 80 anos do avô*”, “*Neta de Luciano Pavarotti canta música na TV e arranca aplausos*”, “*Neta de Agnaldo Rayol canta Ave Maria junto com avô*” e “*Neta de Roberto Carlos canta A Montanha (Obrigado, Senhor), mostra vídeo*”.

O tópico 2 associa figuras do entretenimento à política, com a palavra “bolsonaro” sendo a mais importante neste subconjunto. Outras palavras relevantes incluem: “presidente”, “politico” e “apoio”. Nesse caso, existem notícias falsas no *corpus* que indicam apoio de famosos a políticos. O tópico 3 está relacionado a programas de TV e vídeos, com a palavra “globo” sendo a mais importante neste subconjunto. Palavras como “mostrar” e “demitir” também são bastante relevantes aqui, indicando que este tópico tem relação com notícias falsas sobre pessoas que foram demitidas de emissoras de televisão. Por fim, o tópico 4 está ligado a notícias falsas envolvendo celebridades, onde as palavras “flagrar”, “foto” e “circular”, evidenciam falsos acontecimentos sobre famosos.

A categoria de notícias relacionada à tecnologia possui seis tópicos em sua modelagem. O tópico 0 traz como assunto principal falsas ofertas de brindes e promoções oferecidos por empresas e sites online durante o período de Natal. Os termos mais relevantes associados a esse tópico incluem: “dar”, “site”, “boticario”, “natal” e “perfume”. A oferta de brindes e promoções é comum nessa época do ano, e muitos boatos falsos são disseminados pela internet com a intenção de aplicar

golpes que prometem esses benefícios em troca de informações pessoais[33]. O tópico 1 está relacionado a golpes que visam enganar usuários de smartphones e outros dispositivos móveis, geralmente através de mensagens ou links maliciosos. Os termos mais relevantes associados a este tópico incluem: “celular”, “virus”, “ganhar”, “liberar” e “compartilhar”. As pessoas que caem nessas armadilhas podem acabar perdendo dinheiro ou tendo seus dados pessoais comprometidos. Portanto, é importante analisar esse tipo de notícia falsa.

O tópico 2 está relacionado à propagação de vírus através de mensagens e links maliciosos enviados em redes sociais. Os termos mais relevantes associados a este tópico incluem: “virus”, “mensagem”, “foto”, “online” e “pedir”. Esses golpes podem ser muito convincentes e difíceis de detectar, especialmente se parecem ter sido enviados por amigos e familiares[34]. O tópico 3 está relacionado a notícias falsas sobre políticos e suas campanhas eleitorais espalhadas nas redes sociais. Os termos mais relevantes associados a este tópico incluem: “facebook”, “ganhar”, “bloquear”, “post” e “bolsonaro”. A manipulação de informações políticas é um problema sério em muitos países, e as redes sociais têm sido amplamente utilizadas como plataforma para disseminação de notícias falsas e desinformação[35].

O tópico 4 traz como assunto principal, os golpes que visam enganar as pessoas através da oferta de crédito e veículos, geralmente solicitando informações pessoais e bancárias em troca. Os termos mais relevantes associados a este tópico incluem: “celular”, “credito”, “carro”, “receber” e “pedir”. As pessoas que são vítimas desses golpes podem acabar perdendo dinheiro e tendo seus dados pessoais comprometidos. Por fim, o tópico 5 está relacionado com golpes que usam o aplicativo de mensagens *WhatsApp* para enviar links maliciosos e coletar informações pessoais dos usuários. Os termos mais relevantes associados a este tópico incluem: “whatsapp”, “virus”, “mensagem”, “celular” e “liberar”. Os golpes no WhatsApp têm se tornado cada vez mais frequentes[36], e as pessoas precisam estar atentas com os mesmos.

---

[33] Disponível em: https://canaltech.com.br/apps/e-golpe-campanha-maliciosa-no-whatsapp-oferece-brindes-de-natal-128951/
[34] Disponível em: https://mundoconectado.com.br/artigos/v/17821/novo-golpe-duplica-whatsapp-e-pede-dinheiro-para-contatos
[35] Disponível em: https://agenciabrasil.ebc.com.br/geral/noticia/2019-09/estudo-aponta-manipulacao-politica-pela-internet-em-70-paises-em-2019
[36] Disponível em: https://einvestidor.estadao.com.br/comportamento/golpes-celular-crescem-50-psafe/

Portanto, ao analisar a modelagem realizada sobre a categoria de notícias sobre tecnologia foi possível constatar o meio de propagação de boatos e notícias falsas que podem agir até mesmo como golpes para obtenção de informação. Além disso, os termos encontrados por essa modelagem constituem uma grande ferramenta para identificação dos principais vetores de disseminação de desinformação.

Na modelagem associada a categoria de notícias sobre ciência há 7 tópicos no total. O tópico 0 tem como destaque os termos “mostrar”, “vídeo”, “foto”, o que destaca uma informação importante em relação as notícias falsas sobre ciência, ou seja, boa parte das desinformações nesse sentido vem acompanhada de vídeos e fotos falsos para embasar as informações falsas. O tópico 1 traz os termos “russia”, “concurso” e “nuclear”, que provavelmente estão associados ao seguinte boato: “*Brasileira Marcela Pereira ficou em 2° lugar no concurso de física nuclear na Rússia*”. Os termos “gigante”, “interior”, “meteorito” e “pirassununguo” (versão lematizada de “Pirassununga”) estão associados ao seguinte boato: “*Foto mostra buraco gigante feito por meteorito em Pirassununga, interior de São Paulo*”.

O tópico 2 traz um assunto específico, isto é, os termos “maquina”, “quântico”, “bolinha”, “cor” e “separar” reforçam o seguinte boato: “*Máquina separa bolinhas por cores por causa da física quântica*”. O tópico 3 traz entre os principais termos: “mundo”, “mostrar” e “setembro” não havendo um assunto claro. O tópico 4 traz termos como: “reproduzir”, “crescer”, “berinjela”, “dentro” e “escorpioes” que, por sua vez, reforçam o seguinte boato: “*Escorpiões cresceram e se reproduziram dentro de berinjela, mostra vídeo*”.

O tópico 5 trata sobre um boato a respeito de um suposto vídeo que mostra um teste simples para verificar se o pó de café é puro ou impuro. Entre os principais termos que destacam esse boato, é possível observar: “café”, “agua”, “copo”, “teste”, “definir”, “impuro” e “puro”. Por fim, o tópico 6 traz entre os principais termos: “ilusão” e “criar” relacionados com o boato: “*Dr. Yamamoto, professor de neurologia, criou imagem que mostra se você está cansado*”. Além dos termos “cientista” e “invisível” relacionados com o boato: “*Tecido que torna invisível é inventado por cientista na China*”. Portanto, o tópico traz como assunto principal invenções científicas falsas.

A modelagem relacionada as notícias da categoria sobre esporte possui 4 tópicos. O tópico 0 está relacionado ao futebol, com termos como: “time”, “premio”, “flamengo”, “goleiro” e “brasileirao”. O termo “goleiro” aprece em notícias falsas, tais

como: “*Hoax: Goleiro mexicano Ochoa, que fechou gol contra o Brasil, tem seis dedos*”, “*Goleiro Bruno se filia ao PSDB e é liberado da prisão*”, entre outras. Portanto, de maneira geral esse tópico reflete notícias falsas sobre questões particulares sobre atletas.

O tópico 1 aparece com termos como “sushi” e “sopa” que aparentemente não têm relação direta com o esporte. Entretanto, ao analisar as notícias falsas presentes no *corpus* é possível identificar o seguinte boato: “*Brasil pode ser expulso das Olimpíadas por causa de sopa de sushi*”. O tópico 2 tem termos como "temer", “federal” que estão relacionados com informações falsas sobre políticas públicas relativas ao esporte, como exemplo: “*Governo Temer vai suspender bolsa a atletas após as Olimpíadas*”. O único termo que se refere a algum atleta neste tópico é “neymar” que, por sua vez, também aparece relacionado ao termo “temer” no seguinte boato: “*Neymar tem dívida de R$ 200 milhões com a Receita Federal perdoada por Temer*”. O tópico 3 traz termos como “cigano” e “briga”, que estão relacionados ao seguinte boato: “*Júnior Cigano briga em bar*”. Portanto, o tópico traz desinformações sobre atletas de uma maneira geral.

A categoria de notícias relacionada a assuntos internacionais possui 5 tópicos em sua modelagem. O tópico 0 remete a notícias falsas com conteúdo visual, como fotos e vídeos. Os termos mais frequentes incluem: “mostrar”, “foto” e “video”. Também há menções de “homem”, “mulher”, “carro” e “presidente”. Ao observar as principais notícias do *corpus*, é possível observar que esse tópico abrange notícias falsas que utilizam imagens para difundir informações enganosas.

O tópico 1 apresenta termos como: “bolsonaro”, “biden”, “senar”, “terrorista” e “amazonio” (versão lematizada de “Amazônia”). Esses termos sugerem uma conexão com política internacional associada a questões ambientais, possivelmente com foco na relação entre o Brasil e os Estados Unidos. Os termos “prender” e “ameacar”, realçam notícias falsas que visam desacreditar figuras políticas. O tópico 2 se concentra em termos como: “comunista”, “socialista”, “trump”, “guerra”, “armar”, “declarar” e “voto”. Isso sugere um possível tema relacionado à polarização política e conflitos ideológicos, com menções ao presidente americano Donald Trump e à retórica socialista/comunista. Os verbos “armar” e “declarar” indicam uma possível associação com conflitos armados.

O tópico 3 traz termos como: “capar”, “destacar”, “prejudicar”, “argentina” e “quarentena”. Esses termos trazem notícias falsas associadas à pandemia COVID-19,

com foco em suas implicações econômicas para a Argentina. Notícias falsas, tais como: “*Argentina anuncia confisco de carros de cidadãos do país e povo protesta*”, “*Argentina faliu, quebrou todas empresas e implantou o comunismo com o apoio da China*”, “*Argentina anuncia confisco de carros de cidadãos do país e povo protesta*”, “*Governo da Argentina bloqueia todas contas bancárias do país*” e “*Argentina não existe mais porque virou a China comunista*”, comprovam a análise realizada. Além disso, temos que em todas essas notícias falsas é citada a COVID-19 como um dos fatores da falsa falência da Argentina.

Por fim, o tópico 4 apresenta termos relacionados ao coronavírus, como “genético”, “coronavirus” e “matar”. Portanto, é possível explicar a ocorrência desses termos devido ao grande número de notícias falsas relativas a pandemia de COVID-19 estarem presentes na categoria sobre assuntos internacionais.

Finalmente, na modelagem associada a categoria de notícias sobre religião há 4 tópicos. O tópico 0 traz termos como “mostrar”, “igreja” e “demonio”, que remetem a notícias falsas presentes no *corpus*, tais como: “*Diabo aparece na porta de igreja no meio do culto, mostra foto*”. O tópico 1 está relacionado com profecias bíblicas e teorias conspiratórias sobre eventos históricos. Entre os termos, temos: “livro”, “século”, “previu”, “Isaio” (lematização da palavra “Isaías”) e “guerra”. A notícia falsa “*Bíblia previu guerra da Síria do Século XXI no livro Isaías 17*” é um exemplo que comprova o sentido do tópico.

O tópico 2 está relacionado a notícias falsas sobre movimentos sociais e políticos. Alguns dos termos listados são: “criticar”, “social”, “movimento”, “destruir” e “estatua”. Embora esses termos não pareçam estar diretamente relacionados entre si, existem notícias falsas sobre movimentos sociais citando nome de religiosos no *corpus*. Por fim, o tópico 3 está associado a teorias conspiratórias envolvendo símbolos religiosos e eventos históricos. Os cinco termos mais importantes são “simular”, “vinda”, “estatua”, “destruir” e “corpo”. A notícia “*Blue Beam: Nasa vai simular 2ª vinda de Jesus Cristo*” é um exemplo de desinformação do *corpus* que utiliza os termos.

6.3 ABORDAGEM VIA LSA OTIMIZADA COM AJUSTE DE HIPERPARÂMETROS

Nesta abordagem foi utilizado o algoritmo LSA para a modelagem de tópicos. O algoritmo LSA foi anteriormente definido na seção 2.3, porém é importante ressaltar

ao contexto de aplicação no presente capítulo que o algoritmo utiliza semântica distribucional, isto é, analisa as relações entre um conjunto de documentos e os termos que eles possuem, e dessa forma produz os tópicos que, por sua vez, se traduzem como um conjunto de conceitos relacionados aos documentos e termos.

Nesse caso, temos que o LSA parte da hipótese distributiva, ou seja, as palavras com significados próximos ocorrem em textos semelhantes. Desse modo, ao considerar uma coleção de documentos (C), com d documentos, e sabendo que n é o número de palavras únicas na coleção C, então temos que o funcionamento do algoritmo está atrelado a construção de uma matriz M de dimensão $d \times n$, contendo o número de ocorrências de palavras por parágrafo, ou seja, a matriz possui linhas representando palavras únicas e colunas representando cada documento. Para isso, é necessário utilizar os textos dos documentos do *corpus* estudado. Após a construção da matriz, torna-se necessário utilizar a técnica matemática de decomposição de valor singular (SVD) que, por sua vez, tem por objetivo reduzir o número de linhas da matriz conservando a estrutura de similaridade entre colunas da mesma. Este processo pode ser feito tanto utilizando o cosseno do ângulo entre os dois vetores, quanto utilizando o produto escalar entre as formas normalizadas dos dois vetores, formado por duas linhas escolhidas. Entretanto, em linhas gerais a ideia consiste em decompor a matriz M da seguinte forma:

$$M = U\sum V^T, \text{ onde:}$$

- U: Distribuição de palavras entre os diferentes contextos
- $\sum$: Matriz diagonal de associação entre os contextos
- V^T: Distribuição de contextos entre diferentes documentos.

Nesse caso, ao realizar o cálculo, valores que se aproximam de um representam palavras muito semelhantes, enquanto valores que se aproximam de zero representam palavras bem distintas. Adicionalmente, é importante destacar que uma característica importante dessa decomposição consiste no fato de que a mesma permite realizar o truncamento de alguns contextos que não sejam necessários. Logo, a matriz $\sum$ traz em sua diagonal, valores que representam a significância do contexto. Portanto, ao utilizar os valores é possível reduzir a dimensão selecionando os k valores mais altos da diagonal da matriz $\sum$, e com isso é possível obter M_k, tal que:

$$M_k = U_k\sum_k V_K^T, \text{ onde:}$$

- M_k: Matriz aproximada de M.
- $U_k, \sum_k, V_K^T$: Matrizes contendo apenas os k contextos selecionados.

6.3.1 Modelagem de Tópicos e Otimização de Hiperparâmetro

Na abordagem descrita nesta seção foi utilizada a técnica *Latent Semantic Analysis* (LSA), onde toda a implementação do algoritmo foi realizada via Python com o uso da biblioteca *gensim*. Utilizando os dados resultantes do processo descrito na seção 6.1, foi criado o dicionário *id2word* com o uso da classe *Dictionary* da biblioteca *gensim* que, por sua vez, mapeia o identificador de uma palavra com seu respectivo *token*, e além disso foi criado a entrada (*bag-of-words*) para a utilização na modelagem.

Após isso, foi realizado o treinamento do modelo-base utilizado para experimentação inicial. Nesse sentido, o principal fator a ser considerado no treinamento foi o número K de tópicos, visto que esse é o principal hiperparâmetro para o treinamento do modelo final de tópicos. Dessa forma, em cada modelagem utilizando a técnica LSA foram treinados 17 modelos distintos, isto é, para cada modelo foi alterado o número de tópicos, considerando o intervalo $[4,20]$ como possibilidades de valores para K. Nesse sentido, ao treinar cada modelo foi extraída a coerência respectiva com o uso da classe *CoherenceModel*. Consequentemente, ao encontrar o número de tópicos com maior coerência resultante, o mesmo foi adotado como o valor do hiperparâmetro a ser utilizado no modelo final.

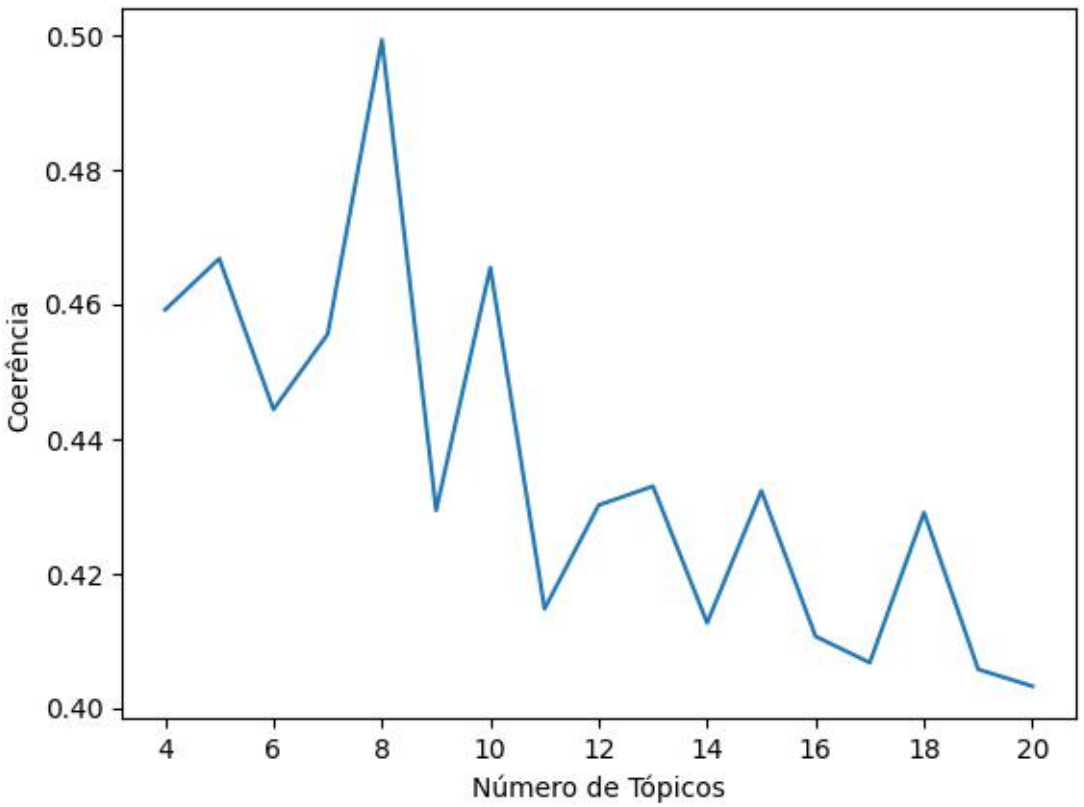
Portanto, ao considerar cada um dos três dicionários de tópicos resultantes descritos nas seções a seguir, temos que os scripts de treinamento possuem código semelhante. Desse modo, os modelos são treinados com o conjunto de dados específico de acordo com a separação detalhada na seção 6.1, e após isso, o valor da métrica de coerência de tópicos é calculado a partir dos dados de validação. Adicionalmente, ao finalizar cada treinamento são obtidos os tópicos apresentados por listas de palavras, onde cada lista representa um tópico. Nesse contexto, os dez termos mais relevantes de cada tópico são considerados nesta etapa. Além disso, temos que os valores associados aos termos dentro de cada tópico também são extraídos neste momento. Por fim, no Apêndice G é possível observar os valores de coerência para todos os valores de (K) usados na otimização da abordagem LSA

realizada tanto temporalmente, quanto categoricamente, e também na abordagem geral.

6.3.2 Dicionário de Tópicos Geral

Esta subseção traz os resultados obtidos, de forma geral. Desse modo, a descrição será realizada de forma a abordar a modelagem de tópicos para todos os documentos presentes no *corpus* de estudo. Tendo em vista o processo detalhado na subseção 6.3.1, foi realizada a modelagem para o *corpus* completo, isto é, considerando todos os documentos (notícias falsas) para a modelagem. Os valores de coerência para K no intervalo $[4, 20]$ podem ser vistos no Gráfico 53.

Gráfico 53 - Valor de coerência por quantidade de tópicos (LSA)



Fonte: Elaborado pelo autor (2023).

Com isso, é possível observar que o número de tópicos ótimo para modelagem é de 8 tópicos. Após a definição da quantidade de tópicos, a etapa seguinte corresponde ao treinamento e execução do algoritmo LSA com 8 tópicos, onde foi possível verificar os tópicos gerados com o respectivo valor atribuído a cada palavra em termos de significância, como pode ser visto no Quadro 6.

Quadro 6 – Modelagem de tópicos geral do corpus via LSA

Tópicos (0 à K-1)
0.294*"pessoa" + 0.228*"falar" + 0.170*"querer" + 0.158*"saber" + 0.155*"gente" + 0.155*"ficar" + 0.138*"passar" + 0.131*"bolsonaro" + 0.130*"presidente" + 0.129*"ano".
0.491*"bolsonaro" + -0.358*"pessoa" + 0.355*"presidente" + -0.231*"falar" + 0.210*"governo" + 0.202*"brasileiro" + -0.145*"tomar" + -0.135*"gente" + -0.126*"agua" + -0.113*"virus".
-0.553*"falar" + -0.298*"bolsonaro" + 0.201*"pessoa" + -0.189*"gente" + -0.175*"tomar" + 0.169*"receber" + 0.161*"mensagem" + -0.154*"querer" + 0.121*"ano" + -0.115*"presidente".
0.545*"pessoa" + 0.447*"bolsonaro" + -0.246*"ano" + 0.127*"virus" + -0.126*"querer" + 0.125*"presidente" + -0.124*"ficar" + -0.124*"mulher" + -0.122*"filho" + 0.122*"mensagem".
0.369*"mensagem" + 0.283*"bolsonaro" + -0.228*"governo" + -0.219*"presidente" + -0.181*"ano" + 0.170*"amigo" + 0.154*"contato" + -0.151*"brasileiro" + -0.149*"vacino" + 0.138*"compartilhar".
0.414*"falar" + -0.259*"bolsonaro" + 0.225*"governo" + 0.219*"mensagem" + -0.216*"ano" + 0.160*"receber" + -0.160*"morrer" + -0.158*"ficar" + 0.143*"site" + -0.136*"vida".
-0.338*"querer" + 0.312*"ano" + 0.265*"bolsonaro" + 0.207*"falar" + -0.195*"governo" + -0.192*"pegar" + -0.180*"presidente" + -0.178*"gente" + -0.159*"pessoa" + -0.139*"olhar".
-0.384*"acabar" + -0.362*"ano" + -0.323*"pessoa" + 0.193*"agua" + 0.173*"passar" + -0.159*"falar" + 0.152*"mensagem" + -0.143*"receber" + -0.131*"aumento" + -0.121*"partido".

Fonte: Elaborado pelo autor (2023).

A partir da determinação dos principais termos representativos, pode-se identificar, por meio de associação, um assunto específico para cada tópico. Assim, cada tópico foi associado ao seu respectivo assunto. É possível observar que apesar do uso da coerência (veja Gráfico 53) para a determinação da quantidade de tópicos ideais, o algoritmo gerou tópicos com assuntos em comum, ou seja, os tópicos 1 e 2 apresentaram termos associados ao tópico 0, bem como, os tópicos 3 e 6 também apresentaram termos relacionados ao tópico 0. Além disso, foram encontrados alguns tópicos que agregaram termos com pouca relação entre si, isto é, os *junk topics* (ALSUMAIT et al., 2009). Porém, com uma observação mais detalhada dos tópicos é possível obter informações relevantes ao presente trabalho. Abaixo, os tópicos T_0 ao T_7 são analisados em maiores detalhes.

- T_0: 0.294 * **pessoa** + 0.228 * **falar** + 0.170 * **querer** + 0.158 * **saber** + 0.155 * **gente** + 0.155 * **ficar** + 0.138 * **passar** + 0.131 * **bolsonaro** + 0.130 * **presidente** + 0.129 * **ano** – este tópico traz conceitos ligados a política, em geral, o grande número de verbos pode ser justificado devido ao processo de lematização

realizado no conjunto de dados com o objetivo de obter melhores resultados na modelagem.

- T_1: 0.491 * **bolsonaro** + -0.358 * **pessoa** + 0.355 * **presidente** + -0.231 * **falar** + 0.210 * **governo** + 0.202 * **brasileiro** + -0.145 * **tomar** + -0.135 * **gente** + -0.126 * **agua** + -0.113 * **vírus** – este tópico tem informações que podem ser divididas em ao menos dois temas relacionados: política e saúde. A presença de palavras como “**governo**”, “**presidente**”, “**bolsonaro**”, entre outras, indicam o primeiro tema, enquanto termos como “**agua**” e “**vírus**” indicam o segundo tema. Em geral, existem alguns termos com pouca conexão com o restante das palavras do tópico, como “**falar**”, “**tomar**” e “**gente**”.
- T_2: -0.553 * **falar** + -0.298 * **bolsonaro** + 0.201 * **pessoa** + -0.189 * **gente** + -0.175 * **tomar** + 0.169 * **receber** + 0.161 * **mensagem** + -0.154 * **querer** + 0.121 * **ano** + -0.115 * **presidente** – este tópico contém muitos verbos que, por sua vez, estão relacionados com a lematização realizada no *corpus* que agregou as conjugações do verbo no seu infinitivo, o que demonstra que as notícias muitas vezes refletem ações e acontecimentos realizados.
- T_3: 0.545 * **pessoa** + 0.447 * **bolsonaro** + -0.246 * **ano** + 0.127 * **virus** + -0.126 * **querer** + 0.125 * **presidente** + -0.124 * **ficar** + -0.124 * **mulher** + -0.122 * **filho** + 0.122 * **mensagem** – este tópico traz informações importantes porque novamente relaciona termos sobre política com termos sobre saúde, isto é, “bolsonaro”, “presidente”, “virus”, etc. Em geral, grande parte das notícias envolvendo vírus estão atreladas a pandemia de COVID-19, o que pode explicar esta relação, visto que durante a pandemia muitas notícias falsas foram divulgadas e endossadas tanto por políticos, quanto por figuras públicas (SHIN; WANG; LU, 2022).
- T_4: 0.369 * **mensagem** + 0.283 * **bolsonaro** + -0.228 * **governo**+ -0.219 * **presidente** + -0.181 * **ano** + 0.170 * **amigo** + 0.154 * **contato** + -0.151 * **brasileiro** + -0.149 * **vacino** + 0.138 * **compartilhar** – este tópico traz informações relevantes, principalmente ao considerar a palavra “**vacino**” que, por sua vez foi alterada pela lematização e, em muitas notícias analisadas nas etapas anteriores do trabalho estão relacionadas com notícias falsas envolvendo as vacinas de COVID-19 em desenvolvimento no período de pandemia.

- T_5: 0.414 * **falar** + -0.259 * **bolsonaro** + 0.225 * **governo** + 0.219 * **mensagem** + -0.216 * **ano** + 0.160 * **receber** + -0.160 * **morrer** + -0.158 * **ficar** + 0.143 * **site** + -0.136 * **vida** – este tópico possui palavras, tais como: “morrer”, utilizada ostensivamente para disseminação de boatos sobre mortes de figuras públicas. Além disso, novamente é possível perceber a palavra “Bolsonaro” em outro tópico.
- T_6: -0.338 * **querer** + 0.312 * **ano** + 0.265 * **bolsonaro** + 0.207 * **falar** + -0.195 * **governo** + -0.192 * **pegar** + -0.180 * **presidente** + -0.178 * **gente** + -0.159 * **pessoa** + -0.139 * **olhar** – novamente é possível perceber a palavra “Bolsonaro” entre os tópicos modelados, e no geral existe pouca ligação clara entre os termos agregados.
- T_7: -0.384 * **acabar** + -0.362 * **ano** + -0.323 * **pessoa** + 0.193 * **agua** + 0.173 * **passar** + -0.159 * **falar** + 0.152 * **mensagem** + -0.143 * **receber** + -0.131 * **aumento** + -0.121 * **partido** – neste tópico é possível perceber principalmente a palavra “partido” que traz o viés político para o tópico, entretanto novamente existe baixa ligação entre os termos agregados.

Portanto, de maneira geral, temos que a modelagem para o *corpus* completo obteve seu valor máximo de coerência com 8 tópicos que, corresponde ao número de categorias presentes no *corpus*. Entretanto, os tópicos não necessariamente trouxeram assuntos de cada categoria distintamente presente no *corpus*, isto é, em alguns tópicos foi perceptível a presença de termos que fazem referência a assuntos de uma mesma categoria de notícias. Além disso, é possível perceber que em alguns tópicos distintos existe a associação de assuntos semelhantes.

6.3.3 Dicionário de Tópicos por Período de Tempo

Esta subseção traz os resultados obtidos, de forma temporal. Logo, a descrição será realizada de forma a abordar de maneira cronológica a modelagem de tópicos dos períodos de tempo abrangidos no *corpus* de estudo. Adicionalmente, foram realizadas análises quantitativas e análises qualitativas dos resultados alcançados.

Considerando o processo descrito na subseção 6.3.1, e realizando o cálculo das coerências para os modelos treinados para cada ano presente no *corpus* foram encontrados os valores de coerência descritos na Tabela 3 que descreve os resultados quantitativos para a modelagem.

Tabela 3 - Evolução do valor da métrica de coerência de acordo com o número de tópicos para os períodos temporais

Número de Tópicos	2013-2015	2016	2017	2018	2019	2020	2021
4	0,5675	0,3601	0,5397	0,6170	0,4893	0,3932	0,3795
5	0,4687	0,4200	0,4012	0,5306	0,4306	0,4278	0,4146
6	0,6105	0,3783	0,5252	0,5510	0,4506	0,4990	0,4759
7	0,6384	0,4814	0,4332	0,5793	0,4025	0,3385	0,4394
8	0,6304	0,3393	0,4176	0,5753	0,4232	0,3700	0,3614
9	0,6329	0,3631	0,5237	0,5403	0,4538	0,3307	0,4322
10	0,6807	0,4367	0,4563	0,5167	0,3747	0,4554	0,3866
11	0,5906	0,4390	0,4288	0,4654	0,4393	0,3918	0,3682
12	0,5662	0,4089	0,4086	0,5218	0,3731	0,4098	0,3753
13	0,5910	0,3884	0,4869	0,5596	0,3823	0,4082	0,3622
14	0,6276	0,3759	0,4646	0,4982	0,4163	0,3787	0,3914
15	0,6260	0,3805	0,3662	0,4747	0,4190	0,3960	0,3545
16	0,5873	0,3929	0,4389	0,4754	0,3879	0,3682	0,4005
17	0,6791	0,3788	0,3930	0,4687	0,4038	0,3653	0,3614
18	0,5634	0,3579	0,3575	0,4436	0,3951	0,3518	0,3847
19	0,5312	0,3977	0,4530	0,4431	0,3689	0,3441	0,4007
20	0,5989	0,3149	0,4271	0,4616	0,3780	0,3744	0,4124

Fonte: Elaborado pelo autor (2023).

Os valores de coerência encontrados refletem a qualidade da modelagem realizada para cada quantidade de tópicos. Portanto, é importante ressaltar que mesmo utilizando esse valor como parâmetro para seleção da quantidade de tópicos (K) a ser utilizada na modelagem, existem casos onde o maior valor de coerência está vinculado a um número alto de tópicos. Porém, dependendo do *corpus* e seus respectivos documentos analisados, temos que a densidade dos dados pode ser baixa, e consequentemente a modelagem pode trazer termos repetidos em diferentes tópicos, o que impossibilita a modelagem com alto número de tópicos. Para esses casos, é importante verificar outros valores para K, que tragam alto valor de coerência.

Do ponto de vista qualitativo, temos na Tabela 4, a modelagem realizada para os anos presentes no *corpus*. Para cada tópico identificado (com as *Top* 10 palavras do tópico), é apresentado o ano em que se enquadra, a coerência máxima, a média das coerências, e o percentual de ganho de qualidade na modelagem considerando a métrica de coerência. Dessa forma, é possível identificar o quanto a abordagem voltada para análise de coerência pode influenciar na melhora da modelagem via LSA em relação à média. A modelagem com informações adicionais, tal como o valor singular de cada termo pode ser vista no Apêndice E.

Tabela 4 - Modelagem de tópicos temporal via LSA com estudo de coerência

Ano	Tópico	Máxima coerência	Coerência média	Percentual de melhora em relação à média
2013-2015	fazenda, querer, falar, animal, pessoa, tomar, vaca, saber, presidente, lugar	0,6807	0,5994	13,56%
	fazenda, querer, animal, falar, vaca, lugar, mandar, possível, filho, dizer			
	cama, gazin, saquinho, terra, comprar, cemiterio, vídeo, produto, dormir, tirar			
	atleta, gesto, presidente, brasileiro, proibir, imprensar, respeito, berro, lugar, manifestar			
	pessoa, passar, cama, tomar, enviar, problema, pesquisador, fazenda, doença, comprar			
	enviar, crime, referir, direito, jornalista, governador, cooperativismo, carta, artigo, apenas			
	filme, malevola, pessoa, doença, pesquisador, problema, tomar, filho, tratamento, deixar			
	evitar, mude, rumo, colisão, favor, filme, norte, curso, navio, falar			
2016	pessoa, falar, ficar, gente, ano, passar, saber, caso, governo, brasileiro	0,4814	0,3890	23,75%
	falar, causar, filha, quimico, senhor, microcefalia, produto, ministerio_saude, vacina, gente			
	falar, filha, senhor, caso, shopping, governo, causar, forte, estrutura, juntar			
	mulher, caso, direito, denunciar, pico_maximo, químico, vacino, produto, explicar, provocar			
	governo,brasileiro,mulher,congresso, boliviano,ficar,criança,gente,empresa, presidente			
	mulher, direito, considerar, ano, filha, senhor, ficar, saudita, denunciar, vacino			
	frango, carne, arsênico, comer, nitro, saude, nível, novo, risco, grupo			
2017	falar, pessoa, saber, mensagem, acontecer, gente, passar, ano, querer, entrar	0,5397	0,4424	21,99%
	globo, temer, jornalismo, mensagem, rede_globo, alinhar, noticia, contato, jornal, vincular			
	desconto, aereo, passagem, idoso, mensagem, contato, empresa, falar, gente, grande			
	fruta, comer, estomago, mensagem, comida, paciente, vazio, cancro, forma, fatia			
2018	bolsonaro, pessoa, ficar, querer, passar, saber, ano, falar, gente, presidente	0,6170	0,5130	20,27%
	forca, area, auxiliar, delegar, militar, intervencao_militar, policia, quartel, estadual, especial			
	bolsonaro, candidato, voto, passar, ajudar, votar, presidente, eleicao, urna, mensagem			
	olhar, gripe, comer, passar, vacino, paciente, chegar, ano, tomar, falar			

2019	acabar, pessoa, ano, falar, presidente, bolsonaro, aumento, salario, partido, ficar	0,4893	0,4110	19,05%
	acabar, aumento, presidente, salario, partido, falar, aprovar, redutor, aposentado, reduzir			
	presidente, bolsonaro, abaixo, amigo, site, pessoa, brasileiro, compartilhar, receber, casa			
	abaixo, site, falar, pessoa, receber, compartilhar, presidente, matar, ganhar", mensagem			
2020	pessoa, falar, vírus, bolsonaro, gente, querer, saber, agua, tomar, ficar	0,4990	0,3884	28,47%
	bolsonaro, agua, falar, café, tomar, moro, virus, pessoa, brasileiro, sangue			
	falar, cafe, bolsonaro, pessoa, sangue, tomar, virus, moro, descarregar, saber			
	agua, pegar, entender, mascara, morrer, querer, corpo, cara, pega_mascara, banho			
	agua, cafe, bolsonaro, pegar, colocar, falar, cadastro, site, novo, tomar			
	pessoa, virus, mundial, chines, empresa, receber, cadastro, ano, mundo, guerra			
2021	pessoa, querer, gente, presidente, bolsonaro, tomar, vacino, ficar, governo, covid	0,4759	0,3941	20,75%
	presidente, bolsonaro, governo, gente, covid, agua, brasileiro, pessoa, governador, vacino			
	vacino, tomar, dose, presidente, vacina, pessoa, produção, bolsonaro, covid, querer			
	governo, presidente, oxigenio, governador, produção, tomar, politico, acabar, dose, hospital			
	bolsonaro, oxigenio, governador, pessoa, covid, medico, receber, entregar, governo_federal, agua			
	cadastro, beneficio, informar, bolsonaro, familiar, agua, gente, site, receber, presidente			

Fonte: Elaborado pelo autor (2023).

Ao observar a Tabela 4 é possível perceber no triênio a presença de tópicos que versam sobre manifestações, e que refletem boatos sobre as manifestações[37] de junho de 2013 que, por sua vez, foram uma série de mobilizações ocorridas simultaneamente em muitas cidades do Brasil no ano de 2013. Além disso, o triênio é caracterizado por tópicos que refletem boatos sobre variados assuntos, o que corrobora as análises realizadas no capítulo 5. Observando de forma mais detalhada, temos no tópico 0 termos relacionados com fazendas, animais e questões rurais. As palavras-chave mais frequentes incluem: “fazenda”, “animal”, “vaca” e “presidente”. É possível que este tópico esteja associado a notícias falsas envolvendo propriedades rurais, como supostas invasões ou desapropriações ilegais, conflitos entre proprietários de terras e agricultores ou mesmo escândalos envolvendo políticos ou familiares, como exemplo, temos as seguintes notícias falsas presentes no *corpus*:

[37] Disponível em: https://agenciabrasil.ebc.com.br/geral/noticia/2023-06/junho-de-2013-entenda-o-cenario-de-insatisfacao-que-levou-a-protestos

"*Lulinha comprou a maior fazenda do mundo*" que faz menção ao filho do político Lula, e também "*Vídeo mostra 40 mil bois em carretas na fazenda de Lula em São Félix do Xingu (PA)*". No tópico 1, é possível constatar a relação dos tópicos com propriedades de terra e conflitos agrários. As palavras-chave mais frequentes incluem: "fazenda", "propriedade" e "possivel". É possível que este tópico esteja associado a notícias falsas envolvendo disputas de terra, como supostas invasões ou desapropriações ilegais, planos governamentais para demarcação de terras ou de preservação ambiental, entre outros. No período em questão, o Brasil vivia uma época de protestos populares contra a corrupção, e muitas notícias falsas podem ter sido criadas para influenciar a opinião pública.

No tópico 2, existem termos ligados a produtos vendidos em lojas ou comércios. As palavras-chave mais frequentes incluem: "cama", "gazin" e "terra". É possível que este tópico esteja associado a notícias falsas envolvendo produtos comercializados em lojas e comércios, como problemas de qualidade ou até mesmo fraudes. No período em questão, o Brasil vivia uma crise econômica (BARBOSA FILHO, 2017) e muitas empresas se envolviam em práticas comerciais questionáveis. O tópico 3 faz menção de termos atrelados a atletas e manifestações esportivas. As palavras-chave mais frequentes incluem: "atleta", "gesto" e "presidente". Esse tópico está associado a notícias falsas envolvendo desempenho de atletas, escândalos em clubes e decisões governamentais relativas ao esporte. O período em questão foi marcado por diversos eventos esportivos importantes no Brasil, como a Copa do Mundo de 2014, o que gerou um ambiente propício para o surgimento de notícias falsas sobre esses temas (KONDLATSCH, 2014).

O tópico 4 está relacionado a problemas de saúde e de diagnósticos médicos. As palavras-chave mais frequentes incluem: "pessoa", "doença" e "problema". Ao observar as principais notícias falsas do *corpus* no período estudado é possível perceber que esse tópico está associado a notícias falsas envolvendo diagnósticos médicos incorretos, tratamentos ineficazes ou até mesmo curas milagrosas para doenças graves. Notícias falsas sobre saúde são bastante comuns, e a disseminação dessas informações pode ter graves consequências para o público em geral. O tópico 5 traz um contexto relacionado com questões legais e judiciais. As palavras-chave mais frequentes incluem: "crime", "jornalista" e "governador". Esse tópico pode estar associado com notícias falsas envolvendo acusações, julgamentos e manipulação de processos judiciais. O período em questão foi marcado por diversos escândalos

políticos e criminais no Brasil, o que pode ter favorecido a criação e disseminação de notícias falsas sobre esses temas.

O tópico 6 traz termos associados a filmes e produções audiovisuais. As palavras-chave mais frequentes incluem: “filme” e “malevola”. Portanto, esse tópico está associado a notícias falsas envolvendo lançamentos de filmes e informações enganosas sobre a produção. Como exemplo, temos no *corpus* a seguinte notícia: “*Filme Malévola é uma propaganda do inferno, diz teoria da conspiração*”, que traz falsas informações no âmbito religioso. O tópico 7 traz como palavras-chave mais frequentes: “policia”, “onibus” e “acidente”. Entre as notícias envolvendo os termos, temos: “*História falsa: jovem pregador quase apanha no ônibus e faz mudo falar*” e “*Cobra é encontrada dentro de ônibus*”. No tópico 8 existem termos relacionados a críticas e avaliações de filmes. As palavras-chave mais frequentes incluem: “filme”, “critica” e “sucesso”. É possível que esse tópico esteja associado a notícias falsas envolvendo críticas negativas ou positivas feitas por especialistas em cinema, ou até mesmo informações equivocadas sobre o sucesso ou fracasso de determinados filme.

Por fim, o tópico 9 está relacionado ao fracasso e sucesso em determinados setores, como negócios e esportes. As palavras-chave mais frequentes incluem: “fracasso”, “negocio” e “crise”. Ao ler as principais notícias do *corpus* no período de estudo, é possível constatar que esse tópico está associado a notícias falsas envolvendo empresas em crise, falência e supostos escândalos financeiros, e até mesmo informações enganosas sobre o desempenho de equipes esportivas. O período em questão foi marcado por diversas crises econômicas e políticas no Brasil, o que pode ter levado à disseminação de notícias falsas para influenciar a opinião pública.

A modelagem para o ano de 2016 traz tópicos que envolvem o termo “microcefalia” e “ministério_saude” que, por sua vez, refletem o avanço do vírus Zika e da microcefalia no país[38], porém como as notícias presentes no *corpus* são falsas, temos que a maior parte das notícias relacionadas a esse assunto trazem falsas informações sobre a doença, e falsos acontecimentos relacionados. Além disso, novamente é possível constatar em 2016 tópicos que refletem boatos variados, tais como boatos sobre: “*Os frangos possuírem substância cancerígena, tal como o arsênico*”. Analisando de forma mais detalhada, temos no tópico 0 termos que

[38] Disponível em: https://agenciabrasil.ebc.com.br/geral/noticia/2016-12/avanco-do-zika-e-da-microcefalia-assustam-o-mundo-em-2016

remetem a questões gerais do cotidiano, como pessoas falando sobre suas vidas e experiências. As palavras-chave mais frequentes incluem: “pessoa”, “falar”, “ficar” e “gente”. Em geral, as notícias falsas associadas a esse tópico trazem informações enganosas compartilhadas em redes sociais ou aplicativos de mensagens instantâneas. No tópico 1, é possível constatar claramente a relação com à saúde pública, especificamente à epidemia do Zika vírus no Brasil em 2016. As palavras-chave mais frequentes incluem: “filha”, “quimico” e “microcefalia”. Esse tópico está associado a notícias falsas envolvendo a transmissão e prevenção do Zika vírus, bem como possíveis causas da microcefalia.

No tópico 2, temos que as palavras-chave mais frequentes incluem “caso” e “shopping”, o que traz associação com boatos isolados com baixa frequência no período de estudo. No tópico 3, existem menções a questões de gênero e violência contra a mulher. As palavras-chave mais frequentes incluem: “mulher”, “caso” e “direito”. Portanto, esse tópico está associado a notícias falsas envolvendo casos de violência contra a mulher, bem como discussões sobre direitos das mulheres. O tópico 4 está conectado com questões envolvendo a política brasileira em 2016, especificamente ao congresso e ao presidente. As palavras-chave mais frequentes incluem: “governo”, “brasileiro” e “presidente”. É possível que esse tópico esteja associado a notícias falsas ou informações enganosas relacionadas à situação política do Brasil na época, que inclusive remetem ao processo de *Impeachment* ocorrido no referido período.

No tópico 5, temos novamente questões de gênero e violência contra a mulher como pauta principal. As palavras-chave mais frequentes incluem: “mulher”, “direito” e “denunciar”. Portanto, existe clara associação a notícias falsas envolvendo casos de violência contra a mulher, bem como à conscientização sobre os direitos das mulheres. Por fim, temos no tópico 6 questões de saúde pública, especificamente à segurança alimentar. As palavras-chave mais frequentes incluem: “frango”, “carne” e “arsenico”. Ao analisar as principais notícias, é notável a relação do tópico com notícias falsas e informações equivocadas relacionadas à qualidade da carne de frango vendida no Brasil, bem como possíveis riscos à saúde dos consumidores. Como exemplo, temos o seguinte boato: “*75% dos frangos têm uma substância cancerígena, o arsênico*”.

Em 2017, é possível constatar tópicos que possuem termos sobre política como, por exemplo, termos que ocorrem em boatos, tais como: “*Rede Globo se alinha*

com a esquerda e passa a adotar a cor vermelha com o intuito de ajudar o partido PT". Além disso, é possível perceber a presença de boatos sobre saúde, isto é, "*comer frutas de estômago vazio evitar o câncer*"; sobre assuntos nacionais, ou seja, "*idosos terem direito a 50% de desconto em passagens aéreas*", entre outros. De forma mais detalhada, temos novamente no tópico 0, questões gerais do cotidiano. As palavras-chave mais frequentes incluem: "falar", "pessoa", "saber" e "mensagem". Portanto, podemos afirmar que o tópico traz informações enganosas compartilhadas em redes sociais e aplicativos de mensagens instantâneas. No tópico 1, existem menções específicas ao governo Temer e à Rede Globo. As palavras-chave mais frequentes incluem: "globo", "temer" e "jornalismo". Ao analisar as principais notícias falsas temos boatos sobre a cobertura jornalística da Rede Globo sobre o governo Temer na época.

O tópico 2 traz palavras-chave que incluem: "desconto", "aéreo" e "passagem". Analisando as principais notícias falsas, temos informações equivocadas relacionadas a promoções de passagens aéreas para idosos. Por fim, no tópico 3 temos questões de saúde pública, especificamente à alimentação saudável. As palavras-chave mais frequentes incluem: "fruta", "comer" e "estomago". Nesse sentido, existem no *corpus* notícias falsas com informações equivocadas sobre os benefícios de comer frutas para a saúde.

Em 2018, a modelagem traz no tópico 0, questões relacionadas à política brasileira em 2018, especificamente à eleição presidencial e ao candidato Jair Bolsonaro. As palavras-chave mais frequentes incluem: "bolsonaro", "pessoa", "ficar" e "presidente". Portanto, existe a associação com notícias falsas relacionadas à campanha eleitoral de Bolsonaro. No tópico 1 existem termos relacionados à segurança pública no Brasil e ao papel das forças militares na manutenção da ordem. As palavras-chave mais frequentes incluem: "força", "area" e "militar". Portanto, este tópico remete a notícias falsas relacionadas à atuação das forças militares e policiais em operações de combate à criminalidade. Nesse sentido, é importante destacar a intervenção federal na segurança pública do Estado do Rio de Janeiro[39] ocorrida nesse período.

O tópico 2 novamente se relaciona com assuntos relativos à política brasileira em 2018, especificamente às eleições presidenciais e à votação em Jair Bolsonaro. As palavras-chave mais frequentes incluem: "bolsonaro", "candidato" e "voto".

[39] Disponível em: https://www12.senado.leg.br/noticias/materias/2018/02/21/senado-autoriza-intervencao-na-seguranca-publica-do-estado-do-rio-de-janeiro

Realizando a leitura das principais notícias falsas presentes no *corpus* no referido período, temos informações enganosas relativas a eleição presidencial. Por fim, temos o tópico 3 abordando questões de saúde pública, especificamente à vacinação contra a gripe. As palavras-chave mais frequentes incluem: “gripe”, “vacina” e “paciente”. Entre as notícias falsas existem informações equivocadas sobre os benefícios da vacinação contra a gripe, bem como possíveis riscos à saúde de pessoas que não se vacinam.

No ano de 2019, a modelagem de tópicos traz termos como: “salario”, “aumento”, “aposentado” que remetem a notícias falsas sobre o reajuste para aposentados, pensões e demais benefícios pagos pelo INSS realizado em 2019. Além disso, é possível perceber o termo “compartilhar” contido no tópico 3 que, em boa parte das notícias falsas está associado ao intuito de disseminação da notícia, ou seja, o termo frequentemente está contido em sua própria estrutura. Detalhadamente, temos no tópico 0, termos relacionados à política brasileira em 2019, especificamente ao presidente Jair Bolsonaro e a questões econômicas como salários e aumento de preços. As palavras-chave mais frequentes incluem: “acabar”, “pessoa”, “ano” e “presidente”. Esse tópico está associado a notícias falsas e informações enganosas relacionadas às políticas econômicas do governo Bolsonaro.

O tópico 1 também está relacionado a questões econômicas, especificamente ao aumento de salário para aposentados e a redução de benefícios. As palavras-chave mais frequentes incluem: “acabar”, “aumento” e “salario”. Dessa forma, o tópico tem conexão com notícias falsas sobre políticas econômicas e sociais em relação aos aposentados. No tópico 2, existe menção ao presidente Jair Bolsonaro e sua imagem pública. As palavras-chave mais frequentes incluem: “presidente”, “bolsonaro” e “brasileiro”. De maneira geral, temos que esse tópico está associado com notícias falsas relacionadas à imagem pública do presidente e possíveis polêmicas envolvendo seu nome. Por fim, o tópico 3 traz termos sobre violência e crime no Brasil. As palavras-chave mais frequentes incluem: “falar”, “pessoa” e “matar”. Portanto, ao observar as principais notícias falsas presentes no *corpus*, existem boatos sobre crimes violentos e ameaças à segurança pública no país.

Em 2020, é importante destacar tanto a presença constante do termo “bolsonaro” em boa parte dos tópicos obtidos na modelagem, bem como a presença do termo “vírus” que, por sua vez, tem intrínseca relação com a pandemia de COVID-19, o que pode ser constatado ao ler as notícias falsas contidas no *corpus* nesse

período. De forma específica, é possível observar no tópico 1, a presença dos termos “vírus” e “café” no mesmo tópico, que nos remete para boatos, tais como: “*Café ou chá curam e previnem o coronavírus*”, que reforça os resultados achados no capítulo 5 do presente trabalho onde foi constatado que a maior parte das notícias de 2020 relacionadas a saúde tinham como tema assuntos relacionados a COVID-19, e em boa parte trazendo falsos métodos de cura para a doença. Observando cada tópico separadamente, temos o tópico 0 relacionado à pandemia de COVID-19 em 2020 e suas implicações na vida cotidiana das pessoas incluindo palavras-chave mais frequentes, tais como: “pessoa”, “falar”, “vírus” e “Bolsonaro”. Portanto, existe relação com informações equivocadas sobre a pandemia e sobre medidas de prevenção ao vírus.

No tópico 1, entre as palavras-chave mais frequentes temos: “Bolsonaro”, evidenciando notícias falsas relacionadas às políticas do governo em relação à pandemia e a possíveis declarações polêmicas do presidente. O tópico 3 está relacionado a questões de saúde e à pandemia de COVID-19, com destaque para os problemas enfrentados pelos profissionais de saúde na linha de frente incluindo termos como: “falar” e “Bolsonaro”. Entre os outros termos, existem indícios de que tratem de notícias falsas sobre a pandemia e sobre as medidas de proteção aos profissionais de saúde, bem como a possíveis controvérsias envolvendo o governo e os trabalhadores da saúde. No tópico 3, novamente temos a pandemia de COVID-19 e suas implicações na saúde pública, com destaque para a necessidade de medidas preventivas como o uso de máscaras e a higiene pessoal. As palavras-chave mais frequentes incluem: “agua”, “corpo” e “banho”. Portanto, os boatos trazem informações falsas sobre os cuidados preventivos durante a pandemia, bem como polêmicas envolvendo medidas governamentais de prevenção.

O tópico 4 também relacionado à pandemia de COVID-19 e às implicações econômicas e sociais da crise, traz destaque para o cadastro de pessoas em sites para receber ajuda financeira. Entre os termos mais frequentes, está em destaque “cadastro”. Notadamente, as notícias do *corpus* que envolvem esse termo fazem referência a medidas governamentais para ajudar as pessoas afetadas pela pandemia, bem como a possíveis problemas no processo de cadastro e distribuição dos recursos. Finalmente, o tópico 5 traz destaque para notícias falsas sobre a origem do vírus. As palavras-chave mais frequentes incluem: “china”, “virus” e “mundo”. Entre as informações equivocadas existem menções sobre a origem e disseminação do

vírus, bem como a possíveis controvérsias envolvendo relações internacionais e cooperação na luta contra a pandemia.

Em 2021, nos tópicos 0, 1 e 2 é possível perceber os termos: “covid”, “bolsonaro”, “vacino” (versão lematizada da palavra “vacina”) que, novamente corrobora os resultados obtidos no capítulo 5 onde foi constatado que a maior parte das notícias de 2021 relacionadas a saúde tratavam sobre a pandemia de COVID-19 com foco em boatos sobre as vacinas que naquele momento estavam em desenvolvimento e também começaram a ser aplicadas na população[40]. De forma detalhada, temos que o tópico 0, refere-se às implicações da pandemia na vida cotidiana das pessoas e traz temas ligados à vacinação contra a COVID-19, medidas de prevenção e contágio, bem como possíveis controvérsias envolvendo o presidente Bolsonaro. A palavra “pessoa” é uma das mais citadas nesse tópico, o que sugere que as discussões estão centradas nas experiências individuais dos brasileiros em meio à crise sanitária. Além disso, a palavra “vacina” aparece entre as mais frequentes, indicando que o processo de imunização contra a COVID-19 é um tema central entre as informações falsas.

O tópico 1 está relacionado ao governo federal e suas políticas em relação à pandemia de COVID-19. As palavras-chave mais frequentes incluem: “presidente”, “bolsonaro” e “brasileiro”. Assim, as discussões abordam desde as medidas adotadas pelo governo para combater o vírus até o posicionamento do presidente frente à crise sanitária. O tópico 2 se concentra na vacinação contra a COVID-19, com destaque para a produção e distribuição da vacina. As palavras-chave mais frequentes incluem: “vacina”, “tomar” e “dose”. Portanto, temos que as informações falsas estão centradas no processo de vacinação do país e em possíveis desafios e obstáculos encontrados ao longo desse processo. Além disso, a palavra “producao” aparece como uma das mais frequentes, indicando que os boatos abordam questões relacionadas à fabricação da vacina e sua disponibilidade no mercado. O tópico 3 traz a crise sanitária e à gestão governamental da pandemia incluindo entre os termos mais frequentes: “governo”, “oxigenio” e “hospital”. As principais notícias falsas do *corpus* sobre saúde nesse período trazem informações falsas sobre o fornecimento de oxigênio hospitalar durante a crise sanitária e às políticas do governo em relação à saúde pública. Assim, abordam desde a disponibilidade de recursos para os hospitais até possíveis medidas

[40] Disponível em: https://agenciabrasil.ebc.com.br/saude/noticia/2021-01/vacinacao-contra-covid-19-começa-em-todo-o-pais

adotadas pelo governo para minimizar os impactos da pandemia na saúde dos brasileiros.

O tópico 4 está relacionado ao presidente Jair Bolsonaro e possíveis controvérsias políticas. As palavras-chave mais frequentes incluem: “bolsonaro”, “covid” e “governo federal”, sugerindo informações falsas centradas nas políticas do governo federal em relação à pandemia. Dessa forma, abordam tanto declarações polêmicas do presidente quanto outras medidas adotadas pelo governo que geraram controvérsia entre a população. Por fim, o tópico 5 traz questões sociais e econômicas, com destaque para programas de auxílio financeiro durante a pandemia. As palavras-chave mais frequentes incluem: “cadastro”, “beneficio” e “informar”. Com isso, as discussões podem se concentrar nas medidas adotadas[41] pelo governo para ajudar as pessoas afetadas pela crise econômica decorrente da pandemia de COVID-19. Além disso, a palavra “familiar” também aparece entre as mais frequentes, indicando que o debate aborda questões relacionadas às famílias e às suas necessidades durante esse período difícil. Em resumo, a análise das palavras-chave em cada tópico relacionado à pandemia de COVID-19 no Brasil em 2021 permite compreender os principais temas discutidos na sociedade brasileira em meio à crise sanitária. As preocupações com a vacinação, as políticas do governo federal e os impactos sociais e econômicos da pandemia estão presentes em todas as discussões, evidenciando a complexidade desse cenário desafiador para todos.

Portanto, é possível concluir com a modelagem temporal que a análise exploratória de publicações, categorias e entidades voltada aos períodos de tempo presentes em um *corpus* pode atuar como ponto de apoio na interpretação de tópicos em abordagens temporais. Dessa forma, a análise exploratória realizada no capítulo 5 atuou como uma ferramenta de auxílio na interpretação dos tópicos obtidos nessa modelagem. Dessa forma, é possível entender que um estudo exploratório detalhado do *corpus* antes da modelagem pode facilitar o processo de modelagem e interpretação de tópicos a ser realizado. Principalmente, devido ao fato de que a análise temporal pode ressaltar os assuntos mais abordados em notícias falsas nos períodos estudados, facilitando a interpretação da modelagem que pode não ser clara em uma visualização inicial.

[41] Disponível em: https://www.poder360.com.br/economia/pagamento-do-auxilio-emergencial-comeca-nesta-3a-feira-veja-o-calendario/

6.3.4 Dicionário de Tópicos por Categoria

Esta subseção traz a descrição da aplicação da abordagem já descrita e os resultados obtidos, de forma categórica. Dessa forma, a descrição será realizada de forma a abordar de maneira temática a modelagem de tópicos das categorias abrangidas no *corpus* de estudo. Além disso, é importante destacar que esta seção também traz a análise e avaliação dos resultados alcançados tanto qualitativamente por meio da descrição e interpretação dos tópicos, quanto quantitativamente realizando a análise da métrica de coerência para validação do modelo.

Os valores de coerência encontrados refletem a qualidade da modelagem realizada para cada quantidade de tópicos. Portanto, é importante ressaltar que mesmo utilizando esse valor como parâmetro para seleção da quantidade de tópicos (K) para utilização na modelagem, existem casos onde o maior valor de coerência está vinculado a um número alto de tópicos. Porém, dependendo do *corpus* e seus respectivos documentos analisados, temos que a densidade dos dados pode ser baixa, e consequentemente a modelagem pode trazer termos repetidos em diferentes tópicos, o que impossibilita a modelagem com alto número de tópicos. Para esses casos, é importante verificar outros valores para K, que tragam alto valor de coerência.

Do ponto de vista qualitativo e quantitativo, temos na Tabela 5, a modelagem realizada para as categorias presentes no *corpus*. Para cada tópico identificado (com as *Top* 10 palavras do tópico), é apresentada a categoria em que se enquadra, a coerência de maior valor, a média das coerências para os modelos treinados com quantidade de tópicos no intervalo $[4, 20]$, e o percentual de ganho de qualidade na modelagem considerando a métrica de coerência. Dessa forma, é possível identificar o quanto a abordagem voltada para análise de coerência pode influenciar na melhora da modelagem via LSA em relação à média. A modelagem com informações adicionais, tal como o valor singular de cada termo pode ser vista no Apêndice E.

Tabela 5 - Modelagem de tópicos categórica via LSA com estudo de coerência

Categoria	**Tópico**	**Máxima coerência**	**Coerência média**	**Percentual de melhora em relação à média**
Política	bolsonaro, presidente, governo, acabar, querer, pessoa, brasileiro, saber, falar, pai	0,4335	0,3855	12,45%
	bolsonaro, acabar, ano, pessoa, pai, dinheiro, aumento, governo, salario, falar			

	presidente, acabar, bolsonaro, aumento, partido, governo, aprovar, pessoa, brasileiro, temer			
	globo, presidente, temer, falar, jornalismo, noticia, alinhar, rede_globo, jornal, governo			
	presidente, governo, acabar, federal, partido, mandar, aumento, brasileiro, governador, medico			
	querer, falar, governo, ano, mandar, lugar, povo, acabar", tomar, filho			
	governo, politico, pessoa, congresso, acabar, presidente, moro, dinheiro, mensagem, governador			
	moro, morar, morer, dizer, deputado, voto, acabar, congresso, querer, brasileiro			
Brasil	falar, pessoa, passar, ficar, saber, ano, gente, casa, acontecer, crianca	0,4979	0,4289	16,08%
	falar, filha, senhor, shopping, juntar, estrutura, forte, passar, mensagem, crianca			
	forca, area, auxiliar, delegar, aereo, partir, crianca, estadual, intervencao_militar, secretario			
	desconto, passagem", aereo, idoso, empresa, forca, ano, grande, adquirir, area			
	crianca, governo, receber, matar, saber, brasileiro, trabalhar, ficar, aereo, desconto			
Saúde	pessoa, falar, virus, tomar, gente, agua, pegar, vacino, causar, passar	0,4476	0,3929	13,92%
	causar, falar, zika, pegar, virus, gente, vacina, tomar, quimico, mascara			
	agua, pegar, mascara, falar, tomar, entender, governo, sangue, banho, corpo			
	agua, tomar, pegar, mascara, vacino, virus, entender, cafe, caso, ficar			
	falar, cafe, tomar, comer, fruta, virus, pessoa, vacino, hospital, sangue			
Entretenimento	querer, pessoa, morrer, globo, ano, ficar, vida, hoje, programa, deixar	0,4520	0,4220	7,10%
	globo, programa, vida, hoje, bolsonaro, aprender, dizer, tempo, rede, pensar			
	gente, cara, mundo, pessoa, luciano_huck, investimento, chegar, globo, querer, discutir			
	luciano_huck, investimento, produto, formular, globo, amostra, pessoa, tornar, pesquisa, ano			
	ano, morrer, globo, ator, cantor, dizer, bolsonaro, programa, deixar, falar			
	programa, presidente, ratinho, globo, podiar, ficar, apresentador, querer, falar, plateio			
	falar, podiar, globo, presidente, ficar, ator, ratinho, programa, saber, morrer			

	pedir, falar, podiar, homem, capitao, confederar, soldado, presidente, filho, cantor			
	cantor, brasileiro, querer, deixar, vida, show, presidente, viagem, capitao, dinheiro			
	falar, podiar, viagem, ficar, pedir, cantor, pai, levar, mensagem, filho			
	querer, achar, virtude, capacidade, vida, ano, pessoa, sonho, momento, bolsonaro			
Tecnologia	mensagem, site, receber, compartilhar, contato, amigo, abaixo, ganhar, presente, novo	0,6435	0,5055	27,29%
	mensagem, site, contato, enviar, abaixo, compartilhar, ganhar, presente, cadastro, celular			
	cadastro, presente, informar, site, beneficio, premio, abaixo, compartilhar, programa, novo			
	atualizar, presente, bloquear, contato, mensagem, repassar, amigo, deixar, pedir, ficar			
	presente, atualizar, compartilhar, novo, site, mensagem, abaixo, cadastro, premio, retirar			
	atualizar, informacoes, numero, chave, mensagem, hotel, casa, site, magnetico, poder			
Ciência	espelho, dia, insolacao, ficar, casa, poder, dentro, fenomeno, temperatura, perigoso	0,7869	0,6606	19,11%
	espelho, parte, espaço, refletir, unha, dia, insolação, mulher, imagem, teste			
	terra, climatico, natural, causar, orbitar, mudança, hoje, nasa, depender, ano			
	natural, climatico, terra, corona, espalhar, professor, dizer, orbitar, mudanca, afetar			
	pulso, sangue, cientista, frasco, conseguir, morte, soro, fracassar, valvula, vasilha			
	catastrofe, ilha, corona, trabalho, espalhar, deslizamento, professor, metro, dizer, dentro			
Esporte	jogador, atleta, jogo, flamengo, presidente, querer, falar, brasileiro, ficar, americano	0,6926	0,5394	28,40%
	jogador, falar, atleta, querer, americano, flamengo, mundo, tecnico, ficar, achar			
	atleta, falar, jogo, americano, terrorista, gente, hotel, casa, morar, mineirao			
	flamengo, presidente, gesto, americano, brasileiro, hotel, jovem, falar, jogar, saber			
Mundo	pessoa, mundo, pai, ano, homem, governo, virus, mulher, apenas, direito	0,5302	0,4544	16,68%
	mulher, virus, direito, pessoa, considerar, chines, morcego, riqueza, rato, mundial			
	mulher, direito, pessoa, considerar, especie, pai, veredicto, virus, muculmano, especialista			

	pessoa, muçulmano, pai, canada, mulher, prefeito, golfinho, entender, vida, falso			
Religião	pessoa, falar, querer, deus, igreja, saber, mensagem, senhor, orar, pedir	0,6038	0,4534	33,17%
	falar, bolsonaro, mudança, querer, saber, usar, mensagem, deus, enviar, ajudar			
	cama, gazin, terra, comprar, mensagem, saquinho, cemitério, tirar, vídeo, povo			
	ajudar, irmao, povo, colocar, gente, poder, vida, mensagem, senhor, crianca			
	senhor, igreja, encher, cama, jejum, mensagem, foto, gazin, pessoa, orar			

Fonte: Elaborado pelo autor (2023).

Considerando a abordagem de modelagem de tópicos por categorias, é possível perceber na modelagem da categoria sobre o tema política a presença constante do termo “bolsonaro” nos tópicos. O tópico 0 traz termos relacionados ao presidente Jair Bolsonaro e seu governo. As palavras-chave incluem: “bolsonaro”, “presidente” e “governo”. Também há menção à palavra “acabar”, o que sugere uma abordagem negativa em relação às ações do governo, conforme foi constatado ao ler as principais notícias falsas envolvendo os termos. No tópico 1, é possível perceber esse termo associado aos termos “aumento” e “salario”. Ao observar as notícias sobre política é possível encontrar notícias falsas sobre economia que relacionam os termos, trazendo um falso viés negativo referente as ações políticas do governo. O tópico 2 traz novamente a política brasileira de uma forma mais ampla, com destaque para o tema do aumento salarial. As palavras-chave incluem: “presidente”, “partido” e “aprovar”. Há menção à palavra “acabar”, o que sugere notícias falsas em relação às políticas governamentais.

No tópico 3, novamente é possível perceber os termos: “temer”, “globo”, “alinhar” e “jornalismo”, ao verificar as principais notícias do *corpus*, é possível perceber o alinhamento desses termos em notícias, tais como: “*Diretor do Fantástico revela plano da Globo para trazer Lula de volta*” e “*Globo se alinhou com a esquerda e vai adotar a cor vermelha*”. O tópico 4 está ligado ao papel do governo federal no Brasil. As palavras-chave incluem: “presidente”, “governo”, “federal” e “mandar”. Já o tópico 5 envolve questões de poder e liderança. As palavras-chave incluem: “querer”, “lugar”, “povo” e “tomar”. Há menção à palavra “acabar”, e tem relação a notícias falsas ligadas a golpes de estado. No tópico 6 e 7 é possível perceber a presença do termo “moro” que, por sua vez, teve muitas ocorrências constatadas na análise de entidades

realizada no capítulo 5 do presente trabalho, e entre as principais ocorrências existem notícias sobre corrupção e boatos difamatórios.

Na modelagem realizada sobre a categoria sobre assuntos nacionais é possível perceber tópicos bem definidos. O tópico 0 se concentra em aspectos gerais da vida cotidiana das pessoas, incluindo termos como: “falar”, “passar”, “ficar” e “ano”. Não há muita clareza sobre um tema específico, mas estão relacionados com ações cotidianas frequentemente abordadas em notícias falsas. O tópico 1 traz termos como: “filha”, “senhor”, “shopping” e “estrutura”. No entanto, não fica claro como esses termos estão inter-relacionados, visto que estão presentes em diferentes notícias falsas que abordam assuntos completamente distintos. O tópico 2 traz termos, tais como: “forca”, “estadual”, “intervencao_militar” e “secretario”, e que novamente trazem associações a notícias falsas criadas em decorrência da intervenção federal realizada no estado do Rio de Janeiro com o intuito de amenizar a situação de segurança interna em decorrência no estado no período vigente, entre as notícias existem boatos de intervenção militar e medidas alarmantes e destoantes das decisões realizadas na época.

No tópico 3 é possível encontrar termos, tais como: “desconto”, “passagem”, “aéreo”, “idoso”, que novamente tem relação com notícias falsas, tais como: “*Passagens aéreas devem ser vendidas com 50% de desconto para idosos*” e “*Idosos têm direito a 50% de desconto em passagens aéreas*”, entre outras. Por fim, o tópico 4 traz termos, tais como: “matar”, “criança”, “ficar” e “desconto”, o que não evidencia uma única temática, visto que os termos estão relacionados com diferentes contextos de notícias falsas.

Na categoria relacionada a saúde, temos que a modelagem trouxe tópicos que refletem assuntos de diferentes momentos, isto é, no tópico 1 é possível perceber a presença dos termos “zika”, “vírus” e “causar”, que estão principalmente associados a boatos sobre o vírus zika e a microcefalia. Nos tópicos 0, 2, 3 e 4 é possível perceber temos como: “vírus”, “tomar”, “agua”, “vacina”, “passar”, que estão tanto relacionados com boatos sobre falsas curas para a COVID-19, bem como boatos sobre as vacinas de COVID-19. De forma mais detalhada, temos que o tópico 0 está associado com notícias falsas sobre vírus e doenças, com termos como: “virus”, “agua”, “causar” e “vacina”. Os temas incluem informações enganosas sobre como o vírus se espalha, como se proteger ou tratar a doença, ou sobre a eficácia e segurança das vacinas, se

relacionando tanto com a pandemia de COVID-19 deflagrada no ano de 2020, quanto com o surto do vírus Zika no Brasil em 2015.

No tópico 1, há menção de termos, tais como: "máscara", "vacina" e "tomar", que novamente remetem a notícias falsas ligadas a pandemia de COVID-19, principalmente com a alta frequência de boatos ligados a falsos efeitos da vacina com recomendações errôneas contrárias a vacinação. O tópico 2 traz termos que se relacionam com notícias falsas presentes no *corpus* que trazem desinformação sobre o uso de máscaras e medidas de higiene não serem eficazes na prevenção da disseminação do vírus SARS-CoV-2. O tópico 3 novamente possui termos relacionados a notícias falsas sobre estratégias para prevenção do contágio do vírus e tratamento dos sintomas, com termos como: "água", "máscara" e "vírus". Há menção do termo "vacina" com correlação negativa, o que remete a informações enganosas sobre os riscos ou benefícios das vacinas. Além disso, ao analisar as notícias do *corpus* que possuem esses termos é possível observar a presença de desinformação sobre como as doenças se espalham e como se proteger delas.

O tópico 4 traz termos como: "café" e "fruta". No entanto, há menções de termos de saúde, como: "sangue" e "hospital". Entre as notícias falsas que contemplam esses termos, existem desinformações sobre como esses alimentos afetam a saúde, além de boatos sobre como tratar os efeitos da COVID-19 com falsas receitas.

A modelagem categórica relativa ao tema entretenimento traz o nome de diferentes figuras públicas, tanto de personalidades da mídia, quanto de políticos. Em geral, a maior parte dos tópicos reflete a tendência de criação de boatos a respeito de figuras públicas, o que pode ser comprovado no tópico 4 com a ocorrência de termos como: “luciano_huck”, “investimento”, “produto”, “pesquisa”, que faz referência a notícias falsas, tais como: “*Pílula para emagrecer da Unicamp ganha investimento de Luciano Huck*”.

Ao observar a modelagem de tópicos para a categoria tecnologia, o assunto mais importante e recorrente entre os tópicos tem relação com termos, tais como: “premio”, “presente”, ganhar”, “mensagem”, “repassar”, que na maior parte das notícias estão presentes em boatos envolvendo golpes com o uso da tecnologia. Em geral, a maior parte das notícias traz variados crimes de informação com o uso do pretexto de falsas premiações com o preenchimento de formulários, compartilhamento de informações, entre outras ações. Analisando minuciosamente cada tópico, é possível perceber no tópico 0, algumas palavras-chave relacionadas a mensagens e

compartilhamento em sites, com termos como: “mensagem”, “site” e “compartilhar”. Nota-se no *corpus* grande quantidade de notícias falsas com informações enganosas sobre promoções e brindes para incentivar o compartilhamento ou cadastro em determinados sites. De forma semelhante, o tópico 1 está relacionado a mensagens e contatos, com termos como: “mensagem”, “contato” e “enviar”. Em geral, as notícias que abrangem esses termos no *corpus* contêm desinformação e fraudes envolvendo mensagens falsas e *Sending and Posting Advertisement in Mass* (SPAMs) que solicitam informações pessoais ou financeiras.

No tópico 2, novamente há evidências de notícias falsas relacionadas a cadastros e benefícios, com termos como: "cadastro", "benefício" e "premio". Portanto, abordam informações enganosas sobre benefícios e recompensas oferecidas em troca do preenchimento de cadastros e inscrições em programas. O tópico 3 traz palavras-chave que se relacionam com ameaças e avisos, com termos como: “atualizar”, “bloquear” e “repassar”. Ao verificar as notícias falsas do *corpus* que contemplam esses termos, é possível encontrar informações enganosas e ameaças falsas sobre a necessidade de atualizar aplicativos e sistemas, bem como sobre bloqueios de dispositivos e contas.

O tópico 4 está relacionado a presentes e prêmios, com termos como: “presente”, “premio” e “retirar”. Portanto, remete a boatos que trazem informações enganosas sobre a disponibilidade e condições para resgatar falsos presentes e prêmios anunciados em sites. Por fim, o tópico 5 remete a informações pessoais e de localização, com termos como: “informacoes”, “numero” e “chave”. Ao analisar as notícias associadas a este contexto é possível identificar desinformação e fraude envolvendo coleta de informações pessoais ou rastreamento de localização sem consentimento do usuário.

Na categoria ciência, temos que a modelagem traz termos, tais como: “nasa”, “mudança”, climatico”, “natural”, “terra” no tópico 2, o que remete a notícias falsas sobre informações enganosas relacionadas com causas e consequências das mudanças climáticas, bem como sobre as medidas necessárias para reduzir suas consequências. Como exemplo no *corpus*, vide: “*Nasa admite que mudanças climáticas ocorrem por causa da órbita da Terra e não por ação do homem*”. No tópico 4, temos os termos: “pulso”, “sangue”, “cientista”, “morte”, que se relacionam com boatos presentes no *corpus* que contém informações enganosas e exageradas sobre pesquisas médicas e descobertas científicas, tais como: “*Cientista do Arizona faz*

experimento onde condenado à pena de morte morreu de infarto ao participar de experiência em que achava que perdeu muito sangue".

No tópico 5, existem termos, tais como: "ilha" e "catastrofe" que remetem ao seguinte boato: "*Nasa confirmou profecia de tsunami no Brasil após recuo do mar*". Além disso, ainda no tópico 5 existem palavras-chave relacionadas à disseminação de informações falsas e teorias da conspiração sobre a pandemia de COVID-19, com termos como: "corona", "espalhar" e "dizer". Além disso, temos que a presença do termo "insolação" nos tópicos 0 e 1, bem como a presença dos termos "fenômeno" e "temperatura" no tópico 1 estão relacionados com o seguinte boato: "*Fenômeno do equinócio nos atingirá nos próximos 5 dias e trará calor ao Brasil*". No tópico 3 existe correlação a eventos e tragédias envolvendo desastres naturais, com termos como: "catastrofe", "deslizamento" e "ilha". Dentre as notícias falsas, se destacam assuntos, tais como: a prevenção e o gerenciamento de desastres naturais, bem como sobre a eficácia de medidas de segurança.

A modelagem realizada na categoria esportiva pode ser útil para entender melhor que tipo de informação falsa foi disseminada no período de cobertura do *corpus* sobre o tema esportivo. O tópico 0 traz notícias falsas que tratam de jogadores e atletas, com palavras como "jogador", "atleta" e "brasileiro" aparecendo com alta frequência. Alguns termos negativos também aparecem nesse tópico, como "querer" e "ficar", sugerindo insatisfação com os jogadores ou com o estado atual do futebol. Esses termos sugerem a discussão sobre o desempenho dos jogadores brasileiros e suas transferências entre os clubes. No contexto do futebol, as notícias falsas podem apresentar informações inverídicas sobre jogadores e atletas, criando uma percepção distorcida sobre sua qualidade e valor. Elas também podem exagerar ou inventar informações sobre transferências[42], gerando interesse e engajamento em torno de um evento que pode não ocorrer.

O tópico 1 está associado à opinião das pessoas sobre os jogadores e atletas, com palavras como: "falar", "querer" e "achar" aparecendo com alta frequência. Em resumo, o tópico 1 sugere que as notícias falsas trazem opiniões sobre os jogadores e atletas brasileiros em comparação com os jogadores estrangeiros. Já o tópico 2 está relacionado a eventos esportivos específicos, com palavras como: "jogo", "mineirao" e "hotel" aparecendo com alta frequência. A palavra "terrorista" é um pouco

[42] Disponível em: https://monitor7.r7.com/fake-news-de-dirigente-grandes-contratacoes-que-ficaram-no-quase-07072022

surpreendente, mas pode estar relacionada a algum evento esportivo que ocorreu em meio a preocupações de segurança. Em resumo, o tópico 2 traz a percepção de que as notícias falsas do *corpus* também versaram sobre eventos esportivos específicos e possíveis falsas informações sobre a segurança dos locais onde ocorrem. Finalmente, o tópico 3 parece estar relacionado a um clube específico, o Flamengo, com palavras como: “flamengo” e “jogar” aparecendo com alta frequência. O tópico também menciona outros termos relacionados ao futebol, como: “brasileiro” e “jovem”. As palavras mais comuns mencionadas nesse tópico incluem: “flamengo”, “brasileiro”, “jogador” e “jogar”. Esses termos trazem notícias falsas envolvendo o desempenho dos jogadores e do clube em geral.

Na modelagem sobre assuntos internacionais, temos que o tópico 0 está relacionado com questões gerais de direitos humanos, com destaque para a pessoa e o governo. Alguns outros termos relevantes incluem: “mulher”, “homem” e “direito”. Portanto, trata de possíveis notícias falsas sobre o tema de igualdade de gênero, por exemplo. No tópico 1 existe uma grande relação ao vírus e à pandemia, com destaque para a palavra “virus”. No entanto, também há alguns termos variados, tais como: “morcego”, “rato” e “riqueza”, que indicam a presença de outros temas de desinformação. O tópico 2 se relaciona com questões de direitos humanos, com um foco específico na mulher e no direito. Há também menções de outros termos, embora não esteja claro como esses termos se relacionam com os demais presentes no tópico. A presença da palavra “muculmano” sugere um tema relacionado ao contexto da religião muçulmana. Por fim, o tópico 3 é bastante diverso, com termos aparentemente desconexos, tais como: “golfinho”, “canada” e “vida”. No entanto, a palavra “pessoa” aparece como a mais relevante, o que pode indicar que se trata de um tópico geral, possivelmente envolvendo notícias falsas variadas. A presença da palavra “falso” pode confirmar essa suspeita.

Por fim, ao considerar a modelagem realizada na categoria religiosa podemos obter algumas inferências, ou seja, temos o tópico 0 relacionado com notícias falsas sobre religião. Os termos mais relevantes incluem: “pessoa”, “deus”, “mensagem” e “orar”. No tópico 1, também existem termos relacionados com notícias falsas sobre religião, mas com um foco político. Os termos mais relevantes incluem: “bolsonaro”, “mudanca”, “querer” e “enviar”. Em geral, a maior parte desses termos está presente em notícias falsas que busca influenciar a opinião pública se aproveitando da religiosidade das pessoas.

No tópico 2, temos que o resultado da modelagem não parece estar diretamente relacionado a notícias falsas sobre religião, mas existe evidência de que os tópicos remetem notícias falsas utilizadas como isca para atrair leitores. Os termos mais relevantes incluem: “gazin”, “cama”, “terra”, “saquinho” e “cemiterio”, que podem ter sido utilizados para criar uma história sensacionalista, sem necessariamente ter uma ligação real com a religião. O tópico 3 tem relação com notícias falsas sobre religião que exploram questões sociais e humanitárias. Os termos mais relevantes incluem: “ajudar”, “gente”, “vida” e “crianca”. A presença da palavra “poder” pode indicar a presença de críticas a alguma instituição religiosa ou líder religioso. Por fim, o tópico 4 está relacionado com notícias falsas sobre discursos e mensagens religiosas. Os termos mais relevantes incluem: “senhor”, “igreja”, “jejum” e “orar”. Dessa forma, temos que esse tópico pode estar relacionado com notícias falsas que promovem uma mensagem religiosa específica, com o objetivo de manipular a crença das pessoas.

Concluindo a análise da modelagem de tópicos por categoria realizada, é possível afirmar que a separação do *corpus* por categoria de notícia garante que a análise das informações falsas seja realizada com mais precisão e possibilita um melhor entendimento de como essas notícias são estruturadas. Desse modo, se torna mais fácil identificar padrões específicos de linguagem e estilo em diferentes tipos de notícias falsas, bem como entender quais são os principais temas que estão sendo explorados em cada categoria.

Adicionalmente, as análises exploratórias que contemplaram as publicações de notícias falsas, as categorias e as entidades citadas ao longo do tempo, realizadas no capítulo 5, foram extremamente uteis na interpretação da modelagem de tópicos. Ao identificar as entidades e assuntos presentes nas notícias falsas, foi possível entender melhor como elas estão sendo utilizadas para enganar o público e difundir informações errôneas. Por exemplo, foi possível descobrir que certas entidades são frequentemente utilizadas em notícias falsas de saúde, enquanto outras são mais comuns em notícias falsas políticas, como partidos políticos ou líderes governamentais.

Portanto, a ideia utilizada de unificar as análises realizadas com a modelagem trouxe *insights* importantes sobre como as notícias falsas são produzidas e disseminadas, além de ajudar a desenvolver um processo mais eficaz para o estudo. Desse modo, a modelagem de tópicos baseada nas categorias e a análise exploratória

devem ser consideradas ferramentas fundamentais para qualquer estudo sobre notícias falsas que vise entender melhor esse fenômeno complexo.

6.4 COMPARAÇÃO ENTRE AS ABORDAGENS

Esta subseção tem como objetivo realizar uma breve discussão comparativa entre as duas abordagens utilizadas para modelagem de tópicos no presente trabalho. Neste capítulo, serão apresentados os resultados de duas abordagens diferentes de modelagem de tópicos: LDA e LSA. A modelagem de tópicos é uma técnica amplamente utilizada em diversas áreas para identificar e analisar padrões em conjuntos de dados textuais. Reiterando fatos importantes já abordados no presente trabalho, é importante destacar que a abordagem *Latent Dirichlet Allocation* (LDA) é baseada em probabilidades e assume que um documento é gerado a partir de uma mistura de tópicos, enquanto a abordagem *Latent Semantic Analysis* (LSA) se baseia em técnicas algébricas para encontrar relações semânticas entre termos e documentos.

O objetivo deste capítulo é comparar os resultados obtidos por ambas as abordagens em relação à capacidade de identificação e interpretação dos tópicos encontrados, bem como em relação ao desempenho na classificação de novos documentos. Os resultados desta comparação podem fornecer *insights* relevantes sobre qual abordagem é mais adequada para cada tipo de aplicação e ajudar a orientar futuras pesquisas nessa área. Nesse caso, é importante destacar que as duas abordagens basearam seu processo de otimização mediante o uso da métrica de coerência. Portanto, a presente seção utilizou esta métrica para os estudos comparativos.

Inicialmente é importante observar os valores de coerência para as modelagens realizadas nas divisões categóricas das notícias falsas do *corpus*. Para isso, é possível observar na Tabela 6 os valores de coerência de cada abordagem, bem como a categoria de notícias modelada.

Tabela 6 - Comparação entre a coerência ótima das abordagens LDA e LSA por categoria

Coerência ótima C_V		
Categoria	**LDA**	**LSA**
Política	0,6615	0,4335
Assuntos Nacionais (Brasil)	0,6033	0,4979
Saúde	0,4963	0,4476
Entretenimento	0,6090	0,4520
Tecnologia	0,5481	0,6435
Ciência	0,5096	0,7869
Esporte	0,5800	0,6926
Assuntos internacionais (Mundo)	0,5115	0,5302
Religião	0,5738	0,6038

Fonte: Elaborado pelo autor (2023).

Em geral, ao considerar a métrica de coerência, a abordagem utilizando o método LDA obteve melhor resultado de modelagem nas seguintes categorias: Política, Assuntos Nacionais (Brasil), Saúde e Entretenimento. Por outro lado, a abordagem utilizando o método LSA obteve melhor resultado de modelagem nas seguintes categorias: Tecnologia, Ciência, Esporte, Assuntos internacionais (Mundo) e Religião. Essa informação é relevante, pois ao considerar as categorias com maior quantidade de documentos (notícias), então é possível observar que as mesmas abrangem melhores resultados pelo método LDA. Analogamente, ao considerar as categorias com menor quantidade de documentos (notícias), então é possível constatar que são as que contém os melhores resultados pelo método LSA. Portanto, no contexto desse trabalho, o método LDA obteve melhores resultados em conjuntos de documentos maiores, enquanto que o método LSA obteve melhores resultados em conjuntos de documentos menores.

Após isso, é importante verificar também analisar os valores de coerência obtidos nas modelagens realizadas nas divisões temporais do *corpus* para observar os resultados. A Tabela 7 traz as informações relativas a cada abordagem de acordo com o período de tempo modelado.

Tabela 7 - Comparação entre a coerência ótima das abordagens LDA e LSA por período de tempo

Coerência ótima C_V		
Ano	**LDA**	**LSA**
2013-2015	0,5511	0,6807
2016	0,4952	0,4814
2017	0,4802	0,5397
2018	0,5172	0,6170
2019	0,4735	0,4893
2020	0,4929	0,4990
2021	0,5108	0,4759

Fonte: Elaborado pelo autor (2023).

Considerando novamente a métrica de coerência para comparação, é possível observar na Tabela 7 que a modelagem dos anos de 2016 e 2021 tiveram melhores resultados ao considerar a abordagem que utilizou o método LDA, enquanto que os conjuntos de 2013-2015, 2017, 2018, 2019 e 2020 obtiveram melhores resultados considerando a abordagem que utilizou o método LSA. Além disso, ao considerar as modelagens gerais realizadas com as duas abordagens, então é possível constatar que o método LSA atingiu 0,4992, enquanto que o método LDA atingiu 0,6136. Portanto, mais uma vez é possível observar que o método LDA consegue melhores resultados ao lidar com *corpus* com maiores quantidades de documentos.

Do ponto de vista qualitativo, tanto o método LDA, quanto o método LSA, conseguiram capturar termos que se destacaram nos conjuntos de documentos estudados. Apesar disso, ambos os métodos também capturaram termos iguais em tópicos distintos em determinadas modelagens. Entre os períodos temporais modelados, é possível observar que nos anos de (2013-2015), 2016, 2019, o LDA apresentou resultados mais concisos com termos próximos em cada tópico, apresentando melhor junção de termos relacionados ao tema específico de cada tópico, enquanto que o LSA se destacou nos anos de 2017, 2018, 2020 e 2021, trazendo termos que refletem assuntos mais importantes para o período de tempo estudado. Considerando as modelagens realizadas por meio da separação categórica do *corpus*, é possível observar que o método LDA obteve tópicos concisos nas seguintes categorias: política, assuntos nacionais (Brasil), entretenimento, tecnologia e esporte. Por outro lado, o método LSA obteve melhor associação entre os termos em cada tópico em categorias, tais como: saúde, ciência, assuntos internacionais (Mundo) e religião.

Além disso, ao considerar as modelagens do *corpus* completo para cada abordagem, é possível constatar que o LDA atinge coerência de 0,6136, enquanto que o método LSA atinge coerência de 0,4992. Portanto, quantitativamente a abordagem LDA teve vantagem ao considerar todos os documentos do *corpus*. Além disso, ao observar qualitativamente cada tópico criado em cada modelo é possível confirmar que os tópicos gerados por meio da abordagem LDA estão mais concisos e trazem assuntos mais bem definidos.

Por fim, observando os ganhos de coerência obtidos após a otimização das modelagens de tópicos com a abordagem proposta no presente trabalho foi possível constatar que considerar o período temporal, a categoria (temática) e as entidades

nomeadas (análise) como fatores determinantes para identificação de padrões em notícias falsas tende a ser um processo eficiente, o que comprova nossa hipótese, visto que considerando esses fatores no processo de divisão dos dados para modelagem de tópicos foram obtidos ganhos médios de coerência consideráveis, ou seja, foi possível obter modelagens com melhores resultados no presente trabalho.

7 CONCLUSÃO

O presente trabalho teve como principal justificativa a abordagem da problemática da disseminação de desinformação, ou seja, considerando a princípio o estudo Digital 2022: Global Overview Report, publicado pelo site Datareportal, que mostrou a grande quantidade de tempo dispendido por usuários nas mídias digitais, associado as perspectivas encontradas nos estudos (VOSOUGHI et al., 2018; GUO et al., 2019a; SU et al., 2020), que nos mostrou a potencialidade da propagação de conteúdo de desinformação (notícias falsas, boatos, rumores, etc), é possível compreender os riscos dessa problemática para a sociedade, visto que a mesma tende a resultar em diferentes consequências para sociedade, tais como as possibilidades de: Influência em processos democráticos (CANTARELLA et al., 2022), entre os exemplos, temos: eleições de 2016 nos Estados Unidos da América (EUA), e eleições de 2022 no Brasil; Prejuízos ao contingenciamento de crises de saúde, onde o estudo (ROCHA et al., 2021) nos mostrou o impacto da desinformação no contingenciamento da COVID-19 que, por sua vez, é um exemplo recente da problemática; Possibilidade de ocasionamento de crises sociais, onde o artigo (SHAHEEN, 2022) mostra claramente como a desinformação foi um dos fatores associados ao surgimento da atual polarização no Paquistão.

Dessa forma, o presente estudo se propôs a explorar essa problemática, com foco no desenvolvimento de abordagens computacionais eficazes para o combate à desinformação, tendo como base um *corpus* de estudo não-balanceado. Em vista do aspecto do *corpus* ser não-balanceado, a alternativa mais comum pode ser o balanceamento do mesmo.

Entretanto, o processo de balanceamento requer, tanto recursos temporais, quanto recursos computacionais, visto que depende da coleta de uma quantidade muito grande de dados. Dessa maneira, a abordagem proposta no presente trabalho consistiu na identificação de padrões textuais nos documentos do *corpus* de estudo por meio de técnicas de modelagem de tópicos visando a identificação de tópicos latentes e, consequentemente, foram estudados os avanços temporais dessas notícias por meio da investigação da atualização dos padrões ao longo dos períodos temporais, e também das temáticas existentes nas notícias falsas estudadas. Desse modo, essa metodologia tende a ser extremamente útil no combate a desinformação, na medida em que permite a identificação desses padrões que podem ser usados

como ferramenta tanto para o contingenciamento da disseminação de desinformação, como para o estudo do fenômeno.

O objetivo geral de analisar temporalmente as notícias falsas disseminadas na rede entre os anos de 2013 e 2021 foi desenvolvido com sucesso no presente trabalho, tendo em vista que foi possível explorar temporalmente métricas pertinentes ao entendimento da problemática, tais como: quantidade de publicações de notícias falsas, temáticas envolvidas nos conteúdos desinformacionais, entidades nomeadas mais recorrentemente citadas nesses conteúdos, bem como a relação dessas entidades com cada temática específica.

Adicionalmente, os objetivos específicos de realizar a construção do dicionário de tópicos visando a identificação de padrões e caracterização dos conteúdos de desinformação estudados, bem como a verificação da evolução dos padrões empregados nesses conteúdos também foram realizados com sucesso, onde foi possível realizar a proposição de uma abordagem específica no processo de modelagem de tópicos considerando a otimização desses modelos fazendo uso da métrica de coerência referentes aos mesmos, assim como foi possível realizar tanto a análise qualitativa, quanto a análise quantitativa dos mesmos.

A presente pesquisa partiu da hipótese de que o fator temporal (períodos temporais), o fator categórico (categorias/temáticas) e as entidades nomeadas presentes nos conteúdos de desinformação (notícias falsas, boatos, rumores, etc) estudados são fatores determinantes na identificação de padrões nos textos desses conteúdos. Durante o presente trabalho essa hipótese foi testada e comprovada tanto na análise temporal realizada, onde foi possível observar que as entidades nomeadas moldaram padrões específicos de conteúdo de desinformação que, por sua vez, foram analisados em detalhes nas seções 5.3 e 5.4, bem como no processo de construção do dicionário de tópicos via modelagem de tópicos no capítulo 6, onde foi possível constatar que com o processo proposto no presente trabalho visando considerar o fator temporal e o fator categórico na divisão dos dados estudados foi possível obter ganhos de qualidade nas modelagens de tópicos utilizadas para identificação de padrões considerando a métrica de coerência estudada conforme pode ser visto na Tabela 8.

Tabela 8 - Comparação dos ganhos relativos as divisões e técnicas aplicadas

Divisão do *Corpus*	**LDA**	**LSA**
Geral	12,23%	23,8%
Temporal	25,58%	21,12%
Categórica	42,11%	19,35%

Fonte: Elaborado pelo autor (2023).

Paralelamente, com este estudo foi possível demonstrar a importância de uma análise temporal e categórica relativa ao *corpus* de notícia ao qual se deseja estudar, isto é, a importância do contexto para realizar a análise de notícias, visto que as notícias não são subjetivas em sua construção, e dependem ao meio ao qual foram construídas. Com isto, verificou-se que tanto a categoria principal de uma notícia, quanto as principais entidades envolvidas podem ser determinantes na caracterização de um *corpus* com relação aos seus tópicos, e dentro de seus tópicos as palavras mais frequentes também tem importância significativa na caracterização de um conjunto de notícias.

Além disto, foi possível constatar que o fator temporal é determinante na caracterização de notícias falsas, ou seja, podemos afirmar que o período temporal em que uma notícia falsa foi construída pode tanto trazer informações relevantes para o entendimento de sua propagação, quanto pode ser um fator chave para entender o principal objetivo decorrente da fabricação desta notícia.

Neste sentido, podemos afirmar que a modelagem de tópicos atua como processo central na caracterização e entendimento da fabricação de textos com informações falsas, mais especificamente neste trabalho, na caracterização e entendimento da fabricação de notícias falsas, visto que por meio da modelagem de tópicos podemos categorizar os *tokens* (palavras) das notícias do *corpus* em diferentes tópicos e entender a influência dos *tokens* mais utilizados em cada tópico, e além disto, podemos verificar a relação destes tópicos tanto do ponto de vista temporal, quanto do ponto de vista categórico, se considerarmos as análises temporais e categóricas realizadas no presente trabalho.

Por fim, foi possível constatar no presente trabalho como a divisão categórica e a divisão temporal trouxeram efeitos benéficos para as modelagens de tópicos, e suas respectivas otimizações, visto que ao realizar essas divisões foi possível obter melhores valores de coerência para as modelagens realizadas e, consequentemente, melhores resultados no processo de modelagem de tópicos aplicado a textos de

notícias falsas. Desse modo, considerar o tempo e a temática de uma notícia falsa para sua análise e modelagem de tópicos é um fator determinante para a identificação de padrões nas mesmas.

7.1 CONTRIBUIÇÕES

Nesta seção, descreveremos as principais contribuições resultantes do presente trabalho. Durante o desenvolvimento da pesquisa, foram obtidos resultados significativos que fortaleceram o campo de estudo relacionado ao combate à desinformação e a análise de notícias. As contribuições destacadas são as seguintes:

1. Artigo apresentado na Semana de Integração Acadêmica da UFRJ (SIAC/UFRJ): Como subproduto desta pesquisa, um artigo foi submetido e apresentado no maior evento de ensino, pesquisa e extensão da UFRJ na renomada conferência (SIAC/UFRJ). O referido artigo recebeu avaliações extremamente positivas por parte da banca de avaliação, composta tanto por acadêmicos de ciências humanas quanto por acadêmicos de ciências exatas e tecnologia. Após a apresentação, um debate enriquecedor foi realizado, suscitando reflexões positivas sobre o benefício do artigo no combate à desinformação.
2. Revisão de literatura: Uma contribuição significativa deste trabalho foi a realização de uma revisão abrangente da literatura no campo do combate à desinformação utilizando metodologias computacionais. Essa revisão detalhou os principais estudos recentes e avanços nesse campo, fornecendo um panorama abrangente das abordagens e técnicas mais promissoras. Essa revisão de literatura serve como um recurso valioso para pesquisadores e profissionais interessados em se aprofundar nessa área específica.
3. Fundamentação teórica abrangente: Outra contribuição importante deste trabalho foi a elaboração de uma fundamentação teórica abrangente que explorou os principais conceitos relacionados à desinformação e ao combate à desinformação, bem como às metodologias computacionais associadas a esse campo. A fundamentação teórica foi cuidadosamente estruturada e articulada de forma clara e compreensível, tornando-a acessível a indivíduos com diferentes níveis de conhecimento sobre o assunto. Essa contribuição permite que qualquer pessoa interessada, independentemente de seu nível de expertise, compreenda

os fundamentos essenciais necessários para abordar o tema da desinformação e suas estratégias de combate.

4. Processo de tratamento do *corpus* estudado: O processo de tratamento dos dados neste trabalho consistiu em um sólido sistema de filtragem de informações não relevantes, utilizando expressões regulares. Essa abordagem permitiu remover conteúdos indesejados, tais como: ruídos e informações irrelevantes para a análise, garantindo a qualidade dos dados utilizados na pesquisa.

5. Aprimoramento do *corpus* Fakepedia: Outra contribuição significativa deste trabalho é o aprimoramento do *corpus* Fakepedia, que consiste em um conjunto de dados e metadados sobre as notícias estudadas. Esse aprimoramento foi realizado por meio da técnica de *Webscraping*, permitindo a coleta de informações mais abrangentes e detalhadas. O *corpus* Fakepedia aprimorado se tornou uma valiosa fonte de dados para análises futuras relacionadas ao combate à desinformação. O *corpus* aperfeiçoado está disposto no apêndice H.

6. Análise dos dados: A análise dos dados coletados concentrou-se em aspectos temporais e questões sobre as notícias investigadas. Foram realizadas análises das seguintes métricas:

- Quantidade de publicações de notícias ao longo do tempo, identificando possíveis flutuações e tendências.
- Quantidades de notícias por categoria (temática de notícia), visando compreender a distribuição e importância relativa de cada categoria.
- Identificação das entidades nomeadas mais frequentes nas notícias, permitindo uma compreensão mais aprofundada dos principais atores mencionados no contexto da desinformação.
- Identificação das entidades nomeadas mais frequentes entre categorias, analisando possíveis conexões e padrões entre diferentes temas.

7. Dicionário de Tópicos construído por meio da modelagem de tópicos: Uma das principais contribuições deste trabalho é a construção de um dicionário de tópicos utilizando duas abordagens amplamente reconhecidas (*Latent Dirichlet Allocation* e *Latent Semantic Analysis*). O processo de construção do dicionário proposto incluindo a metodologia de aplicação das duas técnicas mais tradicionais de modelagem de tópicos neste estudo é inédito considerando todas

as particularidades e especificidades, e trouxe benefícios relevantes para a modelagem de tópicos no contexto do combate à desinformação.

Todas essas contribuições significativas estão disponíveis ao decorrer do presente trabalho e nos apêndices do presente trabalho, facilitando o desenvolvimento de pesquisas futuras no campo de estudo voltado ao combate à desinformação com o uso de técnicas computacionais. A organização e formalidade na descrição dessas contribuições oferecem uma base sólida para futuros estudos e avanços científicos nessa área em constante evolução.

Portanto, concluímos que o presente trabalho alcançou resultados relevantes e proporcionou contribuições substanciais para o campo do combate à desinformação, englobando a submissão e apresentação de um artigo na SIAC/UFRJ, a construção de uma base sólida de referencial teórico, a análise detalhada de trabalhos pertinentes ao campo de pesquisa na revisão de literatura, o tratamento do *corpus*, o aperfeiçoamento do *corpus*, a análise das notícias estudadas e a construção do dicionário de tópicos via modelagem de tópicos.

7.2 TRABALHOS FUTUROS

Com o objetivo de aplicar a análise realizada, bem como utilizar o dicionário de tópicos construído neste trabalho, podem ser realizados trabalhos futuros no sentido de realizar a construção de um classificador probabilístico com o uso do dicionário de tópicos construído e descrito no capítulo 6 do presente trabalho, para realizar a detecção de notícias falsas, com a aplicação de algoritmos de *Machine Learning* voltados para a classificação, tais como: Regressão linear, *Random Forest*, *Support Vector Machine* (SVC), *Gradient Boosting*, *Naive Bayes*, entre outros. Para isto, seria necessário realizar a construção de um *dataset* balanceado (contendo uma fração de notícias falsas, bem como uma fração de notícias verdadeiras), realizando o trabalho de coleta de novas notícias, com o uso de técnicas de *Web Scraping* para balancear o *corpus* usado como base neste estudo.

Além dessa vertente, se torna possível a utilização do dicionário de tópicos para realização de pesquisas qualitativas com humanos, com o intuito de verificar o quanto as palavras advindas das modelagens de tópicos realizadas no presente trabalho podem ser relevantes para o auxílio na detecção manual de notícias falsas feitas por humanos. Nesse sentido, pode ser útil a realização de pesquisa qualitativa via

formulários, questionários e entrevistas com o objetivo de apresentar notícias falsas dos períodos de tempo estudados com o intuito de obter respostas dos entrevistados sobre as suas percepções em relação a veracidade das notícias sem o uso do dicionário, bem como respostas sobre a percepção de veracidade das notícias com o auxílio dos dicionários. Por conseguinte, pode ser feita a validação da eficiência dos dicionários construídos no presente trabalho, comparando os resultados obtidos nessas pesquisas qualitativas antes e após a inserção dos dicionários como ferramentas de auxílio aos entrevistados.

Por fim, outro possível trabalho futuro pode se concentrar na utilização do *corpus* que foi aprimorado no presente trabalho para a criação de uma nova base de dados integrados, que contenha em sua construção informações mais específicas relativas as notícias armazenadas. Desse modo, pode ser útil a adição de modelagens de tópicos nos conjuntos de dados que podem ser utilizados para estudos futuros, principalmente devido ao entendimento criado no presente trabalho em relação ao benefício das análises exploratórias como ferramenta auxiliar para interpretação de tópicos, bem como o uso das modelagens de tópicos para contextualização e entendimento dos principais assuntos usados para disseminação de notícias falsas tanto do ponto de vista temporal, quanto do ponto de vista categórico.

REFERÊNCIAS

ALSUMAIT, Loulwah; BARBARA, Daniel; GENTLE, James; *et al*. Topic Significance Ranking of LDA Generative Models. *In*: [s.l.: s.n.], 2009, p. 67–82.

BARBOSA, Fernando de Holanda. A crise econômica de 2014/2017. **Estudos Avançados**, v. 31, p. 51–60, 2017.

BASTICK, Zach. Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. **Computers in Human Behavior**, v. 116, p. 106633, 2021.

BHATIA, Sanjiv. Regular Expressions. **Computer Apex**, 2005.

BLEI, David M. Latent Dirichlet Allocation.

BOYD-GRABER, Jordan; HU, Yuening; MIMNO, David. Applications of Topic Models.

CANALTECH. **É golpe | Campanha maliciosa no WhatsApp oferece brindes de Natal**. Canaltech. Disponível em: <https://canaltech.com.br/apps/e-golpe-campanha-maliciosa-no-whatsapp-oferece-brindes-de-natal-128951/>. Acesso em: 19 jul. 2023.

CANTARELLA, Michele; FRACCAROLI, Nicolò; VOLPE, Roberto. Does fake news affect voting behaviour? **Research Policy**, v. 52, n. 1, p. 104628, 2023.

CAUCASO, Osservatorio Balcani e. **Information Manipulation: A Challenge for Our Democracies**. Media Freedom Resource Centre OBCT. Disponível em: <https://www.rcmediafreedom.eu/Publications/Reports/Information-Manipulation-A-Challenge-for-Our-Democracies>. Acesso em: 30 jun. 2023.

CCI/ENSP. **Conheça 6 “fake news” sobre as vacinas contra a Covid-19**. Disponível em: <https://informe.ensp.fiocruz.br/noticias/51261>. Acesso em: 15 jul. 2023.

CHOWDHURY, Gobinda G. Natural language processing. **Annual Review of Information Science and Technology**, v. 37, n. 1, p. 51–89, 2003.

COLOMINA, Carme; MARGALEF, Héctor SÁNCHEZ; YOUNGS, Richard. The impact of disinformation on democratic processes and human rights in the world.

FALEIROS, Thiago De Paulo. **Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais**. Doutorado em Ciências de Computação e Matemática Computacional, Universidade de São Paulo, São Carlos, 2016. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-10112016-105854/>. Acesso em: 19 jul. 2023.

GAN, Jingxian; QI, Yong. Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. **Entropy**, v. 23, n. 10, p. 1301, 2021.

GANESAN, Kavita. What is Term-Frequency? Disponível em: <https://kavita-ganesan.com/what-is-term-frequency/>. Acesso em: 3 maio 2023.

GELFERT, Axel. Fake News, False Beliefs, and the Fallible Art of Knowledge Maintenance. *In*: BERNECKER, Sven; FLOWERREE, Amy K.; GRUNDMANN, Thomas (Orgs.). **The Epistemology of Fake News**. [s.l.]: Oxford University Press, 2021, p. 0. Disponível em: <https://doi.org/10.1093/oso/9780198863977.003.0015>. Acesso em: 3 maio 2023.

GUO, Bin; DING, Yasan; SUN, Yueheng; *et al*. The mass, fake news, and cognition security. **Frontiers of Computer Science**, v. 15, n. 3, p. 153806, 2021.

HARDE, Pooja. An Experimental Technique on Text Normalization and its Role in Speech Synthesis. 2019.

HASTINGS, Peter. Latent Semantic Analysis. *In*: [s.l.: s.n.], 2004.

JELODAR, Hamed; WANG, Yongli; YUAN, Chi; *et al*. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. 2018. Disponível em: <http://arxiv.org/abs/1711.04305>. Acesso em: 3 maio 2023.

KONDLATSCH, Rafael. A venda da Copa do Mundo: uma análise do boato que se perpetua a cada mundial da Fifa. **Dito Efeito - Revista de Comunicação da UTFPR**, v. 5, n. 7, 2014. Disponível em: <https://periodicos.utfpr.edu.br/de/article/view/2705>. Acesso em: 15 jul. 2023.

LAZER, David M. J.; BAUM, Matthew A.; BENKLER, Yochai; *et al*. The science of fake news. **Science**, v. 359, n. 6380, p. 1094–1096, 2018.

LOVINS, Julie Beth. Development of a stemming algorithm.

MAY, Chandler; COTTERELL, Ryan; VAN DURME, Benjamin. An Analysis of Lemmatization on Topic Models of Morphologically Rich Language. 2019. Disponível em: <http://arxiv.org/abs/1608.03995>. Acesso em: 3 maio 2023.

MAY, Chandler; COTTERELL, Ryan; VAN DURME, Benjamin. An Analysis of Lemmatization on Topic Models of Morphologically Rich Language. 2019. Disponível em: <http://arxiv.org/abs/1608.03995>. Acesso em: 3 maio 2023.

MELO, Tiago de; FIGUEIREDO, Carlos M. S. Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach. **JMIR Public Health and Surveillance**, v. 7, n. 2, p. e24585, 2021.

MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; *et al*. Efficient Estimation of Word Representations in Vector Space. 2013. Disponível em: <http://arxiv.org/abs/1301.3781>. Acesso em: 3 maio 2023.

MONTEIRO, Rafael A.; SANTOS, Roney L. S.; PARDO, Thiago A. S.; *et al*. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. *In*: VILLAVICENCIO, Aline; MOREIRA, Viviane; ABAD, Alberto; *et*

al (Orgs.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2018, v. 11122, p. 324–334. (Lecture Notes in Computer Science). Disponível em: <http://link.springer.com/10.1007/978-3-319-99722-3_33>. Acesso em: 3 maio 2023.

NETTLETON, David. Chapter 11 - Text Analysis. *In*: NETTLETON, David (Org.). **Commercial Data Mining**. Boston: Morgan Kaufmann, 2014, p. 171–179. Disponível em: <https://www.sciencedirect.com/science/article/pii/B978012416602800011X>. Acesso em: 3 maio 2023.

NEWMAN, David; CHEMUDUGUNTA, Chaitanya; SMYTH, Padhraic; *et al*. Analyzing Entities and Topics in News Articles Using Statistical Topic Models. *In*: MEHROTRA, Sharad; ZENG, Daniel D.; CHEN, Hsinchun; *et al* (Orgs.). **Intelligence and Security Informatics**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, v. 3975, p. 93–104. (Lecture Notes in Computer Science). Disponível em: <http://link.springer.com/10.1007/11760146_9>. Acesso em: 3 maio 2023.

NEWMAN, Matthew; PENNEBAKER, James; BERRY, Diane; *et al*. Lying Words: Predicting Deception from Linguistic Styles. **Personality & social psychology bulletin**, v. 29, p. 665–75, 2003.

NWANKWO, Ezinne; OKOLO, Chinasa; HABONIMANA, Cynthia. **Topic Modeling Approaches for Understanding COVID-19 MisinformationSpread in Sub-Saharan Africa**.

PENNEBAKER, J. W.; KING, L. A. Linguistic styles: language use as an individual difference. **Journal of Personality and Social Psychology**, v. 77, n. 6, p. 1296–1312, 1999.

PÉREZ-ROSAS, Verónica; KLEINBERG, Bennett; LEFEVRE, Alexandra; *et al*. Automatic Detection of Fake News. 2017. Disponível em: <http://arxiv.org/abs/1708.07104>. Acesso em: 3 maio 2023.

PODER360. **Pagamento do auxílio emergencial começa amanhã; leia o calendário**. Poder360. Disponível em: <https://www.poder360.com.br/economia/pagamento-do-auxilio-emergencial-comeca-nesta-3a-feira-veja-o-calendario/>. Acesso em: 19 jul. 2023.

PORTER, M.F. An algoritm for suffix stripping. **Program**, v. 14, p. 130–137, 2006.

PRITZKAU, Albert; BLANC, Olivier; GEIERHOS, Michaela; *et al*. NLytics at CheckThat! 2022: Hierarchical multi-class fake news detection of news articles exploiting the topic structure.

R7.COM. **Fake news de dirigente: Grandes contratações que ficaram no "quase"**. R7.com. Disponível em: <http://monitor7.r7.com/fake-news-de-dirigente-grandes-contratacoes-que-ficaram-no-quase-07072022>. Acesso em: 19 jul. 2023.

ŘEHŮŘEK, Radim; SOJKA, Petr. Software Framework for Topic Modelling with Large Corpora. *In*: [s.l.: s.n.], 2010, p. 45–50.

REIS, Julio C. S.; BENEVENUTO, Fabrício. Towards Automatic Fake News Detection in Digital Platforms: Properties, Limitations, and Applications. *In*: **Anais do Concurso de Teses e Dissertações (CTD)**. [s.l.]: SBC, 2021, p. 31–36. Disponível em: <https://sol.sbc.org.br/index.php/ctd/article/view/15754>. Acesso em: 3 maio 2023.

ROCHA, Yasmim Mendes; DE MOURA, Gabriel Acácio; DESIDÉRIO, Gabriel Alves; *et al*. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. **Journal of Public Health**, v. 31, n. 7, p. 1007–1016, 2023.

RÖDER, Michael; BOTH, Andreas; HINNEBURG, Alexander. Exploring the Space of Topic Coherence Measures. *In*: **Proceedings of the Eighth ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2015, p. 399–408. (WSDM '15). Disponível em: <https://doi.org/10.1145/2684822.2685324>. Acesso em: 3 maio 2023.

SCHMIDT, Helmut. Probabilistic part-of-speech tagging using decision trees. *In*: [s.l.: s.n.], 1994. Disponível em: <https://www.semanticscholar.org/paper/Probabilistic-part-of-speech-tagging-using-decision-Schmidt/bd0bab6fc8cd43c0ce170ad2f4cb34181b31277d>. Acesso em: 3 maio 2023.

SHAHEEN, Salma. Fake News, Escalation, and Polarization: Pakistan's Disinformation Vulnerabilities. 2022.

SHEN, Cuihua; KASRA, Mona; PAN, Wenjing; *et al*. Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. **New Media & Society**, v. 21, n. 2, p. 438–463, 2019.

SHIN, Inyoung; WANG, Luxuan; LU, Yi-Ta. Twitter and Endorsed (Fake) News: The Influence of Endorsement by Strong Ties, Celebrities, and a User Majority on Credibility of Fake News During the COVID-19 Pandemic. **International Journal of Communication**, v. 16, n. 0, p. 23, 2022.

STEPHANIE. **Latent Semantic Analysis: Simple Definition, Method**. Statistics How To. Disponível em: <https://www.statisticshowto.com/latent-semantic-analysis/>. Acesso em: 3 maio 2023.

VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The spread of true and false news online. **Science**, v. 359, n. 6380, p. 1146–1151, 2018.

WALLACH, Hanna M; MIMNO, David; MCCALLUM, Andrew. Rethinking LDA: Why Priors Matter.

WAWERU MUIGAI, Jane Wambui. Understanding Fake News. **International Journal of Scientific and Research Publications (IJSRP)**, v. 9, n. 1, p. p8505, 2019.

ZHAO, Bo. Web Scraping. *In*: [s.l.: s.n.], 2017, p. 1–3.

ZIPITRIA, Iraide; ARRUARTE, Ana; ELORRIAGA, Jon Ander. Observing Lemmatization Effect in LSA Coherence and Comprehension Grading of Learner Summaries. *In*: IKEDA, Mitsuru; ASHLEY, Kevin D.; CHAN, Tak-Wai (Orgs.). **Intelligent Tutoring Systems**. Berlin, Heidelberg: Springer, 2006, p. 595–603. (Lecture Notes in Computer Science).

2014 FIFA World Cup Brazil™. Disponível em: <https://www.fifa.com/tournaments/mens/worldcup/origin1904-p.cxm.fifa.com/tournaments/mens/worldcup/2014brazil>. Acesso em: 21 jul. 2023.

Avanço do Zika e da microcefalia assustam o mundo em 2016. Agência Brasil. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2016-12/avanco-do-zika-e-da-microcefalia-assustam-o-mundo-em-2016>. Acesso em: 3 maio 2023.

beautifulsoup4: Screen-scraping library. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/>. Acesso em: 16 jul. 2023.

Boatos.org. Boatos.org. Disponível em: <https://www.boatos.org/>. Acesso em: 3 maio 2023.

Cenário de revolta e insatisfação social ajudou a compor junho de 2013 | Agência Brasil. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2023-06/junho-de-2013-entenda-o-cenario-de-insatisfacao-que-levou-a-protestos>. Acesso em: 30 jun. 2023.

Coronavirus. Disponível em: <https://www.who.int/health-topics/coronavirus>. Acesso em: 3 maio 2023.

Digital 2022: Global Overview Report. DataReportal – Global Digital Insights. Disponível em: <https://datareportal.com/reports/digital-2022-global-overview-report>. Acesso em: 3 maio 2023.

Eleições 2018. Justiça Eleitoral. Disponível em: <https://www.tse.jus.br/eleicoes/eleicoes-2018>. Acesso em: 21 jul. 2023.

Entenda o caso — Caso Lava Jato. Disponível em: <https://www.mpf.mp.br/grandes-casos/lava-jato/entenda-o-caso>. Acesso em: 3 maio 2023.

Estudo aponta manipulação política pela internet em 70 países. Agência Brasil. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2019-09/estudo-aponta-manipulacao-politica-pela-internet-em-70-paises-em-2019>. Acesso em: 19 jul. 2023.

Fakepedia-Corpus/dataset/fakepedia-corpus-v1.csv at main · andersoncordeiro/Fakepedia-Corpus. GitHub. Disponível em: <https://github.com/andersoncordeiro/Fakepedia-Corpus/blob/main/dataset/fakepedia-corpus-v1.csv>. Acesso em: 19 jul. 2023.

Gensim: topic modelling for humans. Disponível em: <https://radimrehurek.com/gensim/>. Acesso em: 16 jul. 2023.
Golpes por celular crescem 50% em 2022, diz empresa de cibersegurança. Estadão E-Investidor - As principais notícias do mercado financeiro. Disponível em: <https://einvestidor.estadao.com.br/comportamento/golpes-celular-crescem-50-psafe/>. Acesso em: 19 jul. 2023.

Informações sobre as eleições - Eleições 2014. Justiça Eleitoral. Disponível em: <https://www.tse.jus.br/eleicoes/eleicoes-anteriores/eleicoes-2014>. Acesso em: 21 jul. 2023.

Informações sobre as Eleições 2016 para prefeito, vice-prefeito e vereador. Justiça Eleitoral. Disponível em: <https://www.tse.jus.br/eleicoes/eleicoes-anteriores/eleicoes-2016>. Acesso em: 3 maio 2023.

Jogo da Baleia Azul: Até que ponto devemos nos preocupar? **BBC News Brasil**, Disponível em: <https://www.bbc.com/portuguese/internacional-39753889>. Acesso em: 19 jul. 2023.
Matplotlib — Visualization with Python. Disponível em: <https://matplotlib.org/>. Acesso em: 16 jul. 2023.

Motivations, Methodsand Metrics of Misinformation Detection: An NLP Perspective | Journal Articles | Research Output | Research. Faculty of Humanities. Disponível em: <https://www.polyu.edu.hk/fh/research/research-output/journal-articles/motivations-methodsand-metrics-of-misinformation-detection>. Acesso em: 5 jul. 2023.

NLTK :: Natural Language Toolkit. Disponível em: <https://www.nltk.org/>. Acesso em: 16 jul. 2023.

Novo golpe duplica WhatsApp e pede dinheiro para contatos. Mundo Conectado. Disponível em: <https://mundoconectado.com.br/artigos/v/17821/novo-golpe-duplica-whatsapp-e-pede-dinheiro-para-contatos>. Acesso em: 19 jul. 2023.

Número de golpes pela internet quase triplicou em 2020, aponta ISP | Rio de Janeiro | G1. Disponível em: <https://g1.globo.com/rj/rio-de-janeiro/noticia/2021/01/28/numero-de-golpes-pela-internet-quase-triplicou-em-2020-aponta-isp.ghtml>. Acesso em: 3 maio 2023.

NumPy. Disponível em: <https://numpy.org/>. Acesso em: 16 jul. 2023.

O desastre — Caso Samarco. Disponível em: <https://www.mpf.mp.br/grandes-casos/caso-samarco/o-desastre>. Acesso em: 30 jun. 2023.

pandas - Python Data Analysis Library. Disponível em: <https://pandas.pydata.org/>. Acesso em: 16 jul. 2023.

requests: Python HTTP for Humans. Disponível em: <https://requests.readthedocs.io>. Acesso em: 16 jul. 2023.

Senado autoriza intervenção na segurança pública do estado do Rio de Janeiro. Senado Federal. Disponível em: <https://www12.senado.leg.br/noticias/materias/2018/02/21/senado-autoriza-intervencao-na-seguranca-publica-do-estado-do-rio-de-janeiro>. Acesso em: 30 jun. 2023.

Sociedade de Farmacognosia alerta sobre falsos tratamentos naturais para COVID-19. Disponível em: <http://www.unifap.br/sociedade-de-farmacognosia-alerta-sobre-falsos-tratamentos-naturais-para-covid-19/>. Acesso em: 15 jul. 2023.

spaCy · Industrial-strength Natural Language Processing in Python. Disponível em: <https://spacy.io/>. Acesso em: 16 jul. 2023.

Understanding TF-IDF for Machine Learning. Capital One. Disponível em: <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>. Acesso em: 3 maio 2023.

URL Shortener - Short URLs & Custom Free Link Shortener. Bitly. Disponível em: <https://bitly.com/>. Acesso em: 16 jul. 2023.

urllib — URL handling modules. Python documentation. Disponível em: <https://docs.python.org/3/library/urllib.html>. Acesso em: 16 jul. 2023.

Vacinação contra a covid-19 começa em todo o país. Agência Brasil. Disponível em: <https://agenciabrasil.ebc.com.br/saude/noticia/2021-01/vacinacao-contra-covid-19-come%C3%A7a-em-todo-o-pais>. Acesso em: 19 jul. 2023.

What is Misinformation / Disinformation? | Purdue Libraries. Disponível em: <https://www.lib.purdue.edu/misinformation-training/training-module/what-is-misinformation>. Acesso em: 3 maio 2023.

APÊNDICES

APÊNDICE A – CÓDIGO-FONTE DO TRATAMENTO TEXTUAL DO CORPUS

O código-fonte construído no presente trabalho para o tratamento textual do *corpus* está disponível no seguinte repositório:

https://github.com/leonardoemerson/TCC-Leonardo-Emerson/blob/main/Tratamento-Textual/Tratamento-Textual.ipynb

Neste repositório, é possível encontrar todos os arquivos necessários para reproduzir o tratamento textual de dados realizado neste trabalho. O arquivo está contido no diretório Tratamento-Textual, seguindo a estrutura abaixo:

```
TCC–Leonardo-Emerson/
└── Tratamento-Textual/
    └── Tratamento-Textual.ipynb
```

APÊNDICE B – ALGORITMO PARA EXTRAÇÃO DAS DATAS DE PUBLICAÇÃO DAS NOTÍCIAS FALSAS

O algoritmo construído no presente trabalho para extração das datas de publicação das notícias está disponível no seguinte repositório:

https://github.com/leonardoemerson/TCC-Leonardo-Emerson/blob/main/Aperfeicoamento-dos-Dados/Web-Scraping-Datas.ipynb

Neste repositório, é possível encontrar o código necessário para reproduzir a extração das datas de publicação das notícias falsas do *corpus* realizada neste trabalho. O arquivo está armazenado no diretório Aperfeicoamento-dos-Dados, seguindo a estrutura abaixo:

```
TCC-Leonardo-Emerson/
└── Aperfeicoamento-dos-Dados/
    └── Web-Scraping-Datas.ipynb
```

APÊNDICE C – ALGORITMO PARA EXTRAÇÃO DAS CATEGORIAS DAS NOTÍCIAS FALSAS

O algoritmo completo construído no presente trabalho para extração das categorias das notícias está disponível no seguinte repositório:

https://github.com/leonardoemerson/TCC-Leonardo-Emerson/blob/main/Aperfeicoamento-dos-Dados/Web-Scraping-Categorias.ipynb

Neste repositório, é possível encontrar o código necessário para reproduzir a extração das categorias das notícias falsas do *corpus* realizada neste trabalho. O arquivo está armazenado no diretório Aperfeicoamento-dos-Dados, seguindo a estrutura abaixo:

```
TCC-Leonardo-Emerson/
└── Aperfeicoamento-dos-Dados/
    └──Web-Scraping-Categorias.ipynb
```

APÊNDICE D – CÓDIGO-FONTE DA ANÁLISE EXPLORATÓRIA DO CORPUS

O código-fonte completo construído no presente trabalho para análise exploratória do *corpus* está disponível no seguinte repositório:

https://github.com/leonardoemerson/TCC-Leonardo-Emerson/blob/main/Analise-Exploratoria/Analise-Exploratoria.ipynb

Neste repositório, é possível encontrar o arquivo necessário para reproduzir a análise exploratória do *corpus* realizada neste trabalho. Os arquivos estão organizados no diretório Analise-Exploratoria, seguindo a estrutura abaixo:

```
TCC-Leonardo-Emerson/
└── Analise-Exploratoria /
    └── Analise-Exploratoria.ipynb
```

APÊNDICE E – CÓDIGO-FONTE DE MODELAGEM DE TÓPICOS E RESPECTIVOS DICIONÁRIOS

O código-fonte completo construído no presente trabalho para a modelagem de tópicos do *corpus*, bem como os respectivos dicionários gerados no código estão disponíveis no seguinte repositório:

https://github.com/leonardoemerson/TCC-Leonardo-Emerson/tree/main/Dicionario-de-Topicos

Neste repositório, é possível encontrar todos os arquivos necessários para reproduzir a criação do dicionário de tópicos construído neste trabalho. Os arquivos estão organizados no diretório Dicionario-de-Topicos, seguindo a estrutura abaixo:

```
TCC-Leonardo-Emerson/
└── Dicionario-de-Topicos/
    ├── LDA.ipynb
    └── LSA.ipynb
```

APÊNDICE F – RESULTADO DO PROCESSO DE OTIMIZAÇÃO DAS MODELAGENS DE TÓPICOS UTILIZANDO A ABORDAGEM LDA VIA MÉTRICA DE COERÊNCIA

Os resultados do processo de otimização da modelagem de tópicos via LDA por meio do ajuste de hiperparâmetros estão disponíveis no seguinte repositório:

https://github.com/leonardoemerson/TCC-Leonardo-Emerson/tree/main/Otimizacao-LDA

Neste repositório, é possível encontrar os valores de coerência para todas as combinações de hiperparâmetros. Os arquivos estão organizados no diretório Otimizacao-LDA, seguindo a estrutura abaixo:

```
TCC-Leonardo-Emerson/
└── Otimizacao-LDA /
    ├── lda-resultados-otimização-2016.csv
    ├── lda-resultados-otimização-2017.csv
    ├── lda-resultados-otimização-2018.csv
    ├── lda-resultados-otimização-2019.csv
    ├── lda-resultados-otimização-2020.csv
    ├── lda-resultados-otimização-2021.csv
    ├── lda-resultados-otimização-brasil.csv
    ├── lda-resultados-otimização-ciência.csv
    ├── lda-resultados-otimização-entretenimento.csv
    ├── lda-resultados-otimização-esporte.csv
    ├── lda-resultados-otimização-geral.csv
    ├── lda-resultados-otimização-mundo.csv
    ├── lda-resultados-otimização-política.csv
    ├── lda-resultados-otimização-religião.csv
    ├── lda-resultados-otimização-saúde.csv
    ├── lda-resultados-otimização-tecnologia.csv
    └── lda-resultados-otimização-triênio.csv
```

APÊNDICE G – RESULTADO DO PROCESSO DE OTIMIZAÇÃO DAS MODELAGENS DE TÓPICOS UTILIZANDO A ABORDAGEM LSA VIA MÉTRICA DE COERÊNCIA

Os resultados do processo de otimização da modelagem de tópicos via LSA por meio do ajuste de hiperparâmetros estão disponíveis no seguinte repositório:

https://github.com/leonardoemerson/TCC-Leonardo-Emerson/tree/main/Otimizacao-LSA

Neste repositório, é possível encontrar os valores de coerência para todas as combinações de hiperparâmetros. Os arquivos estão organizados no diretório Otimizacao-LSA, seguindo a estrutura abaixo:

```
TCC-Leonardo-Emerson/
└── Otimizacao-LSA /
    ├── lsa-resultados-otimização-categórica.csv
    ├── lsa-resultados-otimização-geral.csv
    └── lsa-resultados-otimização-temporal.csv
```

APÊNDICE H – CORPUS APERFEIÇOADO

O resultado do processo de aperfeiçoamento do corpus está disponível no seguinte repositório:

https://github.com/leonardoemerson/TCC-Leonardo-Emerson/tree/main/Corpus-Aperfeicoado

Neste repositório, é possível encontrar o arquivo do corpus aperfeiçoado no formato CSV. O arquivo está armazenado no diretório Corpus-Aperfeicoado, seguindo a estrutura abaixo:

```
TCC-Leonardo-Emerson/
└── Corpus-Aperfeicoado /
    └── Corpus-Aperfeicoado.csv
```