UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

STRAUSS CUNHA CARVALHO

Um Framework de observação da mídia jornalística digital no Brasil

Monografia de Conclusão de Curso apresentada como requisito parcial para a obtenção do grau de Especialista em Ciência de Dados

Orientador: Prof. M.Sc. Luciana Regina Bencke

Porto Alegre
2024

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Carvalho, Strauss Cunha

Um Framework de observação da mídia jornalística digital no Brasil / Strauss Cunha Carvalho. – Porto Alegre: PPGC da UFRGS, 2024.

68 f.: il.

Monografia (especialização) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2024. Orientador: Luciana Regina Bencke.

1. Viés da Mídia. 2. Modelagem de Tópicos. 3. Classificação de Posicionamento. 4. BERTopic. 5. Sentence Embeddings. 6. Few-Shot Learning. 7. Web Scraping. I. Bencke, Luciana Regina . II. Título.

“A massa mantém as marcas, as marcas mantêm a mídia e a mídia controla a massa.”

— GEORGE ORWELL

RESUMO

Nota-se – sem caminhar por veredas ideológicas de qualquer âmbito – ao longo dos anos, sobretudo após o início da era da internet, a a amplitude da disseminação de informações por meio de uma parcela da mídia jornalística digital. Nesse grande volume de textos de notícias escritos por esses grupos midiáticos, seja por meio de artigos ou reportagens, comumente, observa-se um teor enviesado ou distorcido a interesse dos editores. Sem adentrar em aspectos sociológicos, tais condutas geram parcialidade nos textos produzidos que, consequentemente, informam os leitores dentro de um contexto criteriosamente predefinido. Neste contexto, este trabalho visa a investigação e a implementação de um *framework* de observação da mídia jornalística digital brasileira, a fim de identificar os seus principais temas de interesse, bem como, também, analisar dois níveis de posicionamento dos textos das notícias produzidas.

Assim, resumidamente, em sua etapa inicial, ele se baseia em coletar, por meio de *web scrapers*, texto de notícias de relevantes portais de notícias nacionais. Em seguida, armazená-las e pré-processá-las com o intuito de se treinar e aplicar modelos de aprendizado não-supervisionados para modelagem de tópicos. Posteriormente, comparam-se os tópicos obtidos entre cada mídia, evidenciando os principais temas de interesse de cada mídia jornalística, propiciando, assim, uma interpretação dos respectivos interesses editoriais. E, fornecendo interpretações acerca sazonalidade das notícias coletadas, evidenciando, ao longo do tempo, possíveis mudanças de interesse nos temas. Por fim, propiciando, à cada notícia dentro de um tópico específico, uma classificação binária acerca do respectivo posicionamento.

Palavras-chave: Viés da Mídia. Modelagem de Tópicos. Classificação de Posicionamento. BERTopic. Sentence Embeddings. Few-Shot Learning. Web Scraping.

ABSTRACT

Without walking along ideological paths of any scope – over the years, it is observed especially after the beginning of the internet era, the extensive dissemination of information through a portion of digital news media. In this large volume of news texts written by these media groups, whether through articles or reports, commonly biased or distorted content is observed in the interest of the editors. Without delving into sociological aspects, such conduct generates partiality in the texts produced, which, consequently, inform readers within a carefully defined context. In this context, this work aims to investigate and implement a framework for observing Brazilian digital journalistic media to identify its main themes of interest. Thus, in its initial stage, it is based on collecting news text from relevant national news portals through web scrapers. Then, store and pre-process them with the aim of training and applying unsupervised learning models to topic modeling. Subsequently, the topics obtained between each media are compared, highlighting the main themes of interest to each journalistic media, thus providing an interpretation of the respective editorial interests. And, providing interpretations about the seasonality of collected news, highlighting, over time, possible changes in interest in themes. Finally, providing each piece of news within a specific topic with a binary from stance classification.

Keywords: Media Bias, Topic Modeling, Stance Classification, BERTopic, Sentence Embeddings, Few-Shot Learning, Web Scraper.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BERT	Biderectional Encoder Representation from Transformers
UOL	Universo On-Line
PUW	Proportion of Unique Words
UMAP	Uniform Manifold Approximation Projection Dimension Reduction

SUMÁRIO

1 INTRODUÇÃO E MOTIVAÇÃO

Nota-se – sem caminhar por veredas ideológicas de qualquer âmbito – ao longo dos anos, sobretudo após o advento da internet, a amplitude da disseminação de informações por meio de uma parcela da mídia jornalística digital. Em uma parte desse grande volume de textos escritos por essas entidades produtoras de notícias — hegemônicas ou independentes — comumente, observa-se um teor enviesado ou distorcido a interesse dos respectivos editores:

> Most biased choices in the media arise from the preselection of right-thinking people, internalized preconceptions, and the adaptation of personnel to the constraints of ownership, organization, market, and political power. Censorship is largely self-censorship, by reporters and commentators who adjust to the realities of source and media organizational requirements, and by people at higher levels within media organizations who are chosen to implement, and have usually internalized, the constraints imposed by proprietary and other market and governmental centers of power. (HERMAN; CHOMSKY, 2010, p. 56) [1]

Deste modo, sem adentrar em aspectos sociológicos, tais condutas têm a capacidade de gerar parcialidades nos textos produzidos que, consequentemente, conduzem os leitores a uma interpretação limitada e dentro de um contexto criteriosamente predefinido. Atualmente, com um amplo uso de tecnologia, tais entidades midiáticas jornalísticas têm explorado as redes sociais. Destas redes, podemos destacar Twitter, Facebook, YouTube e, recentemente, a Threads, da empresa Meta. Elas compõem um contemporâneo vetor de disseminação de informações, porém, neste caso, em tempo real e atingindo um público heterogêneo em menor tempo.

O avanço tecnológico proporcionado pelas redes sociais facilitou a disseminação de notícias em tempo real pelas entidades de mídia, alcançando assim um público heterogêneo mais rapidamente. Isso ocorre, sobretudo, no que se refere às entidades midiáticas jornalísticas independentes, outrora, incapazes de atingir um grande público, devido ao investimento inicial necessário para operacionalização e, em alguns casos, por questões de direitos de concessão. Deste modo, levanta-se a hipótese de que o aumento da produção de notícias em tempo real e o crescente número de entidades envolvidas, pode culminar no aumento da propagação de informações enviesadas.

[1] Grande parte das escolhas tendenciosas feitas pela mídia surge da pré-seleção de pessoas influentes, de preconceitos internalizados, de modo a se adaptar às restrições organizacionais, mercado e poder político. A censura é, em grande parte, autocensura, feita por repórteres e comentaristas que ajustam às realidades dos requisitos editoriais da fonte e da mídia. Isto ocorre, sobretudo, por pessoas em níveis mais altos dentro das organizações de mídia que são escolhidas para implementar e internalizar as restrições impostas por interesses organizacionais, provenientes de mercados e de centros governamentais de poder. (HERMAN; CHOMSKY, 2010, p. 56)

Tal fenômeno, na literatura, denomina-se viés da mídia, embora sem uma definição geral. Segundo Fert (2022), o termo tem origem no século XIX, contrapondo-se ao real propósito da mídia, ou seja, produzir textos imparciais. Em seu trabalho, o autor descreve a necessidade dos textos serem escritos por um profissional, a fim de transmitir informações, de modo a refletir os fatos reais.

De acordo com Fert (2022), embora tenham sido conduzidos numerosos estudos sobre o viés da mídia, eles partem, naturalmente, em sua maioria, das ciências sociais. Isto é, trabalhos subjetivos de âmbito qualitativo. Em contrapartida, este presente trabalho visa propor uma abordagem em termos quantitativos, por meio de uma análise estatística descritiva e de modelagem de tópicos, baseada em algoritmos de inteligência artificial.

Neste cenário, este trabalho visa a investigação e a implementação de um *framework* de observação da imprensa, por meio do treinamento e utilização de modelos de aprendizado não-supervisionados para modelagem de tópicos. Assim, a partir dos tópicos gerados, evidenciar os principais temas de interesse de cada mídia jornalística. E, propiciando, assim, uma interpretação dos respectivos interesses editoriais e analisando, em dois níveis de posicionamento, os textos de notícias produzidas.

O *framework* proposto neste trabalho coleta, por meio de *web scrapers*, texto de notícias oriundas de relevantes portais de notícias nacionais. Em seguida, realiza o armazenamento e o processamento com o intuito de se treinar e aplicar modelos de aprendizado não-supervisionados para modelagem de tópicos. Posteriormente, comparam-se os tópicos obtidos entre cada mídia, evidenciando os principais temas de interesse de cada mídia jornalística, propiciando, assim, uma interpretação dos respectivos interesses editoriais. E, fornecendo interpretações acerca da sazonalidade das notícias coletadas, evidenciando, ao longo do tempo, possíveis mudanças de interesse nos temas. Por fim, propiciando, à cada notícia dentro de um tópico específico, uma classificação binária acerca do posicionamento.

1.1 Objetivos da Pesquisa

O principal objetivo desta pesquisa consiste em investigar e implementar um *framework* visando a coleta, o processamento e o agrupamento dos textos de notícias produzidos por mídias jornalísticas digitais no Brasil, evidenciando os principais temas de interesse de das mídias jornalística, propiciando, assim, uma interpretação dos respectivos interesses editoriais.

E, como objetivos secundários: a) capturar informações sazonais nas notícias coletadas, a fim de detectar, ao longo do tempo, possíveis mudanças de interesse nos temas obtidos; b) classificar as notícias por meio de dois níveis de posicionamento, visando, dentro de um tema, tratar a subjetividade associada.

1.2 Escopo da Pesquisa

Neste trabalho, sugere-se um *framework* para tratar os problema de se identificar os principais temas de interesse das mídias jornalísticas digitais brasileira. Deste modo, propondo uma abordagem de cunho prático, quantitativo, baseado em dados históricos de notícias veiculadas na internet por meio de suas páginas html.

Não faz parte do escopo deste trabalho: a) propor ou investigar as notícias por meio de abordagens de cunho teórico e qualitativo de ordem sociológica; e b) a otimização exaustiva de hiperparâmetros na tarefa de treinamento dos modelos de tópicos.

1.3 Requisitos da Pesquisa

Esta pesquisa deve ser capaz de propiciar:

1) uma revisão bibliográfica acerca dos recentes avanços da modelagem de tópicos usando técnicas de agrupamento, no âmbito de textos de notícias;
2) uma investigação de trabalhos relevantes acerca do uso de abordagens quantitativas para o tratamento do problema do viés da mídia;
3) a proposição de um *framework*, a fim de se analisar os textos das notícias produzidos pela mídia jornalística digital no Brasil;
4) a aplicação do *framework* proposto, por meio de um estudo de caso, composto por:

    4.1) desenvolvimento de cinco coletores *web scraper* de notícias em páginas html, a fim de coletar um significativo e representativo conjunto de noticias; e

    4.2) pré-processamento e análise, por meio de estatística descritiva, dos textos das notícias a serem coletadas;

    4.3) treinamento a avaliação dos resultados de cinco modelos de tópicos correspondendo às mídias jornalísticas foco deste trabalho;

    4.4) análise e discussão dos principais resultados obtidos, por meio dos experimentos a serem realizados.

5) a produção de um *corpus* inédito e amplo composto por notícias de mídias jornalísticas brasileiras.

1.4 Estrutura do Texto

No Capítulo 1, introduz-se as motivações, o tema e o problema. Seguindo pelos objetivos, o escopo, os requisitos e a estrutura deste trabalho. No Capítulo 2, descreve-se a fundamentação teórica necessária ao entendimento das etapas do *framework* proposto, no âmbito da da modelagem de tópicos, com os recentes avanços no processamento de linguagem natural. O Capítulo 3 apresenta uma revisão bibliográfica, por meio de uma investigação de relevantes pesquisas relacionadas com este trabalho. Um Framework de observação da mídia jornalística digital Brasileira, principal contribuição desta pesquisa, encontra-se descrito no Capítulo 4. O Capítulo 5 apresenta a aplicação do *framework* proposto em um estudo de caso envolvendo 45.036 notícias de cinco mídias jornalísticas. Os respectivos resultados são apresentados no Capítulo 6. Por fim, no Capítulo 7, apresenta-se a conclusão da pesquisa e as suas principais contribuições, recomendações e sugestões para trabalhos futuros, bem como alguns aspectos complementares abordados nesta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Recentemente, devido a capacidade de se representar e analisar a linguagem humana computacionalmente, o Processamento de linguagem natural (PLN), tem atraído a atenção do público geral, nos mais diversos contextos do dia a dia. De acordo com Khurana Aditya Koli e Singh (2019), o PLN é um ramo da Inteligência Artificial e Linguística, capaz de propiciar aos computadores entenderem as declarações e as palavras escritas em línguas humanas.

No âmbito do PLN, a Modelagem de Tópicos tem possibilitado a descoberta de temas e suas respectivas relações contidas nos textos em linguagem natural. Segundo (CHURCHILL, 2022), um tópico pode ser definido como um padrão recorrente de co-ocorrência de palavras. Ou seja, um agrupamento de palavras relacionadas dentro de um contexto linguístico.

Neste trabalho, aborda-se a modelagem de tópicos no escopo das descoberta de padrões nas notícias veiculadas por cinco mídias jornalísticas digitais brasileiras. Nas seções a seguir, descrevem-se, resumidamente, os principais conceitos teóricos das técnicas, *frameworks* e modelos relacionados com o trabalho proposto. Na última seção deste Capítulo, descreve-se o conjunto de métricas empregadas nas etapa de avaliação dos modelos treinados.

2.1 Notícia

De acordo (TANDOC; LING, 2018) apud (MOSER et al., 2022), se define o termo notícia como sendo uma produção do jornalismo, na qual espera-se que informações independentes, precisas, abrangentes e confiáveis sejam fornecidas. Deste modo, deseja-se que a noticia seja representação correta e imparcial dos acontecimentos.

(MOSER et al., 2022), compara o termo notícia com o termo reportagem, descrevendo que há uma semelhança em ambos, entretanto, sutilmente diferenciados pelo fator tempo. Ou seja, enquanto a reportagem trata sobre fatos não necessariamente atuais, enfatizando acontecimentos que reatualizem os fatos, a notícia tem caráter imediatista, ou seja, se referindo ao atual, ao novo e agora.

2.2 BERTopic

Cerne deste trabalho, a biblioteca *Topic Modeling from BERT* (BERTopic), proposta por (GROOTENDORST, 2022), permite efetuar modelagem de tópicos. BERTopic aplica uma abordagem de *clustering* para identificação dos tópicos, utilizando representações do textos extraídas dos grandes modelos de linguagem. A biblioteca utiliza modelos de linguagem baseados na arquitetura *Transformers* e usa a técnica *Class-Based Term Frequency–Inverse Document Frequency* (c-TF-IDF) para selecionar um conjunto de palavras representativo de cada tópico (*cluster*).

2.3 BERT

O *Bidirecional Encoder Representations from Transformers* (BERT) (DEVLIN et al., 2019) é um modelo de redes neurais artificiais pré-treinado para a representação de linguagem. Ele foi projetado para lidar, em todas as camadas da rede, com representações bidirecionais profundas de texto, sendo capaz de apreender o contexto de uma palavra com base em todas as suas palavras vizinhas à esquerda e à direita do trecho de texto em questão.

O BERT, um marco importante no estado da arte, foi treinado em dois *corpus*, o *Book Corpus* e as palavras da Wikipedia, ambos compondo um conjunto de dados com mais de 33 milhões de tokens. Assim, desde a sua criação, o BERT tem inspirado uma grande família de modelos de linguagem, tal como o BERTopic, foco principal deste trabalho.

2.4 Sentence Transformer from BERT

O *Sentence Transformer from BERT* (SBERT) (REIMERS; GUREVYCH, 2020), é um *framework* de aprendizado de máquina de código-livre, desenvolvido em PyTorch, na qual se utiliza de modelos baseados na arquitetura *Transformers* [1]. SBERT oferece uma vasta coleção de modelos pré-treinados e ajustados, em mais de 100 idiomas, para diversas tarefas, dentre as quais: similaridade textual semântica, pesquisa semântica e

[1] *Transformers* é uma arquitetura de aprendizado profundo, proposta em 2017, que utiliza o mecanismo de atenção, ou seja, se baseia na quantificação da influência de diferentes partes dos textos de entrada.

mineração de paráfrases. Para isso, é possível extrair *embeddings* dos modelos, que são representações do texto, e efetuar comparações de sentenças por meio da similaridade do cosseno das respectivas embeddings.

Para tarefas de verificação de similaridade, os modelos SBERT são computacionalmente mais eficientes do que utilizar os modelos originais BERT, pois utilizam, no seu treinamento, uma arquitetura de redes neurais siamesas [2]. Os modelos SBERT permitem que vetores de tamanho fixo representem sentenças de entrada comparando-as por meio de medidas de similaridade, tal como por cosseno, distância Euclidiana ou distância de Manhattan.

2.5 Sentence Embeddings

As *Sentence Embeddings*, de acordo com (REIMERS; GUREVYCH, 2019), são representações vetorial esparsas e de tamanho fixo capazes de armazenar informações acerca do contexto e significado dos tokens. Elas são capazes de representar sentenças inteiras, diferentemente representação individual de palavras das *Word Embeddings*.

2.6 Count Vectorizer

Count Vectorizer, neste trabalho, se refere a uma tradicional técnica utilizada há décadas no âmbito do PLN que tem o objetivo de converter o texto de documentos em representações numéricas esparsas, contabilizando as ocorrências de cada token presente no *corpus*. A técnica gera uma matriz na qual as n linhas representam os documentos e as m colunas representam os tokens de todo o *corpus*, sem repetição. Deste modo, cada célula da matriz define a frequência do token no respectivo documento.

Neste trabalho, o uso da técnica de *Count Vectorizer* se dá durante o processo de seleção de palavras para representação dos tópicos (*clusters*). Para isso, a utilização desta técnica permite: a) definir os tamanho dos N-gramas que serão utilizados na representação; e b) remover *stop words*, previamente definidas; e c) tratar os eventuais acentos presentes nos tokens.

[2] De acordo com (SOUZA G.; NAZARé, 2019) *apud* (BROMLEY J.; SHAH, 1994), a Rede Neural Siamesa - RNS consiste em duas redes que compartilham pesos idênticos e ligados por uma ou mais camadas de entrada, resultando em dois vetores de características que são comparados por meio de medidas de similaridade entre as diferentes instâncias. No geral, uma RNS é aplicada à tarefa de classificações e verificações de autenticidade.

2.7 HDBSCAN

A tarefa de Modelagem de tópicos implica no uso de técnicas de agrupamento de dados. Isto é, por meio de uma função de distância pré-definida, visa incluir elementos com graus de semelhança próximos nos mesmos conjuntos, minimizando a distância intragrupo e maximizando a distância intergrupo.

(CAMPELLO; MOULAVI; SANDER, 2013) propuseram um algoritmo baseado em densidade para descoberta de grupos em grandes banco de dados espaciais com ruídos, denominado em inglês, *Density-Based Clustering Based on Hierarchical Density Estimates* - (HDBSCAN). Utilizado neste trabalho, ele cria os *clusters* baseando-se em cada concentração de elementos em um espaço amostral. Assim, no processo de formação de *clusters*, os elementos pertencences às regiões de baixa densidade, são interpretados como *outliers*.

2.8 UMAP

Uma redução de dimensionalidade visa, no âmbito computacional, comprimir o conjunto original de dados, ocupando menos espaço em memória e armazenamento, aumentando a velocidade de processamento e cálculos e removendo componentes redundantes ou irrelevantes.

Neste trabalho, para a redução da dimensionalidade, utilizou-se a *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) (MCINNES; HEALY; MELVILLE, 2020). De acordo com os autores, UMAP é uma técnica denominada aproximação e projeção de coletor uniforme para redução de dimensão, que tem como propósito principal a redução geral da dimensão não linear. Os autores destacam que a técnica propicia uma maior preservação de características estruturais globais do conjunto de dados.

2.9 Few-Shot Learning

De acordo com (TUNSTALL et al., 2022), os métodos de rotulagem para treinamento de dados não anotados surgiram como uma solução para cenários de escassez de rótulos, pois a rotulagem tende a ser demorada e cara. No caso do método de rotulagem

denominado *few-shot*, ele utiliza um classificador, geralmente baseado em uma adaptação de modelos de linguagem, a fim de ser treinado em um pequeno número de exemplos previamente rotulados. Posteriormente, aplica-se o clasificador treinado a todo conjunto de dados sem rótulo.

2.10 c-TF-IDF

De acordo com (JURAFSKY; MARTIN, 2021), Frequência do Termo, em inglês, *Term Frequency* (TF) consiste na frequência de uma palavra p em um documento d, que denota-se por $tf_{p,d}$.

A inversa da frequência do termo, em inglês, *Inverse Document Frequency* (IDF), como o próprio nome sugere é utilizada para calcular o peso de palavras em todos os documentos do corpus. Assim, palavras com frequências muito baixas (raras) têm um alto escore IDF. Por outro lado, palavras com altas frequências, têm baixo score IDF, sobretudo *stop words* (palavras vazias, palavras das classes gramaticais, conjunções e artigos). A equação da IDF da palavra p é apresentada na Equação 2.1, onde N é o número total de documentos na coleção e df_p é o número de documentos nos quais a palavra p ocorre.

$$idf_p = log_{10}(\frac{N}{df_p}) \tag{2.1}$$

Finalmente, o produto da TF com a IDF é o score TF-IDF, descrito na Equação em 2.2, onde $w_{p,d}$ que corresponde ao score da palavra p no documento d. Assim, produzindo, para cada documento, um vetor de pesos de cada palavra do vocabulário com relação ao documento em questão.

$$w_{p,d} = tf_{p,d} \cdot log_{10}(\frac{N}{df_p}) \tag{2.2}$$

O BERTopic, em seus hiperparâmetros, permite a utilização uma variação do TF-IDF, denominado Class-Based Term Frequency–Inverse Document Frequency - c-TF-IDF[3]. Segundo (GROOTENDORST, 2022), tal variação permite o ajuste para funcionar em nível de *cluster* (tópico), a invés de nível de documento, diferenciando, para cada cluster, os respectivos critérios de formação. Assim, propiciando uma representação mais precisa dos tópicos dentro da matriz de palavras.

[3]<https://maartengr.github.io/BERTopic/api/ctfidf.html>

$$w_{x,c} = |tf_{x,c}| \cdot log(1 + \frac{A}{f_x}) \tag{2.3}$$

Onde:

$tf_{x,c}$, a frequência de uma palavra x em uma classe c.

f_x, a frequência de uma palavra x em todas as classes.

A, a média do total de palavras por classe.

2.11 Métricas

As subseções seguintes descrevem as métricas quantitativas empregadas na mensuração dos resultados obtidos nos modelos de tópicos treinados.

2.11.1 Similaridade por Cosseno

A similaridade por cosseno, aplicada ao PLN, apresentada na Equação 2.4, fornece uma medida (*score*) da similaridade entre duas sentenças representadas por meio de dois vetores A e B — num espaço vetorial, na qual se avalia o valor, compreendido no intervalo $[-1, 1]$ em função do cosseno do ângulo compreendido entre elas.

$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.4}$$

Assim, quando o valor da similaridade tende a 0, significa que o angulo é maior e os dois vetores são ortogonais ou perpendiculares entre si. Como exemplo um ângulo entre os dois vetores sendo de 90 graus, a similaridade do cosseno terá valor $\cos(90) = 0$. Ou seja, significando que os dois vetores são ortogonais ou perpendiculares entre si. Por outro lado, quando o valor da similaridade tender a 1, significa que o ângulo é menor e as sentenças são mais próximas (ou parecidas). Como exemplo um ângulo entre os dois vetores sendo de 12 graus, a similaridade do cosseno terá valor $\cos(12) = 0.978$

2.11.2 Índice de similaridade de Jaccard

Comumente utilizado, no século passado, para se medir a similaridade de componentes químicos, o índice de similaridade de Jaccard (JACCARD, 1901), aplicado ao

PNL, representado em 2.5, mede a similaridade entre duas sentenças.

O índice de Jaccard é obtido dividindo a interseção das duas sentenças, representadas por dois conjuntos, dividindo-se a sua interseção pela sua união. Como resultado, o índice tende a 1, caso ambos conjuntos tenham uma quantidade n de elementos em comum, ou seja, a interseção dos dois conjuntos. Por outro lado, o coeficiente tende a 0, em condição contrária.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.5}$$

2.11.3 Coerência

A métrica de coerência, no contexto da modelagem de tópicos, avalia, a partir de um *corpus* de referência, o quão aderente as palavras pertencentes a um tópico gerado têm significados em comum e se inserem em um mesmo contexto interpretativo. Deste modo, no geral, uma coerência se caracteriza pelo relacionamento semântico próximo das n palavras principais que representam um determinado tópico.

Observa-se, no parágrafo anterior, o quão qualitativo se dá a tarefa de medir coerência de palavras em um determinado tópico, sobretudo pelo viés interpretativo inerente a humanos. No entanto, uma abordagem quantitativa é proposta por (HINNEBURG; RöDER, 2015). Em seu trabalho, os autores propuseram um abordagem por meio de quatro dimensões, conforme listagem a seguir, cada qual com n específicas parametrizações, elevando, assim, a um expressivo número de combinações possíveis.

a) segmentação de subconjuntos de palavras;

b) estimativa de probabilidade;

c) medida de confirmação; e

d) agregação.

Ainda no trabalho de (HINNEBURG; RöDER, 2015), baseado na quatro dimensões citadas, os autores sugerem o uso de um modelo de coerência denominado c_v, devido aos seus resultados se assemelharem à interpretação humana. Desde então, tal modelo de coerência tem sido observado na literatura e vem sendo o modelo *default* para o cálculo de coerência na biblioteca Gensim (REHUREK; SOJKA, 2011).

Resumidamente, no modelo c_v, utilizado neste trabalho, para cálculo de coerência, cada palavra do tópico é comparada com o conjunto de palavras de todos os tópicos.

Posteriormente, uma janela deslizante booleana é utilizada, a fim de se avaliar se duas palavras co-ocorrem.

Assim, para todas as n palavras mais prováveis por tópico, cria-se um vetor de palavras de tamanho n, em que cada célula contém a informação mútua pontual normalizada, em inglês *Normalized Pointwise Mutual Information* - NPMI [4], entre essa palavra todas as $n - i$ palavras seguintes. Então, todos os vetores de palavras em um tópico são agregados em um grande vetor de tópicos.

Por fim, obtém-se o score c_v, por meio da média de todas as similaridade por cosseno entre cada palavra do tópico e seu vetor de tópico.

2.11.4 Proporção de Palavras Únicas

A Proporção de Palavras Únicas, em inglês *Proportion of Unique Word* (PUW) é uma métrica de diversidade baseada em na interseção do conjunto de palavras únicas por tópico com conjunto de todas palavras. O resultado é um coeficiente dado pela razão entre o somatório do total de palavras únicas pelo total de palavras mais relevantes dentro do conjunto de dados.

[4] medida de associação entre duas palavras, normalizando o resultado no intervalo $[-1, 1]$. Assim, O limite inferior -1 significa nenhuma co-ocorrência, 0 significa independência e 1 significa co-ocorrência completa.

3 TRABALHOS RELACIONADOS

De modo geral, de acordo com Bernhardt, Krasa e Polborn (2008), o viés da mídia causa um impacto inegável sobre indivíduos e sociedades, pois a perspectiva do leitor é moldada e a opinião do veículo se torna a deles. Assim, causando más interpretações, culminando em resultados perigosos à sociedade.

De modo específico, insere-se o viés da mídia nos mais diversos contextos, tendo mais abrangência no âmbito político-econômico e político-ideológico. Este último, segundo Barnidge et al. (2019), o principal causador de polaridades dicotômicas ideológicas. E esse enviesamento também se insere em outras searas, como por exemplo no âmbito da saúde pública, por meio da análise das notícias das pandemias de SARS, em 2003, e COVID-19, em 2020.

Nesse sentido, (YIGIT-SERT I. S. ALTINGOVDE, 2016) e (R, 2006) *apud* (AIRES, 2020), definem o viés como uma perspectiva parcial dos fatos, descrevendo uma taxonomia para esse viés, na qual representam preferências, sendo as três principais:

1) Viés de seleção - *gatekeeping*, a preferência em selecionar, ou seja, pautar, certas histórias ou acontecimentos a serem cobertos.

2) Viés de cobertura, a preferência em priorizar tempo de exposição e divulgação a um determinado acontecimento ou pessoa.

3) Viés de declaração ou de posição política, a preferência por expressar, favoravelmente ou desfavoravelmente, comentários ou declarações de uma ideologia.

Neste sentido, a seguir, apresentam-se três trabalhos acadêmicos correlatos acerca do viés da mídia e um relatório anual, produzido por uma agência de notícias, de monitoração dos textos produzidos pelas mídias jornalísticas em todo planeta. Os dois primeiros trabalhos acadêmicos se inserem no contexto político-ideológico, o terceiro no contexto da saúde pública.

Melo (2020) monitorou e comparou, em três eleições presidenciais nacionais consecutivas, quatro entidades midiáticas populares brasileiras, por meio de notícias de dois jornais hegemônicos e duas revistas representantes das ideologias alinhadas à esquerda e à direita. No trabalho, baseando-se nas eleições dos anos de 2010, 2014 e 2018, o autor relata o viés midiático e sua capacidade de afetar deliberadamente os juízos e decisões dos leitores. Deste modo, relatando que a forma com que as notícias eram veiculadas, induzia, aos leitores a uma interpretação pré-definida de perspectiva ideológica. Assim, levando-os a crer que uma ideologia se pareça superior a outra.

Greene (2019) investigou o uso indiscriminado, pelas entidades midiáticas, dos algoritmos das redes sociais na mídia estadunidense e sua relação com a polarização política. Segundo o autor, algoritmos da mídia têm, inconscientemente, moldado as crenças de indivíduos dentro dos Estados Unidos. Em sua tese de doutorado, umas das hipóteses partiu da indagação se as notícias tendenciosas das reportagem teriam contribuído para o aumento da polarização dos Estados Unidos. Como conclusão ele aceitou a hipótese nula. Por outro lado, o autor não afirmou que as redes sociais, em suas essências, tiveram responsabilidades diretas na polarização americana.

No contexto da saúde pública, Xu Ziling Luo e Wang (2022) analisaram notícias referentes às epidemias de SARS, em 2003 e Covid-19, em 2020. Os autores denominaram como infodemia, o viés da mídia no contexto das pandemias, devido a severidade com que notícias enviesadas são potencialmente danosas nesse âmbito. Nas duas ocasiões de pandemia, o medo, as especulações e os rumores foram rapidamente amplificados e espalhados pelo mundo pela moderna tecnologia da informação.

Como conclusão, os autores observaram uma fraca correlação entre a confiabilidade das mídias envolvidas com as suas respectivas imparcialidades. Ou seja, mídias reconhecidas pelo senso comum como confiáveis disseminaram notícias parciais. Como solução e trabalho futuros, os pesquisadores sugeriram, no âmbito educacional, um método em programas de alfabetização midiática.

Além dos contextos abordados até aqui, observa-se uma regionalização dos vieses da mídia, ou seja, tendências de interesses regionais. Melo (2020) destaca fatores geográficos nas notícias disseminadas. Ou seja, o que se refere às notícias regionais, no geral, têm mais atenção da mídia local.

Em outro âmbito, no que se refere à confiabilidade das notícias disseminadas pela mídia, segundo Newman et al. (2021), em seu relatório atual publicado pela *Reuters Institute*, compilando dados de diversos países, apenas 44 % dos os entrevistados confiam na maioria das notícias na maior parte do tempo. Evidenciando, portanto, a desconfiança nas fontes e notícias, ao contrário de outrora, via mídia impressa.

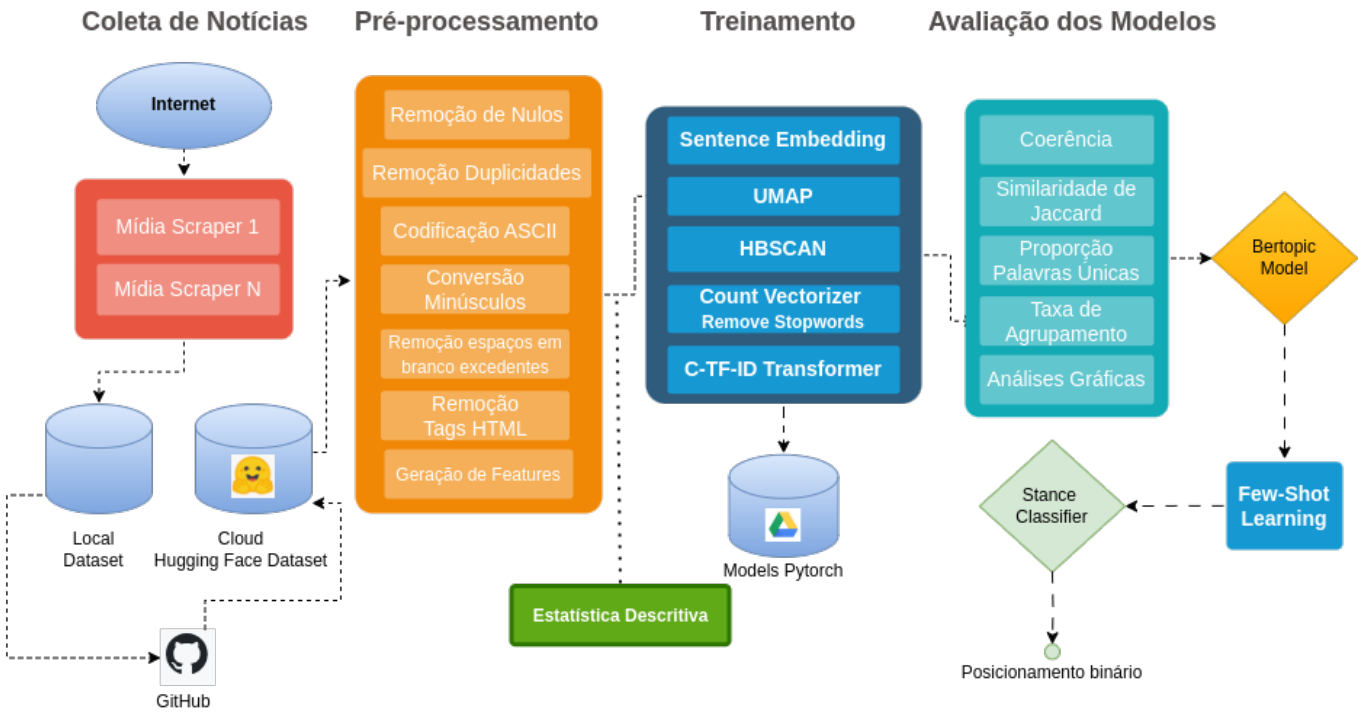
Deste modo, por meio dos três trabalhos relatados anteriormente, observa-se a capacidade de disseminação do viés da mídia em variados contextos. Assim, baseando-se nas conclusões dos trabalhos, constata-se um potencial para a geração de desconfianças no teor das notícias veiculadas. Isto é, possivelmente, culminando em um problema social, ora negando a verdade, ora afirmando a mentira.

4 O FRAMEWORK PROPOSTO

O *framework* de observação da mídia jornalística digital no Brasil é constituído por cinco etapas sequenciais, cada qual, composta por suas respectivas tarefas. A sequência das etapas constituem um *pipeline*, iniciando por meio da raspagem de dados dos portais de notícias, armazenamento, pré-processamento, análise estatística, treinamentos e avaliações dos modelos não-supervisionados, e, por fim, a interpretação e operacionalização dos modelos treinados.

Na Figura 4.1, apresentam-se as etapas e as respectivas tarefas do *framework*, desde a coleta dos textos de notícias à interpretação dos resultados dos modelos de tópicos.

Figura 4.1 – Diagrama do *framework* proposto



Fonte: Autor

Nas seções a seguir, descrevem-se, baseando-se na Figura 4.1, as cinco etapas e as principais tarefas do *framework* proposto neste trabalho.

4.1 Coleta de Notícias

Atualmente, há diversos meios computacionais para a coleta legal de notícias na internet. Dentre eles, alguns meios digitais de comunicação, tais como redes sociais e por-

tais disponibilizam o acesso às notícias por meio de Interface de Programação de Aplicação (*Application Programing Interface* - API). Assim, propiciando uma rápida integração à solução de coleta de dados e minimizando esforços de desenvolvimento. Todavia, tal abordagem, em geral, possui limitação no número de acesso e, em alguns casos, envolve custos de utilização.

Um outro modo de acesso às notícias se dá por meio de raspagem de dados na internet, em inglês *web scraping*. Neste modo, acessa-se diretamente às estruturas das páginas web, escritas em *HyperText Markup Language - HTML*, nos respectivos portais de interesse. Inerentemente, essa abordagem requer mais esforços de desenvolvimento e sustentação. Em contrapartida, não requer custos para utilização e as limitações de acesso se referem ao acesso ao histórico das páginas web publicadas.

Neste cenário, visando obter um inédito e amplo *corpus* de notícias, a primeira etapa do *framework* consiste, majoritariamente, no desenvolvimento e sustentação de um software para a coleta de notícias por meio da raspagem de dados na internet. Para tal, necessita-se escrever um software que compreenda a estrutura do código HTML de uma página web de notícias — e suas alterações durante o tempo — de modo a extrair as *tags* inerentes às notícias.

Assim, se faz relevante ressaltar, conforme estudo de caso a ser descrito no Capítulo 5, a etapa de coleta de notícias por meio de raspagem de dados na internet consumiu significativo tempo de desenvolvimento e sustentação, caracterizando-se na etapa mais onerosa de todo *framework*. Por fim, nesta etapa, após a raspagem, devem-se armazenar as notícias coletadas em um formato estruturado, composto pelo seu título, texto, link, mídia de interesse e a data.

4.2 Pré-processamento

A etapa de pré-processamento tem como objetivo o tratamento dos dados de notícias coletadas e armazenadas na etapa anterior. As tarefas se referem à limpeza e transformação dos dados, visando a mitigação de ruídos no texto das notícias, bem como a criação de covariáveis *(features)*.

No que se refere à limpeza e transformação, aplicam-se as seguintes tarefas: 1) remoção de nulos; 2) remoção de duplicidades; 3) codificação em ASCII; 4) conversão para minúsculos; 5) remoção de espaços em brancos excedentes; 6) remoção de tags HTML.

No que se refere à criação de *features*, sugere-se, visando propiciar insumos às etapas de estatística descritiva e treinamento dos modelos, a criação das seguintes novas variáveis: 1) *year*; 2) *month*; 3) *day*; 4) *dow*; 5) *woy*; e 6) *length*. Onde as cinco primeiras caracterizam marcadores temporais, sendo *dow* uma sigla para *day of week* e *woy* uma sigla para *week on year*. Por último, a sexta variável se refere ao resultado do cálculo do tamanho do texto de cada notícia, informação útil.

4.3 Estatística Descritiva

A terceira etapa no *framework* deste trabalho, fundamental à análise e interpretação quantitativa dos dados, ocorre após a etapa de pré-processamento. Isto, devido às tarefas de transformação nos textos das notícias e a geração de variáveis auxiliares, descritas na seção anterior. Assim, o resultado desta etapa provê um respaldo estatístico mínimo, servindo de insumo à quarta etapa, o treinamento dos modelos de tópicos, como a ser aprensentado no Capítulo 5.

A análise estatística sugerida se refere, basicamente, às principais medidas de tendência central e medidas de dispersão de duas variáveis independentes de interesse: a) o tamanho dos textos das notícias; e b) a taxa de produção de notícias ao longo do tempo. Também, analisa-se a recorrência de palavras mais frequentes da classe gramatical substantivo, comparando-a com as demais classes.

Deste modo, esta etapa de análise descritiva, por meio de medidas estatísticas básicas, é capaz de propiciar relevantes interpretações acerca do teor e natureza dos textos das notícias das mídias envolvidas.

4.4 Treinamento dos Modelos de Tópicos

A quarta etapa, o treinamento dos modelos de tópicos, núcleo central deste trabalho proposto, se baseia no *framerwork* SBERT, que possibilita a extração de *embeddings* pré-treinadas de modelos de linguagem, e na técnica de modelagem de tópicos BERTopic. Encontram-se descritos no Capítulo 2 os conceitos teóricos dos modelos, técnicas e ferramentas inerentes à esta etapa.

Esta etapa se inicia pela tarefa de geração das *embeddings*, a partir dos textos das notícias de um mídia jornalística. Paral tal, utiliza-se um modelo multilíngue pré-

treinado do *HuggingFace*, denominado paraphrase-multilingual-mpnet-base-v2. Nesta tarefa, sugere-se, visando otimizar o tempo de processamento, uso de uma arquitetura computacional composta por *Compute Unified Device Architecture* (CUDA)[1]. Após a extração das embeddings podemos armazenar as mesmas em disco, em formato *numpy*, a fim de reutilizá-las nas tarefas seguintes. Em seguida, ocorre o ajuste de um modelo de redução de dimensionalidade, por meio da técnica de *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP), descrito na seção 2.8.

Na tarefa de agrupamento das notícias foi utilizado o algoritmo *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Cada agrupamento (*cluster*) correponde a um tópico. A representação dos tópicos, ou seja, a escolha de um conjunto de palavras ou expressões coerentes que representem cada um dos tópicos, é feita a partir da conversão dos textos de um *cluster* em representação matricial por meio da técnica de *count vectorizer*, fundamentada em 2.6 . Para a seleção das palavras mais representativas calcula-se o *Class-Based Term Frequency–Inverse Document Frequency* (c-TF-IDF), que é uma variação uma variação do TF-IDF (JURAFSKY; MARTIN, 2021), uma medida estatística de priorização. Antes desse processo, exclui-se as *stop words* no *count vectorizer*, para que as mesmas não apareçam entre os termos mais representativos.

4.5 Avaliação dos Modelos de Tópicos

Após o treinamento, métricas de avaliação de desempenho são necessárias para aferir a sua performance. Na seção 2.11, apresentam-se métricas aceitas na literatura acerca de modelagem de tópicos, usaremos a coerência, o índice de similaridade de Jaccard e a proporção de palavras únicas.

4.6 Few-Shot Learning

Após definidos e avaliados os tópicos de interesse, nesta etapa se executa um treinamento por meio do método *few-shot*, descrito em 2.9, utilizando uma abordagem denomidada SetFit, proposta e implementada por (TUNSTALL et al., 2022). Portanto, rotula-se manualmente algumas notícias previamente selecionadas, sem o uso de *prompts*, de

[1] CUDA é uma arquitetura de computação que, por meio da extensão da linguagem C e do uso de unidade de processamento gráfico (GPU), propicia o uso de computação paralela, visando a otimização do tempo de processamento

modo binário, ou seja, posicionando-se a favor ou contra o texto da notícia relacionando-o a um determinado tópico.

Baseado nos modelos de linguagem pré-treinados oriundos do *sentence transformers from BERT*, segundo os autores, o SetFit provê eficiência no processo de rotulagem manual, minimizando tempo e custo.

4.7 Classificação de Posicionamento

A classificação de posicionamento, em inglês, *stance classification*, neste trabalho, ocorre a partir do agrupamento de tópicos na etapa de modelagem. Ou seja, somente é possível treinar um classificador de posicionamento, a partir de um subconjunto de notícias pertencentes a um tópico previamente descoberto.

Assim, deve-se treinar, como descrito na etapa anterior, um modelo de classificação binária de posicionamento para cada tópico de interesse. Após treinado, o classificador é executado em um subconjunto de dados de notícias pertencentes a um determinado tópico, de modo a classificar o posicionamento como sendo a favor ou contra. No Capítulo 6, serão apresentados os resultados da classificação de posicionamento.

5 ESTUDO DE CASO

Como um estudo de caso, a fim de se aplicar, avaliar e obter os resultados do *framework* proposto, utilizou-se um conjunto de dados composto por 45.036 notícias. Provenientes de cinco grupos de mídias jornalísticas brasileiras, no período de outubro de 2020 até outubro de 2023, esse *corpus* foi extraído de suas respectivas páginas publicadas na internet.

Na Tabela 5.1, apresentam-se, obtidos publicamente, respectivamente, os nomes fantasias, as datas de fundação, os endereços oficiais das páginas web e os nomes das contas na rede social Twitter das mídia jornalística envolvidos nos experimentos destes trabalho.

Tabela 5.1 – Entidades selecionadas para coleta de dados

Nome	Nome Fantasia	Fundação	Twitter	Página Web
Globo	Globo Comunicação e Participação S/A	31/01/1986	@globo	globo.com
Jovem Pan	Radio Panamericana S/A	12/05/1966	@JovemPanNews	jovempan.com.br
Record	Rede Record de Televisão	27/09/1953	@recordtvoficial	recordtv.r7.com
UOL	Universo On Line S/A	21/03/2003	@UOL	uol.com.br/
Gazeta do Povo	Editora Gazeta do Povo S/A	03/02/1919	@gazetadopovo	gazetadopovo.com.br/

Fonte: Autor

Nas próximas seções deste Capítulo, descrevem-se as atividades do estudo de caso inerentes às respectivas etapas do *framework* proposto.

5.1 Coleta de Notícias

A primeira etapa, a mais custosa em termos de tempo de desenvolvimento, teve por objetivo extrair as notícias de páginas na internet, a fim de se utilizar um conjunto de dados primário, inédito e não sintético. Assim, culminando no desafio de se implementar, em linguagem Python 3.10 (ROSSUM; JR, 1995), cinco *scrapers*. Ou seja, um programa de raspagem de dados para a coleta de notícias na internet para cada mídia jornalística envolvida nos experimentos.

Ao final da execução dos *scrapers*, como resultado, gerou-se 36.080 arquivos de notícias coletadas em 179 diretórios, compondo 1.38 GB de dados em espaço em disco do sistema operacional local, conforme Apêndice 9.1. Assim, este conjunto de dados gerado é composto pelos atributos: 1) mídia; 2) título da notícia; 3) texto da notícias; e 4) link da notícia.

Por fim, os dados brutos armazenados localmente foram agrupados por ano e mês

e, posteriormente, transferidos para a nuvem, gerando um *corpus* privado no Hugging Face (HUGGINGFACE, 2023), conforme Apêndice 9.2, a fim de ser formatado na etapa seguinte, conforme a ser descrito na próxima seção.

5.2 Pré-processamento

A segunda etapa, constituída por tarefas de limpeza e transformação dos dados, se iniciou com a leitura do corpus privado de notícias armazenados no Hugging Face, gerado na etapa anterior. realizou-se, nos textos e títulos das notícias, as seguintes tarefas: 1) remoção de nulos; 2) remoção de duplicidades; 3) codificação em ASCII; 4) conversão para minúsculos; 5) remoção de espaços em brancos excedentes; 6) remoção de tags HTML; e 7) criação de *features*. Nesta última tarefa desta etapa, foram gerados, à cada observação, os seguintes novas features: 1) year; 2) month; 3) day; 4) dow; 5) woy; e 6) length.

Por fim, como resultado, obteve-se um *corpus* composto por 36.826 linhas e 9 atributos. Na Figura 5.1, apresenta-se, como exemplo, uma observação com todos os campos formatados.

Figura 5.1 – Exemplo de uma observação no corpus de notícias

media	year	month	day	dow	woy	title	text	length
r7	2020	10	22	3	43	com a nova fase do pronampe em discussao no congresso a expectativa e quadruplicar o alcance mas mantendo a taxa de juros em um digito	programas publicos de credito chegam a r bilhoes cesar conventi fotoarena fotoarena estadao conteudo os programas publicos de credito para micro e pequenas empresas atingiram r bilhoes em outubro e o governo planeja a ampliacao de concessoes segundo o secretario especial de produtividade emprego e competitividade do ministerio da economia carlos da costa sera lancada em breve a terceira fase do pronampe programa nacional de apoio as microempresas e empresas de pequeno porte leia tambem febraban com pronampe carteiras de credito de bancos cresce em mesesa nova fase do programa criado por causa dos impactos da pandemia de coronavirus ainda esta sendo discutida no congresso com a expectativa de quadruplicar a oferta de credito tivemos o pronampe e com pouquissimas diferencas entre eles como o teto para as menores empresas terem acesso ja no pronampe estamos negociando com empresas e congresso alem do banco central e acreditamos que vamos poder ter uma alavancagem aproximadamente vezes maior afirmou costa em live nesta quarta feira a primeira fase destinou r bilhoes aos empreendedores e a segunda r bilhoes a expectativa do governo e quadruplicar o alcance mas mantendo a taxa de juros em um digito com a possibilidade desses recursos serem tratados com creditos tributados pelos bancos com a ferramenta disponibilizada pelo banco central acreditamos que a taxa total se mantera em um digito isso sera suficiente para atender totalmente a demanda existente no mercado hoje disse veja tambem bndes ajuda mil empresas com medidas emergenciais seis em cada empresas foram pouco afetadas por pandemia demanda das empresas por credito registra queda em agosto os emprestimos atualmente tem taxa de juros anual igual a da selic mais a soma de ponto percentual ao ano com um prazo de meses de financiamento de acordo com o secretario mais de das empresas que entraram no programa nao tinham historico de credito formal costa afirma que a retomada rapida de alguns setores foi possivel gracas ao programa de manutencao de empregos que permitiu que empresas mantivessem seus trabalhadores e ao credito que nos ultimos dois meses alcancou valor significativo pesquisa feita pelo sebrae em parceria com a fundacao getulio vargas revelou que no mes de setembro houve uma melhora no acesso ao credito por parte dos pequenos negocios de acordo com o levantamento entre as micro e pequenas empresas que buscaram emprestimos tiveram o pedido aprovado pelas instituicoes financeiras esse e o melhor resultado para a serie iniciada em marco e esta pontos percentuais acima do indicador obtido na pesquisa feita na ultima semana de agosto os programas publicos de credito pronampe programa nacional de apoio as microempresas e empresas de pequeno porte destinado ao desenvolvimento e fortalecimento dos pequenos negocios com concessao de credito para o financiamento da atividade empresarial o emprestimo tem de garantia do governo o prazo para o pagamento do emprestimo e de meses com carencia de oito meses ja a taxa de juros anual maxima aplicada sobre o valor total do credito sera a da selic mais ao ano estao aptas a pedir o financiamento microempresas me com faturamento anual de ate r mil empresas de pequeno porte epp com faturamento entre r mil e r milhoes por ano empresas enquadradas em alguma das categorias acima que nao foram condenadas por condicoes de trabalho analogas a escravidao linha emergencial de credito para folha de pagamentos essa e uma linha de credito para financiamento de folha de pagamento de empregados no ambito do programa emergencial de suporte ao emprego pese regulamentado pela medida provisoria o financiamento podera ser pago em meses com carencia de meses e com juros de ao ano podem aderir ao programa empresarios que tiverem receita bruta anual entre r mil e r milhoes linha de credito da caixa e sebrae para capital de giro com garantias a pequenos negocios a caixa economica federal disponibilizara ate r bilhoes em credito para capital de giro a micro e pequenas empresas e microempreendedores individuais meis a operacao e viabilizada por meio do aporte de r milhoes do sebrae as garantias complementares serao concedidas pelo sebrae por meio do fundo de aval as micro e pequenas empresas fampe todo o credito sera assistido pelo sebrae em todas as etapas desde a liberacao ate a liquidacao bndes credito pequenas empresas o banco nacional de desenvolvimento economico bndes divulgou a expansao da linha bndes credito pequenas empresas que vai beneficiar as empresas para enfrentar as dificuldades de fluxo de caixa e importante destacar que as linhas de credito do bndes sao operadas pelos agentes financeiros credenciados empresas com faturamento ate r milhoes podem obter credito livre sem destinacao especifica de ate r milhoes por ano os recursos do bndes podem financiar ate da operacao a criterio do agente financeiro credenciado e as operacoes contratadas podem ter prazo total de ate anos incluindo um prazo de carencia de ate anos programa emergencial de acesso a credito peac instituido pela medida provisoria podera contar com ate r bilhoes de recursos da uniao podendo garantir ate r bilhoes em operacoes de credito tornando o peac a maior medida de acesso a credito lancada desde o inicio da pandemia os recursos utilizados para as garantias ate agora vieram de um aporte inicial de r bilhoes da uniao aportados pela secretaria especial de produtividade emprego e competitividade sepec do ministerio da economia atualmente agentes financeiros ja estao habilitados para oferecer emprestimos entre r mil e r milhoes cabe a esses agentes a decisao de utilizar a garantia do programa e aprovar ou nao o pedido de credito no momento em que estruturarem cada uma de suas operacoes peac maquininhas modalidade de credito garantido por vendas com maquinas de pagamento digital para meis e mpmes o financiamento e garantido por parte das vendas futuras realizadas por meio de maquininhas limitado ao valor do contrato de emprestimo sendo dispensada a exigencia de aval ou garantia real o valor do emprestimo tambem sera definido com base nas vendas com maquininhas nao podendo ultrapassar o dobro da media mensal das vendas de bens e prestacoes de servico da empresa realizadas entre marco de e fevereiro de limitado a r mil a taxa de juros cobrada pelo agente financeiro nao podera ultrapassar ao ano e o emprestimo que sera depositado na conta do empreendedor tera carencia de seis meses e prazo de meses para pagamento incluindo o tempo de carencia a vigencia do programa e ate de dezembro de alternativeheadline com a nova fase do pronampe em discussao no congresso a expectativa e quadruplicar o alcance mas mantendo a taxa de juros em um digito	1223

Fonte: Autor

5.3 Estatística descritiva do corpus

Entre as etapas de Pré-processamento e Treinamento, há a etapa de estatística descritiva, apresentada, resumidamente na seção 4.3, de suma importância ao entendi-

mento dos textos das notícias, sobretudo por se tratar de um conjunto de dados inédito. Realizou-se uma análise estatística por meio de medidas de tendência central e medidas de dispersão, apresentadas a seguir.

Nas análises a seguir, relatam-se resultados sobre todo o *corpus*, composto pelas cinco mídias. Posteriormente, na seções seguintes, os resultados são apresentados separados por mídias. Em ambos, ressaltam-se os desafios na obtenção das notícias, neste sentido, elencam-se os seguintes fatores que explicam a variância entre total de notícias coletadas por mídia: a) porte do grupo midiático; b) capacidade de produção de notícias; c) leitura das estrutura das páginas html; e d) limitações de acesso às páginas.

Inicialmente, apresenta-se, na Figura 5.2, a proporcionalidade entre as 5 mídias jornalísticas no total de 36.826 de notícias. Nela, observa-se, respectivamente, um valor elevado de notícias à mídia Globo, seguido do UOL e Record e, com valores inferiores, Gazeta e JPan.

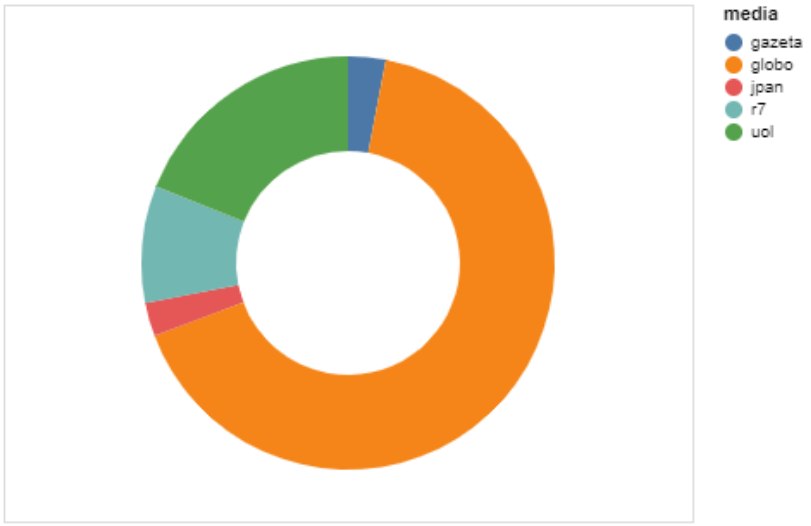
Deste modo, observou-se que a mídia Globo tem, dentro do país, um conjunto regionalizado de propagação de notícias, culminando numa alta produção de notícias diárias. Além de ter, também, segmentos específicos de notícias, como no caso do esporte. Referente às páginas html, as cinco mídias tiveram, ao longo dos 3 anos de observação, alterações nas respectivas estruturas dos códigos-fonte, implicando em desafios na raspagem de dados. Neste contexto, inerente ao conjunto de dados de difícil acesso, realizou-se uma prévia análise estatística dos dados coletados pelos *scrapers* implementados, por meio da variável tamanho do texto e da quantidade diária de notícias produzidas.

De modo a se avaliar, resumidamente, a distribuição intrínseca dos dados à cada uma das cinco mídias, calculou-se, para cada grupo de mídia, as respectivas medidas de tendência central e dispersão. Optou-se por analisar duas variáveis: a) o tamanho do texto das notícias; e b) a quantidade de notícias.

Na análise dos tamanhos dos textos de notícias produzidos, o Gráfico *boxplot* da Figura 5.3, apresenta as respectivas variabilidades das cinco mídias. Observam-se maiores textos de notícias para a Gazeta, sendo que seu valor mínimo se situa acima da mediana das 4 outras mídias. UOL, Globo e Record têm medianas aproximadas. Jpan destoou dos demais, com notícias mais curtas. No entanto, todos os cinco gráficos têm as medianas relativamente próximas às médias.

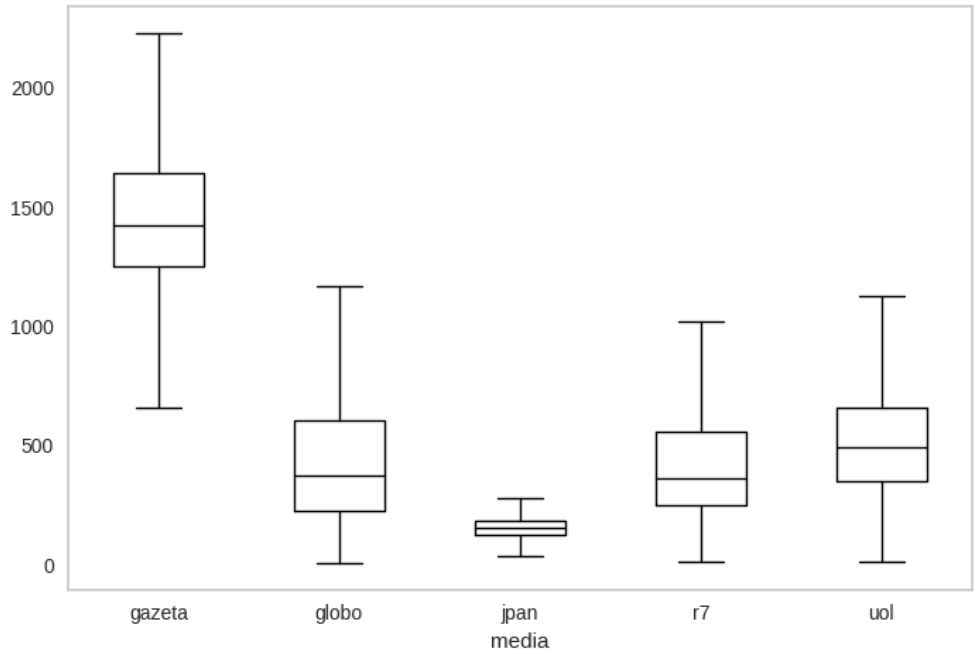
Para a análise da quantidade de notícias coletadas, a Figura 5.4 apresenta seis gráficos, ambos representando medidas de tendência central e de dispersão. O primeiro superior, centralizado, representa medidas de todo o corpus. Os demais Gráficos, posi-

Figura 5.2 – Total de notícias por mídia jornalística



Fonte: Autor

Figura 5.3 – Tamanho das notícias por mídia jornalística



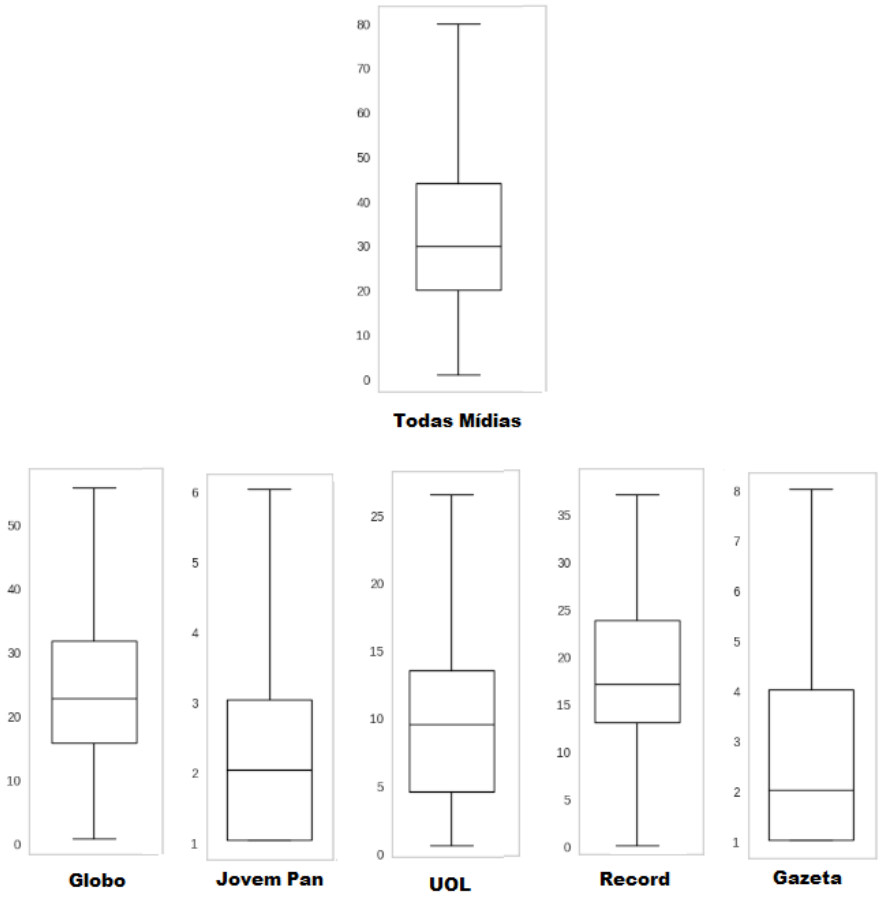
Fonte: Autor

cionados abaixo, representam as medidas de cada uma das cinco mídias. Nesta análise, buscou-se enfatizar as medidas baseadas no intervalo interquartílico, a fim de se medir a variabilidade diária da coleta de notícias sem a influência de *outliers*.

A amplitude e o intervalo interquartílico da Globo se aproximam às respectivas medidas de todo conjunto de dados. Um subconjunto do intervalo interquartílico da Record, com assimetria positiva, faz parte do subconjunto do mesmo intervalo no todo conjunto de dados, indicando proximidade no produção diária de notícias. UOL, a terceira

mídia com maior amplitude, tem assimetria levemente negativa. Gazeta e Jovem Pan têm amplitudes menores, consequentemente, com menores valores de média e medianas de produção diária de notícias.

Figura 5.4 – Variável quantidade diária de notícias coletadas



Fonte: Autor

A Tabela 5.2 apresenta, baseando-se nos *boxplots* da Figura 5.4, as principais medidas de tendência central e dispersão da variável quantidade de notícias diárias, acerca de todo o *corpus* e seus subconjuntos, representados pelas cinco mídias. Observa-se que a Globo possui maior amplitude com média e mediana, de certo modo, próximas. No entanto, com alto desvio-padrão, indicando alta variabilidade. As mesmas características ocorrem com o UOL, porém com menor amplitude. A Record possui amplitude alta, com média e mediana, de certo modo, distantes e alto desvio-padrão. Jovem Pan e Gazeta têm amplitudes baixas. No entanto, Jovem Pan tem média e media próximas e baixo desvio-padrão, indicando baixa variância.

A taxa de coleta da Tabela 5.2 que representa, para cada mídia, o quociente t entre o total de dias de notícias coletadas c sobre o total de dias no período de interesse d. Ou seja, d representado pelos 1.113 dias de coleta de notícias, compreendidos entre o mês de

outubro de 2020 até outubro de 2023 e c pelo total de dias coletados, por mídia.

Portanto, a Globo possui uma maior taxa de coleta $t = 0.87$, caracterizando maior cobertura, seguida pelo UOL, com $t = 0.61$. Record, Jovem Pan e Gazeta têm baixas taxas de coleta, causadas, supostamente, pelos desafios de raspar as respectivas páginas nas internet, como descrito anteriormente, nas seções 4.1 e 5.1.

$$t = \frac{c}{d} \tag{5.1}$$

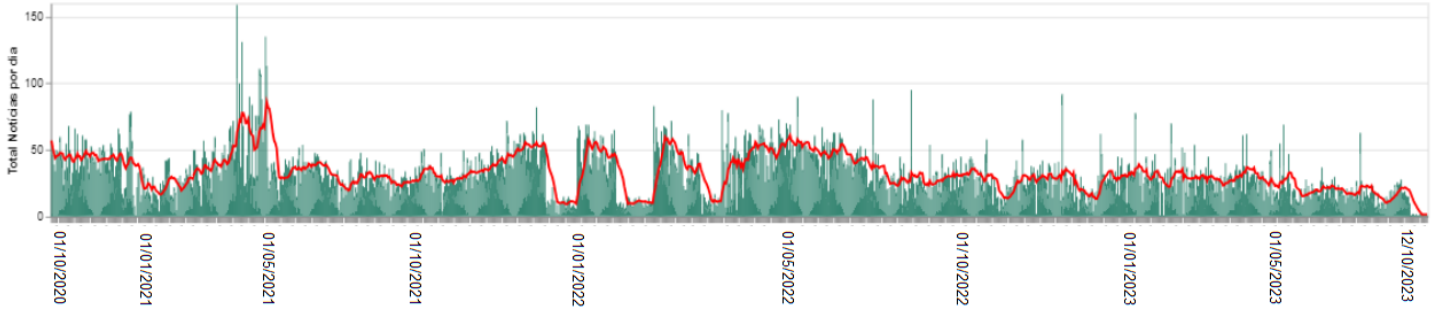
Tabela 5.2 – Quantidade diária de notícias coletadas - Medidas de Tendência Central

Dataset	**Contagem (total dias)**	**Mediana**	**Média**	**Desvio Padrão**	**Min**	**Max**	**Taxa Coleta**
Corpus	1.113	30	33.081	18.36	1	159	1.00
Globo	975	23	25.05	11.61	1	64	0.87
Jovem Pan	443	2	2.16	1.29	1	9	0.39
UOL	685	10	10.19	5.87	1	29	0.61
Record	150	18	22.47	16.33	1	102	0.13
Gazeta	175	2	6.14	10.99	1	61	0.15

Fonte: Autor

No gráfico da Figura 5.5, apresenta-se, ao longo de 3 anos, a distribuição de notícias das cinco mídias, iniciando-se no mês de outubro de 2020 até outubro de 2023. Nele, o eixo x representa o tempo, por meio do intervalo de dias; o eixo y, a respectiva quantidade diária de notícias coletadas. Observa-se, no último mês de coleta, outubro de 2023, um decaimento na quantidade de notícias devido à parada da coleta no início da segunda quinzena deste mês. Os vales se referem aos obstáculos na coleta de notícias, conforme descrito em 4.1. Ressaltam-se, nos dados temporais dos eixos x dos gráficos deste trabalho, uma linha na cor vermelha indicando a tendência média.

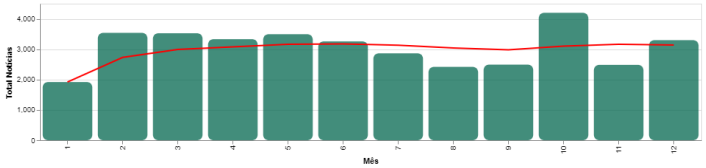
Figura 5.5 – Total diário de notícias coletadas no período de 3 anos



Fonte: Autor

No gráfico da Figura 5.6, a fim de se identificar eventos sazonais, agrupou-se, nos doze meses do ano, todo o conjunto de dados dos três anos. Nele, observa-se um acúmulo de notícias no mês de outubro, caracterizando um indicativo de aumento no número de notícias neste mês de eleições, em 2020 e em 2022.

Figura 5.6 – Total de notícias por meses dos anos

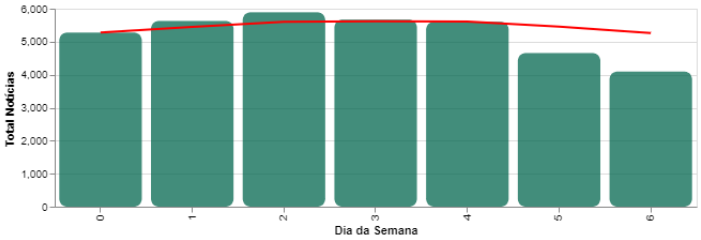


Fonte: Autor

Seguindo a mesma estratégia do agrupamento anterior, o gráfico da Figura 5.7 apresenta uma distribuição por dias da semana, iniciando por zero, representando a segunda-feira e concluindo com seis, representando o domingo. Assim, observa-se um decaimento na produção de notícias nos finais de semana.

De modo a se analisar o teor dos textos das notícias no *corpus* produzidos pelas cinco mídias jornalísticas, nas Figuras 5.9 e 5.8, apresentam-se contagens de palavras mais frequentes. No Gráfico 5.8, representado por todas as classes gramaticais do *corpus*, a partir de um vocabulário de 152.710 palavras, observam-se 3.258.034 instâncias, sendo

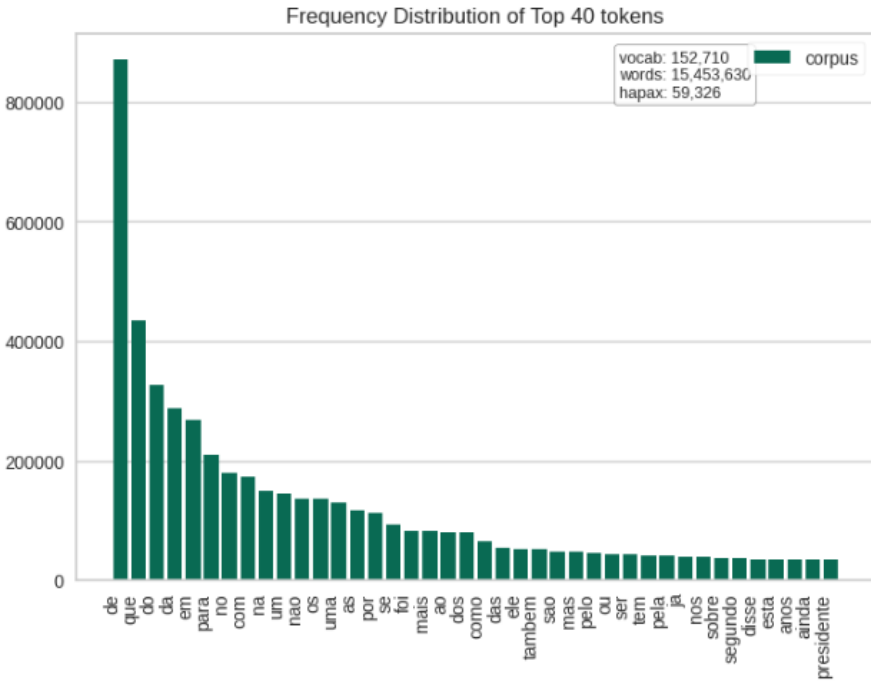
Figura 5.7 – Total de notícias por dias da semana



Fonte: Autor

59.326 com ocorrências únicas. Também, apresentam-se os primeiros quarenta elementos mais frequentes.

Figura 5.8 – Palavras mais frequentes no corpus

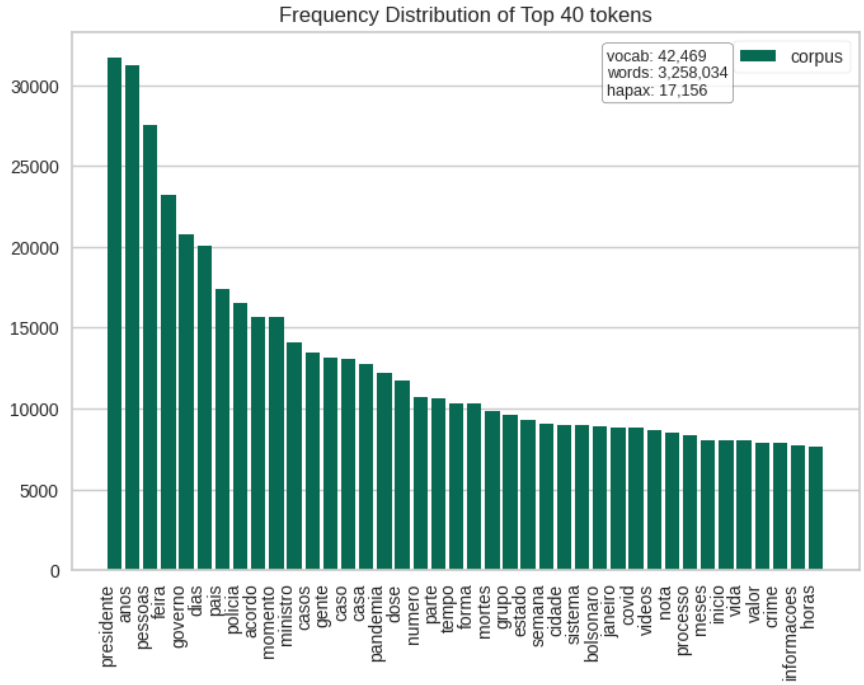


Fonte: Autor

No Gráfico 5.9, selecionando-se por classe gramatical, tem-se um vocabulário de 42.469 substantivos, composto por 15.463.630 instâncias e 17.156 ocorrências únicas. Nele, nas ordenado pelos quarenta primeiros elementos, observa-se os primeiros substantivos indicando um teor relacionado ao tema Política seguido pelo tema Pandemia — ambos fortemente relacionados com os principais tópicos gerados, a serem apresentados no Capítulo 6.

Nota-se, comparando os dois gráficos, que o segundo se aproxima mais à lei de (ZIPF, 1935). Isto é, para qualquer linguagem, a frequência de ocorrência *f(n)* de uma palavra é inversamente proporcional à sua posição na lista global de palavras depois de classificadas por sua frequência de forma descendente.

Figura 5.9 – Substantivos mais frequentes no corpus



Fonte: Autor

5.4 Estatística Descritiva das mídias Jornalísticas

De modo a especificar a análise estatística, resumidamente, nas cinco seções seguintes apresentam-se, separadas pelas respectivas mídias jornalísticas, a distribuição da produção de notícias ao logo dos 36 meses e os respectivos substantivos mais frequentes.

Nos gráficos dos totais de notícias mensais produzidas, apresentados a seguir, observa-se, em alguns casos, um decaimento na quantidade de notícias produzidas. Isto se deve, conforme descrito em 4.1, às mudanças das estruturas das respectivas páginas html.

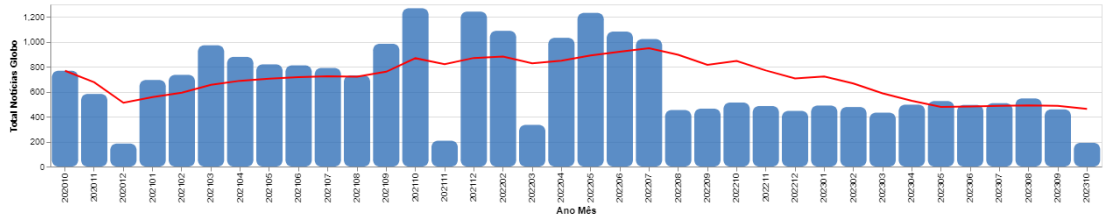
5.4.1 Globo

Dentre as mídias jornalísticas observadas, a Globo possui um maior número de notícias produzidas diariamente. Possivelmente, devido ao fato de possuir um conjunto regionalizado de propagação de notícias, culminando numa alta produção de notícias diárias. Além de possuir portais específicos por temas, como, por exemplo, esportes.

No gráfico da Figura 5.10, apresenta-se a distribuição mensal de notícias coletadas em suas respectivas páginas html, iniciando em outubro de 2020 e finalizando na primeira

quinzena do mês de outubro de 2023.

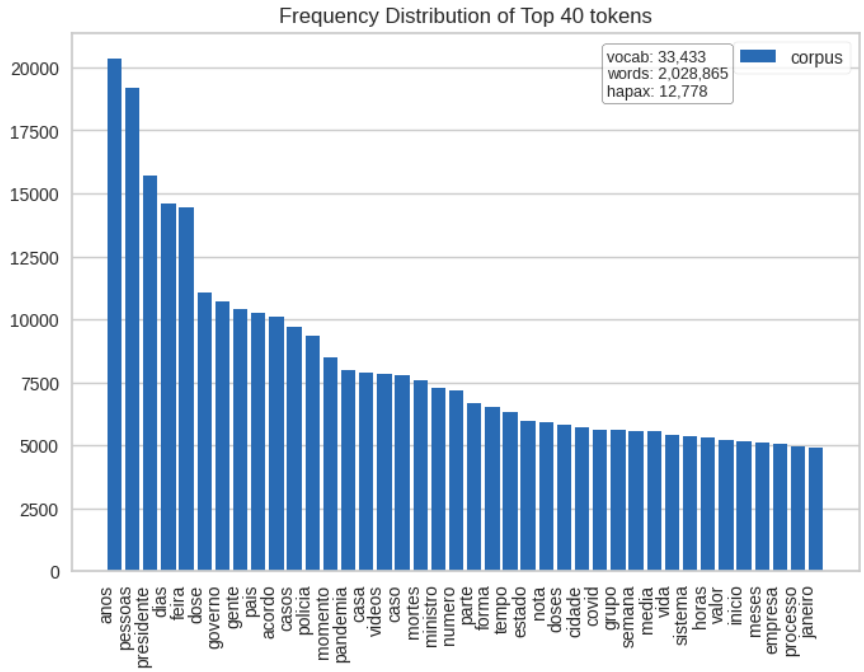
Figura 5.10 – Total de notícias mensais coletadas no período de 3 anos - mídia jornalística Globo



Fonte: Autor

O gráfico da Figura 5.11 apresenta as quarenta palavras mais frequentes, da classe gramatical de substantivos, na mídia Globo, dentro do período de 36 meses. Nele, o conjunto de substantivos é representado por um vocabulário de 33.433 palavras, composto por 2.028.865 instâncias e 12.778 ocorrências únicas. Assim como o gráfico 5.9 referente a todo o *corpus*, os substantivos têm relação com temas ligados à Política seguido pelo tema Pandemia.

Figura 5.11 – Substantivos mais frequentes no corpus - mídia jornalística Globo

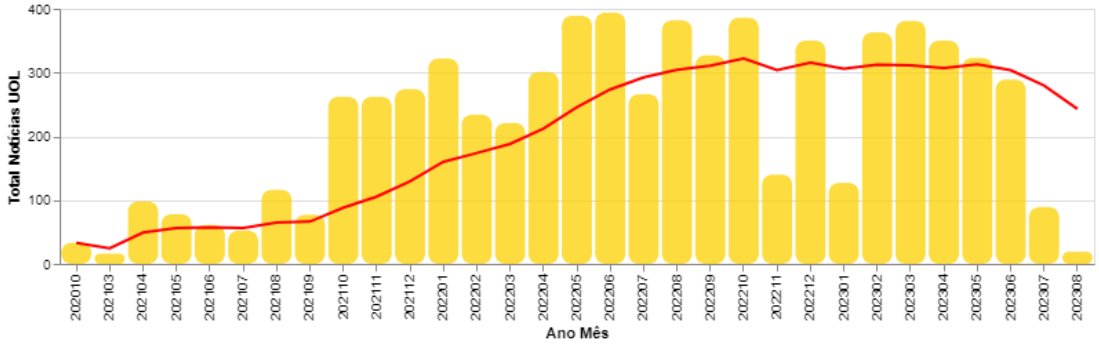


Fonte: Autor

5.4.2 UOL

No gráfico da Figura 5.12 apresenta-se a distribuição mensal, se iniciando em outubro de 2020 e finalizando na primeira quinzena do mês de outubro de 2023, de notícias coletadas nas páginas html da mídia Universo On-Line (UOL) no período de 36 meses.

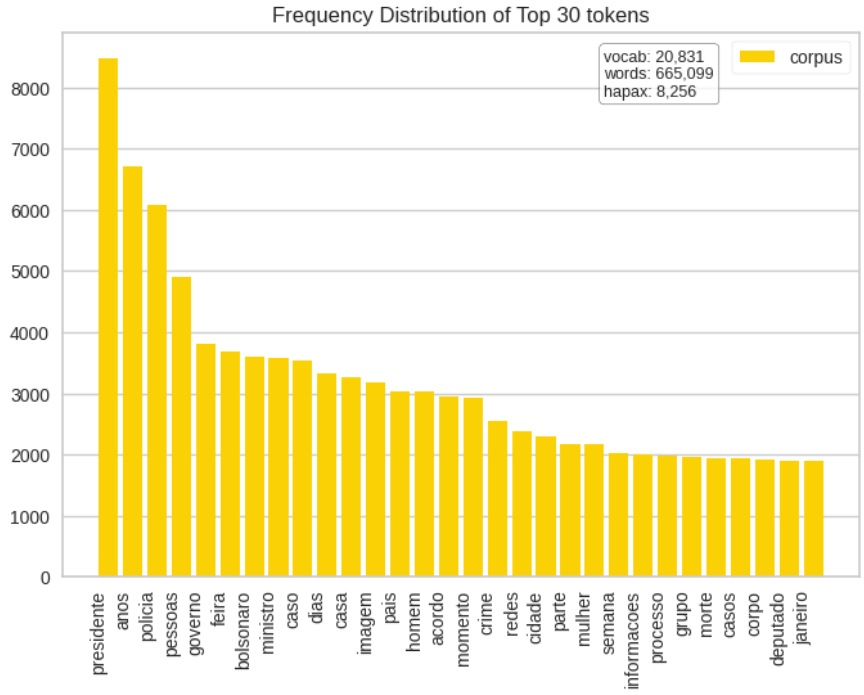
Figura 5.12 – Total de notícias mensais coletadas no período de 3 anos - mídia jornalística UOL



Fonte: Autor

O gráfico da Figura 5.13 apresenta as quarenta palavras mais frequentes, da classe gramatical de substantivos, na mídia UOL, dentro do período de 36 meses. Nele, o conjunto de substantivos é representado por um vocabulário de 20.831 palavras, composto por 665.099 instâncias e 8.256 ocorrências únicas. Grande parte dos substantivos têm relação com temas ligados à Política seguido por palavras sem liagação direta com um tema específico.

Figura 5.13 – Substantivos mais frequentes no corpus - mídia jornalística UOL

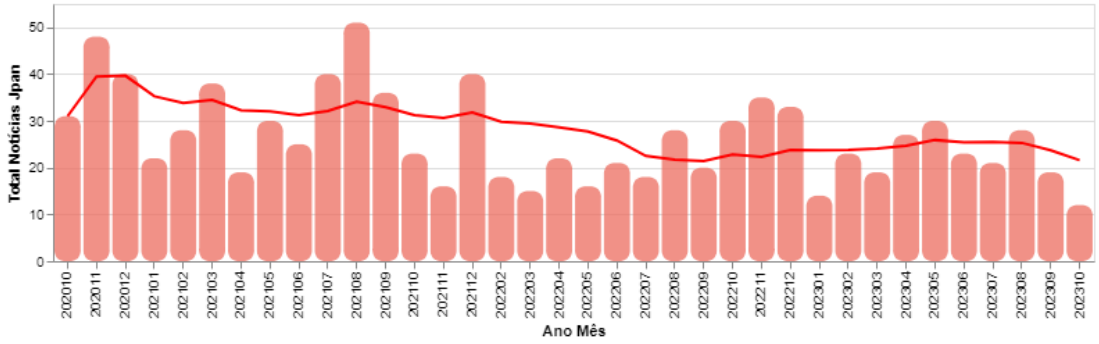


Fonte: Autor

5.4.3 Jovem Pan

No gráfico da Figura 5.14 apresenta-se a distribuição mensal, se iniciando em outubro de 2020 e finalizando na primeira quinzena do mês de outubro de 2023, de notícias coletadas nas páginas html da mídia Jovem Pan no período de 36 meses.

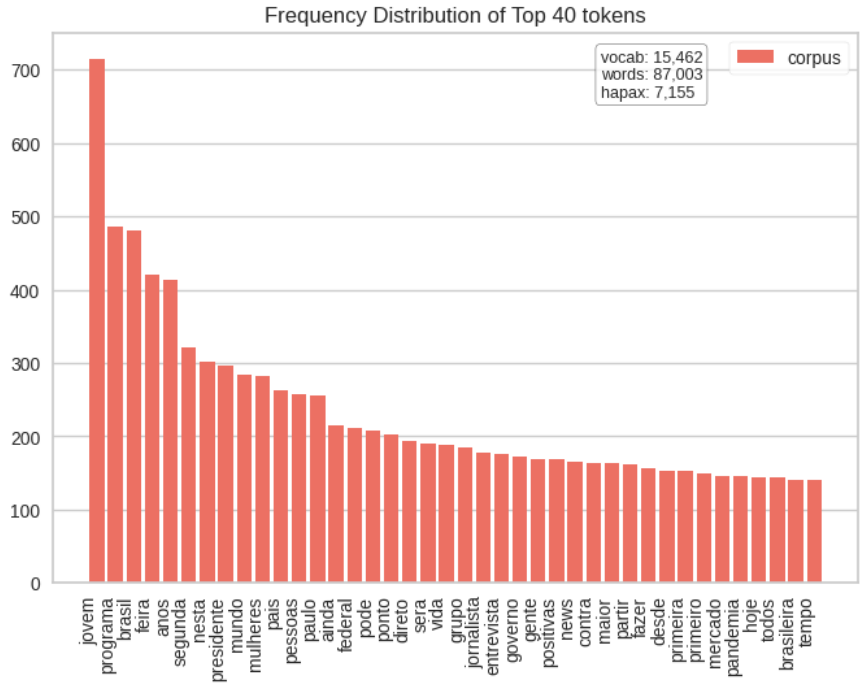
Figura 5.14 – Total de notícias mensais coletadas no período de 3 anos - mídia jornalística JPan



Fonte: Autor

O gráfico da Figura 5.15 apresenta as quarenta palavras mais frequentes, da classe gramatical de substantivos, na mídia Jovem Pan, dentro do período de 36 meses. Nele, o conjunto de substantivos é representado por um vocabulário de 15.462 palavras, composto por 87.003 instâncias e 7.155 ocorrências únicas. Majoritariamente, os substantivos presentes nas quatro primeiras dezenas têm relação com temas ligados à política.

Figura 5.15 – Substantivos mais frequentes no corpus - mídia jornalística Jpan

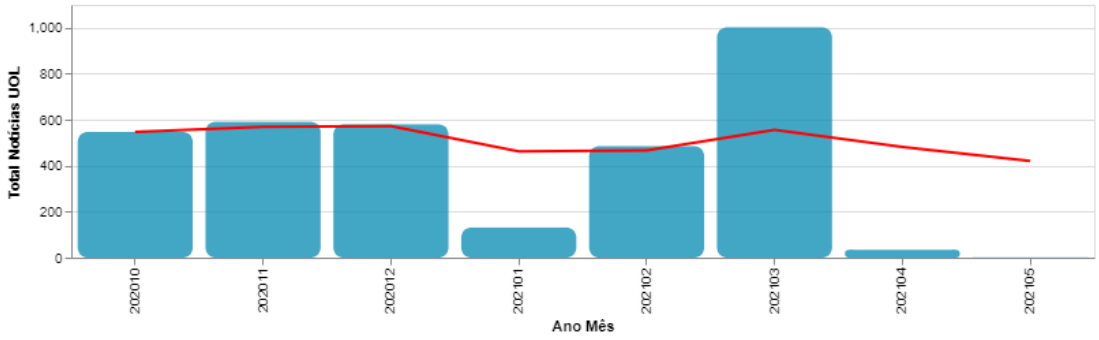


Fonte: Autor

5.4.4 Record

No gráfico da Figura 5.16 apresenta-se a distribuição mensal, se iniciando em outubro de 2020 e finalizando no primeiro semestre de 2021.

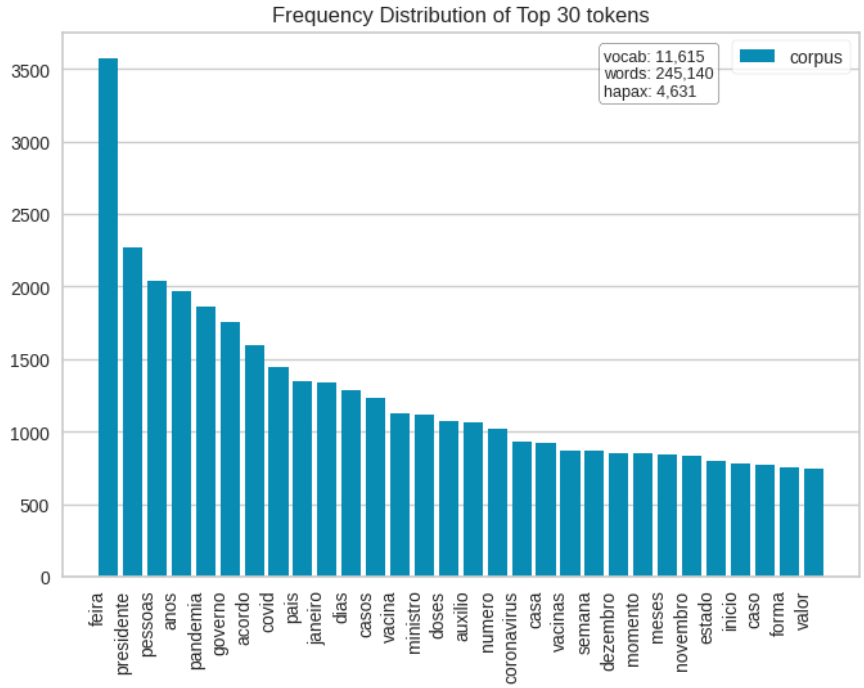
Figura 5.16 – Total de notícias mensais coletadas no período de 3 anos - mídia jornalística Record



Fonte: Autor

O gráfico da Figura 5.17 apresenta as quarenta palavras mais frequentes, da classe gramatical de substantivos, na mídia Record, dentro do período de 36 meses. Nele, o conjunto de substantivos é representado por um vocabulário de 11.615 palavras, composto por 245.140 instâncias e 4.631 ocorrências únicas. Assim como o gráfico 5.9 referente a todo o *corpus*, os substantivos têm relação com temas ligados à Política seguido pelo tema Pandemia.

Figura 5.17 – Substantivos mais frequentes no corpus - mídia jornalística Record

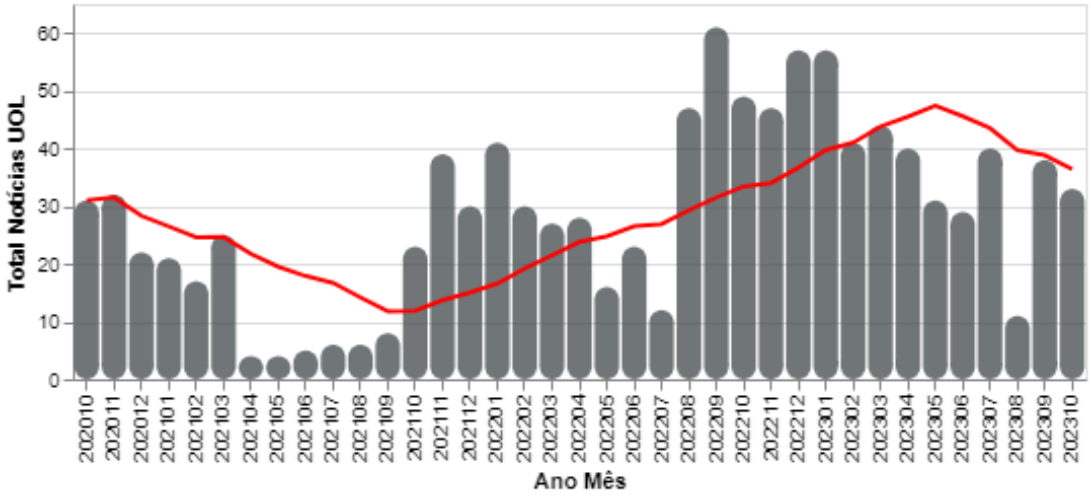


Fonte: Autor

5.4.5 Gazeta

No gráfico da Figura 5.18 apresenta-se a distribuição mensal, se iniciando em outubro de 2020 e finalizando na primeira quinzena do mês de outubro de 2023, de notícias coletadas nas páginas html da mídia Gazeta do Povo no período de 36 meses.

Figura 5.18 – Total de notícias mensais coletadas no período de 3 anos - mídia jornalística Gazeta

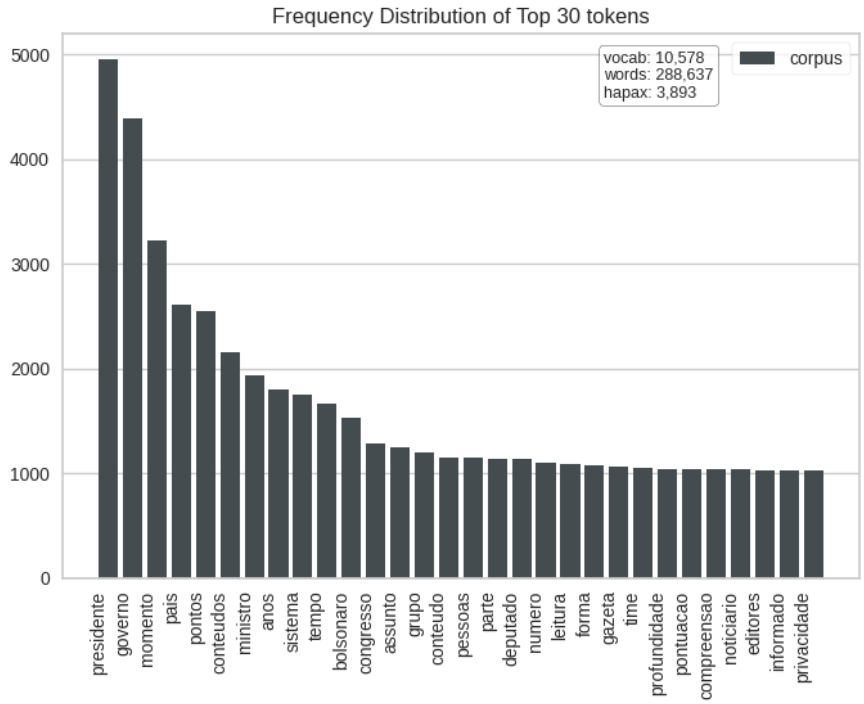


Fonte: Autor

O gráfico da Figura 5.19 apresenta as quarenta palavras mais frequentes, da classe gramatical de substantivos, na mídia Gazeta, dentro do período de 36 meses. Nele, o

conjunto de substantivos é representado por um vocabulário de 10.578 palavras, composto por 288.637 instâncias e 3.893 ocorrências únicas. Majoritariamente, os substantivos presentes nas quatro primeiras dezenas têm relação com temas ligados à política.

Figura 5.19 – Substantivos mais frequentes no corpus - mídia jornalística Gazeta



Fonte: Autor

5.5 Treinamento e Avaliação dos Modelos de Tópicos

Foram realizadas baterias de treinamento dos cinco modelos de tópicos, ou seja, um para cada mídia. Variou-se manualmente alguns parâmetros do BERTopic como os n-gramas no vetorizador, o número mínimo de clusters e de elementos no HDBSCAN e o número de vizinhos e de componentes principais no UMAP, visando obter os melhores valores de coerência dos modelos de tópicos resultantes. Todos os modelos utilizaram a clusterização por meio do algoritmo hierárquico baseado em densidade, o HDBSCAN.

Na Tabela 5.3, apresentam-se as mídias e os respectivos totais de notícias coletadas, relacionando-os com as métricas de avaliação dos modelos de tópicos. Nela, observam-se as métricas de coerência c_v, a proporção de palavras únicas e o índice de similaridade de Jaccard.

Nos resultados obtidos, o modelo treinado com as notícias da Globo obteve maior índice de coerência, e de diversidades, por meio do puw e do pJD. UOL, Record e Jovem

Pan obtiveram maiores scores de diversidade. A Gazeta obteve baixos scores de coerência e de diversidade.

Tabela 5.3 – Métricas de avaliação dos modelos de tópicos

Mídia	Notícias	Tópicos	Coerência c_v	puw	pJd
Globo	24.433	20	0.7578	0.86	0.9854
UOL	6.982	5	0.6177	0.96	0.9834
JPan	959	4	0.6985	0.9	0.9672
r7	3.371	20	0.7197	0.88	0.9909
Gazeta	1.075	3	0.6985	0.63	0.6819

Fonte: Autor

Na Tabela 5.4, apresentam-se as proporções de notícias agrupadas nos cinco primeiros tópicos de cada mídia. No entanto, Jovem Pan e Gazeta não têm um total de cinco tópicos totais, caracterizando, portanto, valores não preenchidos. Observa-se uma maior homogeneidade na distribuição dos cinco primeiros tópicos para as notícias na mídia Globo, seguida pela Record. UOL tem mais de 90% das notícias agrupadas no tópico 1.

Tabela 5.4 – Percentagem de notícias encontradas nos cinco tópicos mais relevantes

Mídia	**da1t**	**da2t**	**da3t**	**da4t**	**da5t**
Globo	23.02%	22.05%	13.02%	9.83%	4.19%
UOL	90.80%	3.92%	2.80%	1.64%	0.81%
JPan	73.12%	16.71%	7.83%	2.32%	-
r7	60.9%	38.68%	18.57%	17.49%	5.042%
Gazeta	59.23%	38.04%	3.72%	-	-

Fonte: Autor

5.6 Classificação de Posicionamento

A classificação de posicionamento, em inglês, *stance classification*, neste trabalho, ocorre a partir do agrupamento de tópicos na etapa de modelagem. Ou seja, somente é possível treinar um classificador de posicionamento, a partir de um subconjunto de nóticias pertencentes a um tópico previamente descoberto.

Assim, para a realização dos experimentos de classificação de posicionamento, elencou-se o tópico denominado Política pois: a) é único comum a todas as cinco mídias; e; b) inerentemente, trata de assuntos relevantes no contexo econômico e social. Ressalta-se, no contexto deste trabalho, a escolha de um único tópico devido às limitações de tempo e escopo.

A construção do classificador se iniciou por meio da rotulagem manual de 64 notícias pertencentes ao tópico Política. Divididas em duas partes iguais, uma metade foi rotulada como "a favor"e outra como "contra". Ou seja, notícias com interpretações favoráveis e contraditórias aos respectivos governos. Nessa rotulagem, observou-se que as notícias remetiam aos mais diversos tipos de governos, sendo federais, estatuais e municipais e, até mesmo, internacionais.

Posteriormente, aplicou-se o método de rotulagem *few-shot learning*, por meio do modelo pré-treinado denominado SetFit (TUNSTALL et al., 2022), às 64 instâncias de notícias rotuladas manualmente. Como resultado do treinamento, obteve-se um classificador binário de posicionamento. Por fim, o classificador foi submetido às demais 12.266 instâncias de notícias não rotuladas, pertencentes ao tópico Política.

6 RESULTADOS

Neste Capítulo, apresentam-se os resultados da modelagem de tópicos e da classificação de posicionamento — respectivamente nas duas seções a seguir 6.1 e 6.2 — obtidos a partir dos experimentos oriundos do estudo de caso descrito no Capítulo 5.

6.1 Resultados da Modelagem de Tópicos

Inicialmente, apresentam-se os resultados da modelagem de tópicos, comparando os cinco modelos treinados para cada mídia. Posteriormente, nas seções seguintes, abordam-se, respectivamente, os resultados individualizados de cada mídia.

Assim, um comparativo dos tópicos gerados a partir das notícias de cada mídia jornalística envolvida nos experimentos — Globo, UOL, Jovem Pan, Record e Gazeta — é apresentado por meio das Figuras 6.1, 6.2, 6.3, 6.4 e 6.5.

Nesse comparativo, apresenta-se a proporção da distribuição dos *n* tópicos encontrados, respectivamente a cada mídia. Observa-se, na constituição dos tópicos, fatores anacrônicos presentes às notícias, como é o caso de Pandemia Covid-19, no ano de 2020, Guerra da Ucrânia, no ano de 2022 e, mais recentemente, em 2023, a Guerra da Palestina.

Observa-se, também, uma variação na quantidade total de tópicos gerados por mídia, sendo a Globo a detentora do maior número de tópicos descobertos, com dezenove ocorrências, seguido por r7, com dezesseis e, em seguida, UOL com cinco ocorrências. Jovem Pan e Gazeta tiveram um quantidade menor de tópicos, quatro e três, respectivamente.

Nota-se, empiracamente, que a quantidade total de tópicos descobertos teve relação com a quantidade de notícias produzidas — e coletadas — por mídia e com interesses editoriais. Neste sentido, houve tópicos comuna às mídias, bem como tópicos individualmente presente em apenas uma mídia, como apresentado na Tabela 6.1 e na Figura 6.6.

Pôde-se analisar a relação de tópicos descobertos entre as mídias por meio de sua visualização através de de grafos. Para tal, há um conjunto de vértices, representados por mídias e por tópicos, relacionando-os por meio de suas respectivas arestas. Deste modo, cada vértice tem um grau de incidência, baseado em seu número de arestas. Ou seja, propiciando mensurar as relações entre uma mídia e os respectivos tópicos de interesse.

Na Tabela 6.1, apresentam-se, ordenadamente, os tópicos representados por vérti-

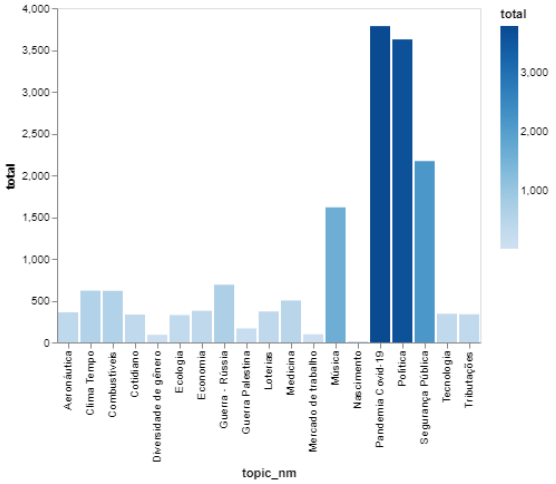


Figura 6.1 – Tópicos encontrados nas notícias da Globo

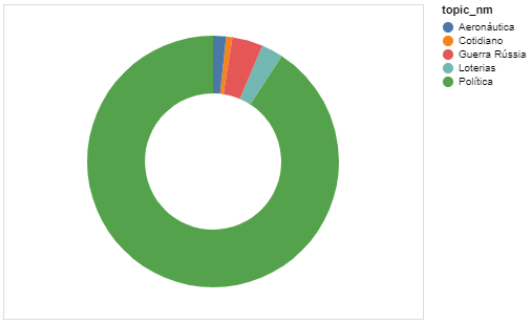


Figura 6.2 – Tópicos encontrados nas notícias do UOL

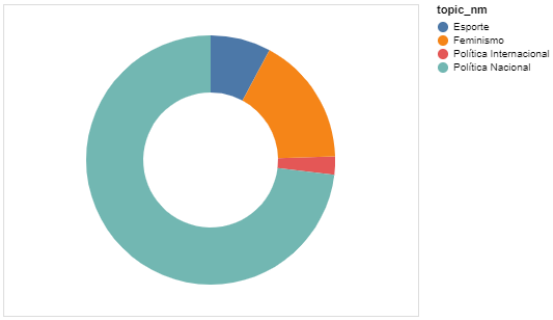


Figura 6.3 – Tópicos encontrados nas notícias da Jovem Pan

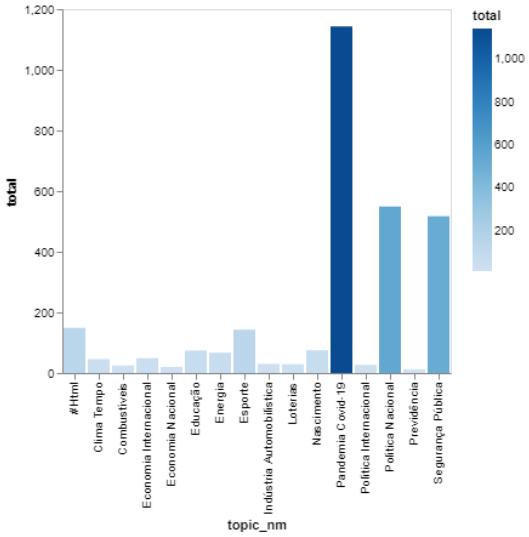
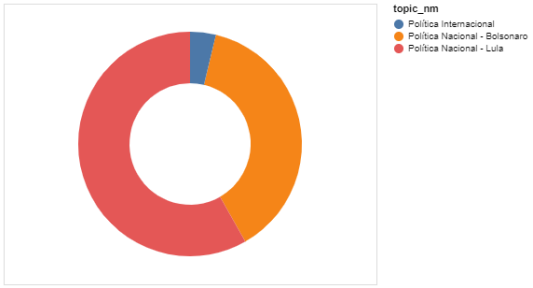


Figura 6.4 – Tópicos encontrados nas notícias da Record

Figura 6.5 – Tópicos encontrados nas notícias da Gazeta



ces do grafo e os respetivos graus. Assim, o tópico presente em todas as mídias possuirá grau igual a cinco. Por outro lado, grau igual a um, representa tópicos de interesse de uma única mídia.

Como exemplo, o tópico denominado Política possui grau igual a cinco, indicando interesse por esse tema em todas as cinco mídias presentes nos experimentos. No caso

de tópicos especificamente encontrado em uma única mídia, como por exemplo o tópico Feminismo, com grau igual a um.

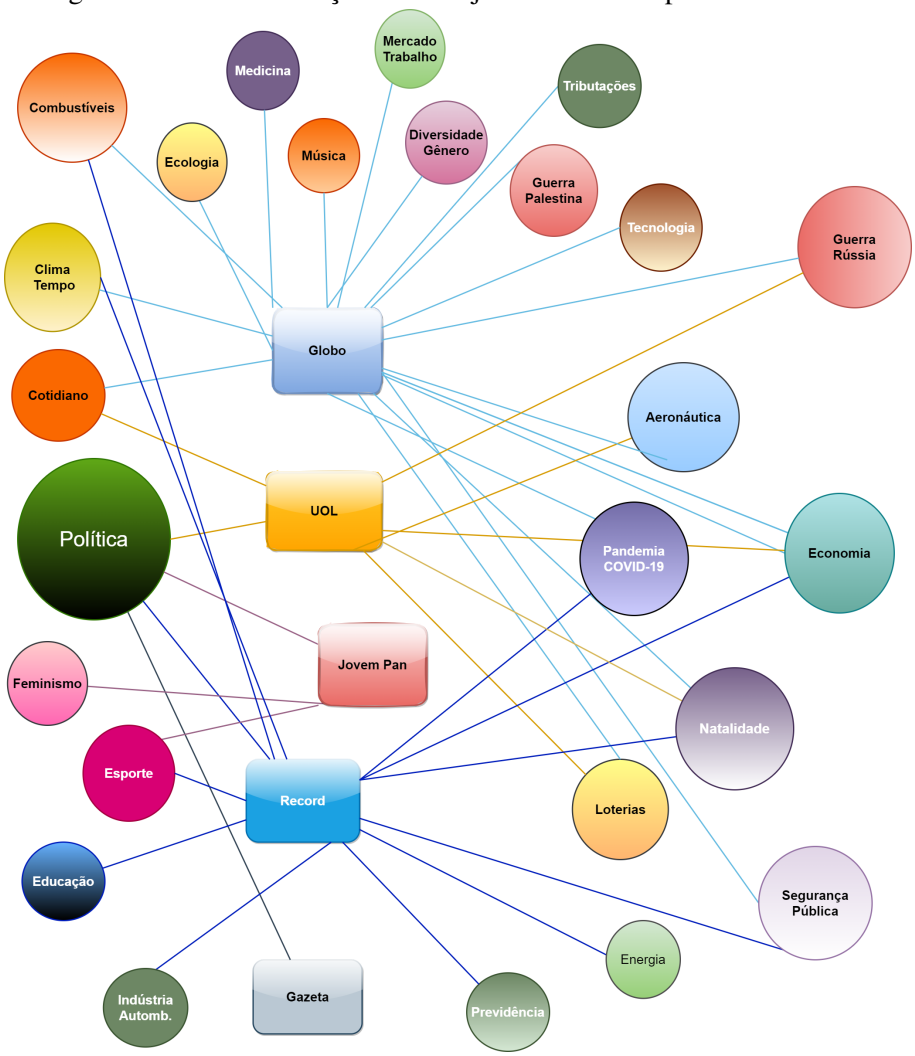
Tabela 6.1 – Vértices e graus de incidência

Vértice	**Grau**	**Vértice**	**Grau**
Política	5	Educação	1
Economia	3	Energia	1
Loterias	3	Diversidade gênero	1
Natalidade	3	Feminismo	1
Aeronáutica	2	Guerra Palestina	1
Clima e Tempo	2	Ind. automobilística	1
Combustíveis	2	Medicina	1
Cotidiano	2	Mercado trabalho	1
Esportes	2	Música	1
Pademia Covid-19	2	Previdência	1
Guerra Rússia	2	Tecnologia	1
Ecologia	1	Tributações	1

Fonte: Autor

Na Figura 6.6, apresenta-se o grafo completo, composto dos dois tipos de vértices V_m e V_t — mídia e tópicos — bem como as respectivas arestas E. A fim de prover uma representação visual, o tamanho do círculo referente ao tópico é diretamente proporcional ao seu grau de incidência.

Figura 6.6 – Grafo - relações mídias jornalísticas x tópicos descobertos



Fonte: Autor

Nas subseções seguintes, separados por mídia jornalística, descrevem-se os respectivos principais resultados encontrados no modelagem de tópicos.

6.1.1 Globo

O modelo da mídia jornalística Globo gerou 19 tópicos. Tendo os dois primeiros, denominados Pandemia Covid-19 e Política, como os mais relevantes, com a quantidade total de notícias próximas. A terceira e quarta posições, respectivamente, Segurança Pública e Música (ou entretenimento) concluem os quatro tópicos mais noticiados, correspondendo a *68.10%* do total de notícias.

Embora os outros 15 tópicos tenham recebido menos relevância, o modelo pôde ser capaz de agrupá-los com uma coerência satisfatória. Ou seja, formando conjunto de

notícias com características comuns dentre os respectivos elementos intra *clusters*.

Na Tabela 6.2, apresenta-se a distribuição ordenada de frequência dos dezenove tópicos gerados pelo modelo da mídia Globo. Nela, têm-se o número, os respectivos nomes atribuídos, o total de notícias e o percentual correspondente a cada tópico.

Tabela 6.2 – Total de notícias por tópico - Globo

Tópico	**Nome**	**Notícias**	**Percentual**
0	Pandemia Covid-19	3785	23.02
1	Política	3626	22.05
2	Segurança Pública	2170	13.20
3	Música	1616	9.83
4	Guerra - Rússia	690	4.19
5	Clima Tempo	619	3.76
6	Combustíveis	617	3.75
7	Medicina	501	3.04
8	Economia	378	2.29
9	Loterias	370	2.25
10	Aeronáutica	360	2.19
11	Tecnologia	342	2.08
12	Tributações	336	2.04
13	Cotidiano	333	2.02
14	Ecologia	327	1.98
15	Guerra Palestina	167	1.01
16	Mercado de trabalho	97	0.59
17	Diversidade de gênero	91	0.55
18	Nascimento	13	0.079

Fonte: Autor

6.1.1.1 Principais Palavras por Tópico

Correspondendo a um subconjunto dos 19 tópicos, os nove gráficos de barras da Figura 6.7 apresentam as dez principais palavras, de diferentes classes gramaticais, que representam os tópicos.

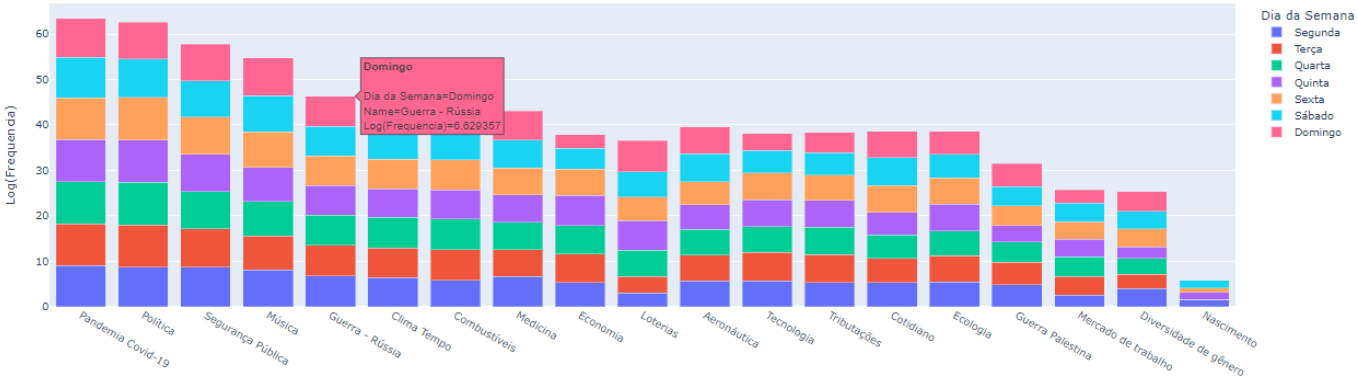
6.1.1.2 Sazonalidade de Notícias por Tópico

O gráfico da Figura 6.8 apresenta uma visualização da produção total de notícias por dia da semana, destacando os respectivos tópicos. Ou seja, a distribuição proporcional dos tópicos pelos sete dias da semana.

Figura 6.7 – Score das palavras em nove tópicos encontrados nas notícias da Globo

Fonte: Autor

Figura 6.8 – Tópicos por dias da semana nas notícias da Globo



Fonte: Autor

6.1.2 UOL

Na Tabela 6.3, apresenta-se a distribuição ordenada de frequência dos cinco tópicos gerados pelo modelo da mídia UOL. Nela, têm-se o número, os respectivos nomes atribuídos, o total de notícias e o percentual correspondente a cada tópico.

6.1.2.1 Principais Palavras por Tópico

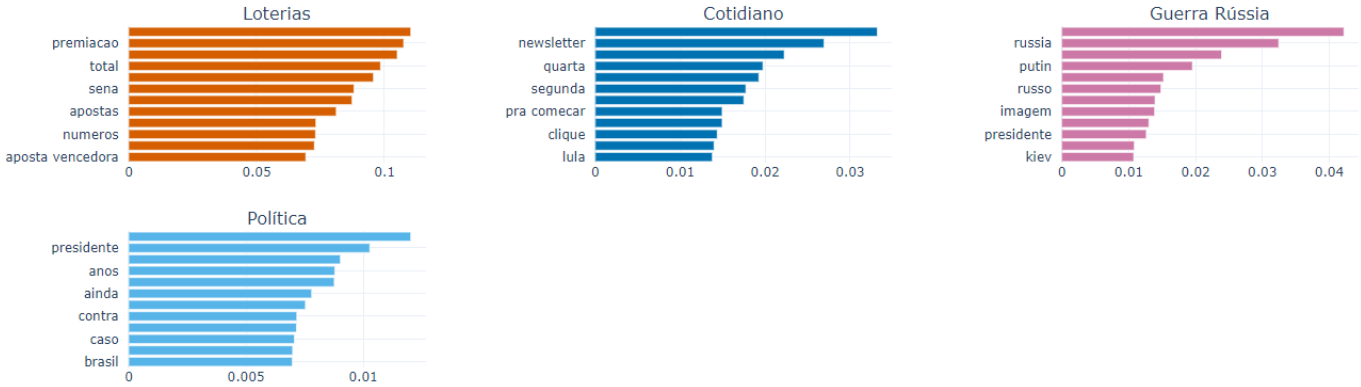
Os quatro gráficos de barras da Figura 6.9 apresentam as seis principais palavras, de diferentes classes gramaticais, que representam os tópicos.

Tabela 6.3 – Total de notícias por tópico - UOL

Tópico	Nome	Notícias	Percentual
4	Política	6340	90.80
2	Guerra Rússia	27	3.92
0	Loterias	196	2.80
3	Aeronáutica	115	1.64
1	Cotidiano	57	0.816

Fonte: Autor

Figura 6.9 – Score das palavras em quatro tópicos encontrados nas notícias do UOL



Fonte: Autor

6.1.2.2 Sazonalidade de Notícias por Tópico

O gráfico da Figura 6.10 apresenta uma visualização da produção total de notícias por dia da semana, destacando os respectivos tópicos.

Figura 6.10 – Tópicos por dias da semana nas notícias do UOL

Fonte: Autor

6.1.3 Jovem Pan

Na Tabela 6.4, apresenta-se a distribuição ordenada de frequência dos quatro tópicos gerados pelo modelo da mídia Jovem Pan. Nela, têm-se o número, os respectivos nomes atribuídos, o total de notícias e o percentual correspondente a cada tópico.

Tabela 6.4 – Total de notícias por tópico - Jovem Pan

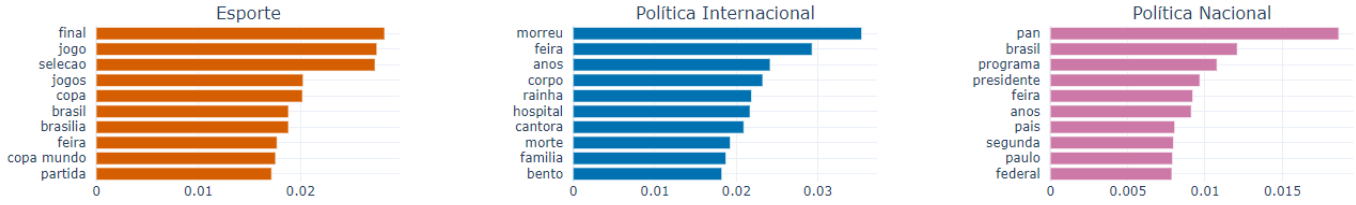
Tópico	Nome	Notícias	Percentual
2	Política Nacional	691	73.12
3	Feminismo	158	16.71
0	Esporte	74	7.83
1	Política Internacional	22	2.32

Fonte: Autor

6.1.3.1 Principais Palavras por Tópico

Os três gráficos de barras da Figura 6.11 apresentam as dez principais palavras, de diferentes classes gramaticais, representativas aos respectivos tópicos.

Figura 6.11 – Score das palavras em quatro tópicos encontrados nas notícias da Jovem Pan



Fonte: Autor

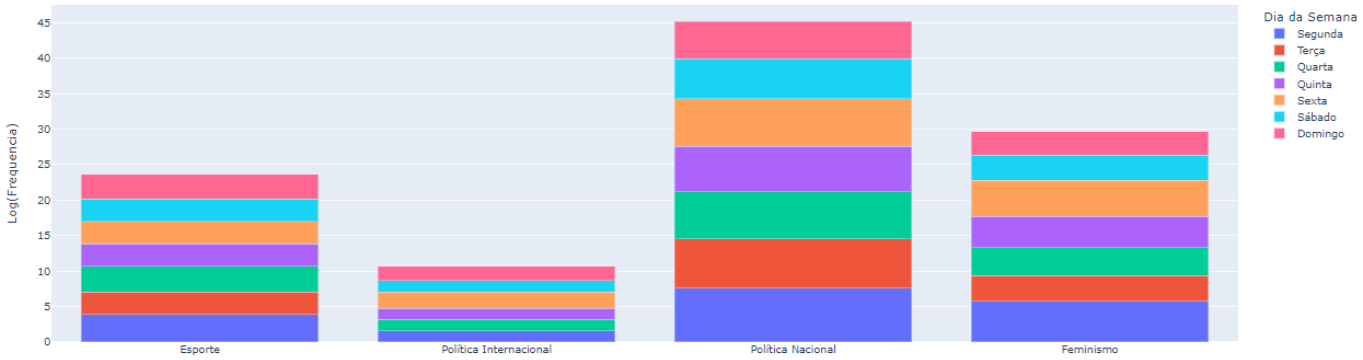
6.1.3.2 Sazonalidade de Notícias por Tópico

O gráfico da Figura 6.12 apresenta uma visualização da produção total de notícias por dia da semana, destacando os respectivos tópicos. Ou seja, a distribuição proporcional dos tópicos pelos sete dias da semana.

6.1.4 Record

Na Tabela 6.5, apresenta-se a distribuição ordenada de frequência dos dezesseis tópicos gerados pelo modelo da mídia Record. Nela, têm-se o número, os respectivos nomes atribuídos, o total de notícias e o percentual correspondente a cada tópico.

Figura 6.12 – Tópicos por dias da semana nas notícias da Jovem Pan



Fonte: Autor

Tabela 6.5 – Total de notícias por tópico - Portal Record

Tópico	Tópico nome	Total notícias	Percentual
0	Pandemia Covid-19	1143	38.68
1	Política Nacional	549	18.58
2	Segurança Pública	517	17.5
3	#Html	149	5.04
4	Esporte	143	4.84
5	Nascimento	75	2.54
6	Educação	74	2.5
7	Energia	67	2.27
8	Economia Internacional	49	1.66
9	Clima Tempo	46	1.56
10	Indústria Automobilística	30	1.02
11	Loterias	29	0.98
12	Política Internacional	27	0.91
13	Combustíveis	25	0.85
14	Economia Nacional	20	0.68
15	Previdência	12	0.41

Fonte: Autor

6.1.4.1 Principais Palavras por Tópico

Os seis gráficos de barras da Figura 6.11 apresentam as dez principais palavras, de diferentes classes gramaticais, representativas aos respectivos tópicos.

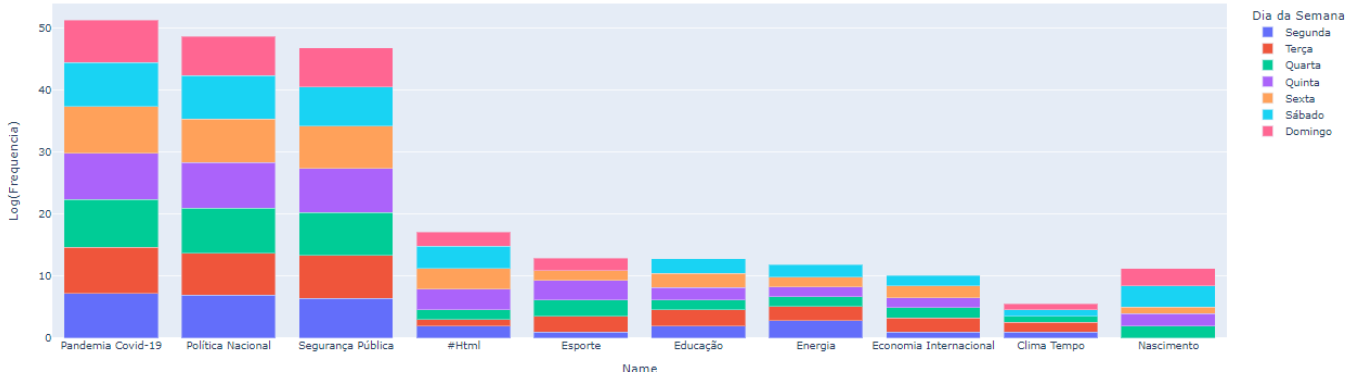
6.1.4.2 Sazonalidade de Notícias por Tópico

O gráfico da Figura 6.14 apresenta uma visualização da produção total de notícias por dia da semana, destacando os respectivos tópicos. Ou seja, a distribuição proporcional dos tópicos pelos sete dias da semana.

Figura 6.13 – Score das palavras em dezesseis tópicos encontrados nas notícias da Record

Fonte: Autor

Figura 6.14 – Tópicos por dias da semana nas notícias da Record



Fonte: Autor

6.1.5 Gazeta

Na Tabela 6.6, apresenta-se a distribuição ordenada de frequência dos três tópicos gerados pelo modelo da mídia Gazeta. Nela, têm-se o número, os respectivos nomes atribuídos, o total de notícias e o percentual correspondente a cada tópico.

Tabela 6.6 – Total de notícias por tópico - Gazeta

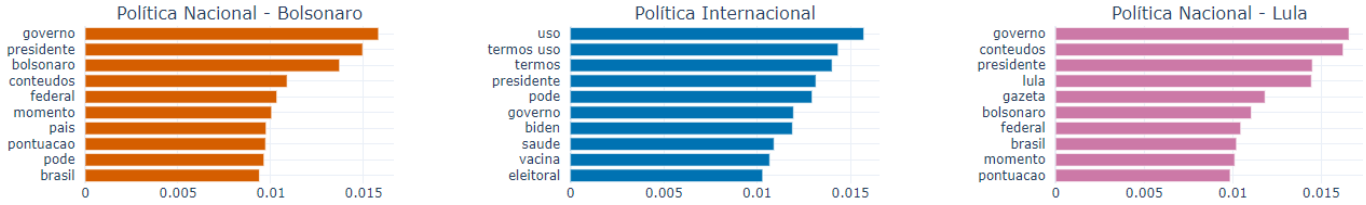
Tópico	Tópico nome	Notícias	Percentual
2	Política Nacional - Lula	626	58.23
0	Política Nacional - Bolsonaro	409	38.046
1	Política Internacional	40	3.72

Fonte: Autor

6.1.5.1 Principais Palavras por Tópico

Os três gráficos de barras da Figura 6.11 apresentam as dez principais palavras, de diferentes classes gramaticais, representativas aos respectivos tópicos.

Figura 6.15 – Score das palavras em três tópicos encontrados nas notícias da Gazeta

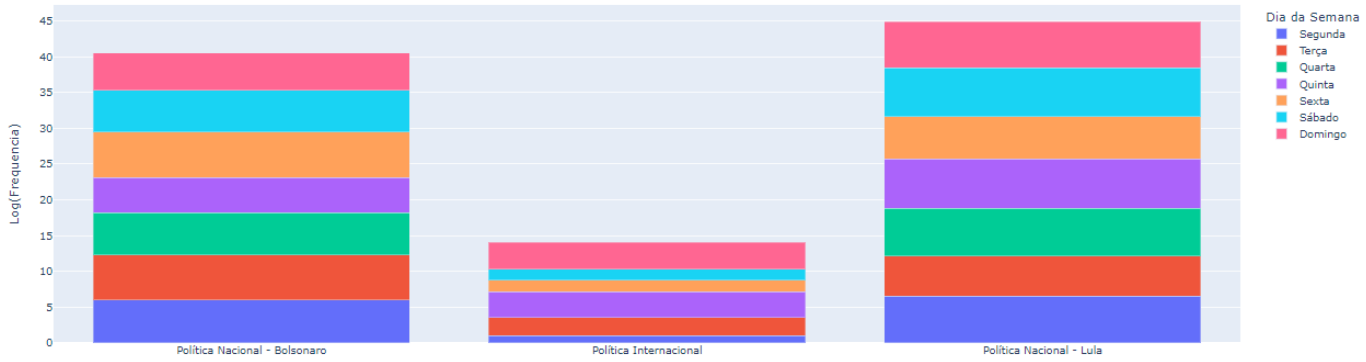


Fonte: Autor

6.1.5.2 Sazonalidade de Notícias por Tópico

O gráfico da Figura 6.16 apresenta uma visualização da produção total de notícias por dia da semana, destacando os respectivos tópicos. Ou seja, a distribuição proporcional dos tópicos pelos sete dias da semana.

Figura 6.16 – Tópicos por dias da semana nas notícias da Gazeta do Povo



Fonte: Autor

6.2 Resultados da Classificação de Posicionamento

Divide-se esta seção de resultados de classificação de posicionamento em duas partes. A primeira se refere à classificação de posicionamentos no contexto do tópico Política, utilizando todo o *corpus*. A segunda parte é composta da primeira acrescentando uma análise temporal, por meio da divisão em dois períodos no tempo, a fim de, porventura, identificar mudanças de posicionamentos, por parte das cinco mídias, na linha do tempo.

6.2.1 Posicionamentos ao Tópico Política

O classificador de posicionamento binário foi submetido às 12.266 instâncias de notícias não rotuladas de todas as 5 mídias, pertencentes ao tópico Política. O resultado da classificação é apresentado na Tabela 6.7 e no gráfico de barras da Figura 6.17. Neles, observa-se uma tendência de equilíbrio entre as duas classes "a favor"e "contra", nas mídias Globo e r7. Ou seja, ambas com distribuições próximas à metade nas duas classes.

No entanto, ocorreu um desbalanceamento nas classes pertencentes às demais mídias. Evidenciado pela mídia Jovem Pan, com $63,81\%$ a mais de posicionamentos favoráveis aos governos, seguindo por Gazeta, com $41,21\%$. Também desbalanceado, em contrapartida às demais mídias, o UOL teve $38,39\%$ a mais de posicionamentos contrários aos governos.

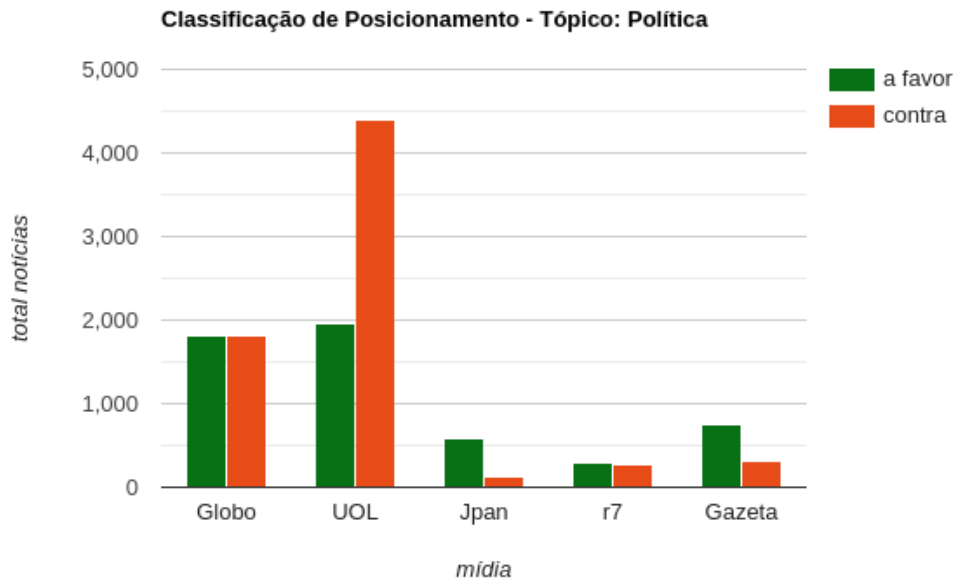
Tabela 6.7 – Posicionamento ao Tópico Política

Tópico	Mídia	Posicionamento	Total	Perc. (%)
Política	Globo	a favor	1,817	50.11
		contra	1,809	49.88
	UOL	a favor	1,953	30.80
		contra	4,387	69.19
	JPan	a favor	584	81.90
		contra	129	18.09
	r7	a favor	302	52.43
		contra	274	47.56
	Gazeta	a favor	759	70.60
		contra	316	29.39

Fonte: Autor

Ressalta-se que, neste trabalho, o classificador de posicionamento não foi capaz de identificar os diversos tipos de governos nos âmbitos, federais, estatuais e municipais

Figura 6.17 – Posicionamento ao Tópico Política



Fonte: Autor

e internacionais.

6.2.2 Posicionamentos sazonal ao Tópico Política em dois períodos de governos

Nesta análise, objetivou-se identificar as possíveis mudanças de posicionamentos, por parte das cinco mídias, na linha do tempo. Para tal, dividiu-se o conjunto de notícias em dois períodos no tempo. O primeiro é compreendido entre outubro de 2020 até o último dia do ano de 2022. O segundo se inicia em janeiro de 2023 até o mês de outubro deste mesmo ano. Esta divisão teve como critério estipular um marco no tempo, baseando-se nas eleições de 2022, a fim de observar os posicionamentos das mídias em relações aos governos antes e após o pleito.

O resultado da classificação de posicionamentos por períodos é apresentado na Tabela 6.8 e no gráfico de barras horizontais da Figura 6.18. Neles, não se observa um consenso, acerca do posicionamento no tempo, por parte das cinco mídias. A Globo aumentou os posicionamentos favóreis aos governos em 7.39%, seguido por Jovem Pan com 5.16%.

Por outro lado, Gazeta e UOL aumentaram os poscionamentos desfavoráveis, respectivamente, em 27.03% e 4.85%. Devido às dificuldades na coleta de notícias do r7, relatado no Capítulo 5, não foi possível, comparar as notícias do segundo período de

tempo.

Portanto, observa-se, como destacado no gráfico de barras da Figura 6.18 que a Globo foi a única mídia que mudou, na totalidade, o seu posicionamento perante os governos. No primeiro período, as notícias desfavoráveis eram majoritarias, invertendo-se no segundo período.

A Gazeta tem o maior desbalanceamento das classes entre os dois períodos. Ou seja, apesar de não mudar o posicionamento na totalidade, observou-se um aumento da notícias desfavoráveis.

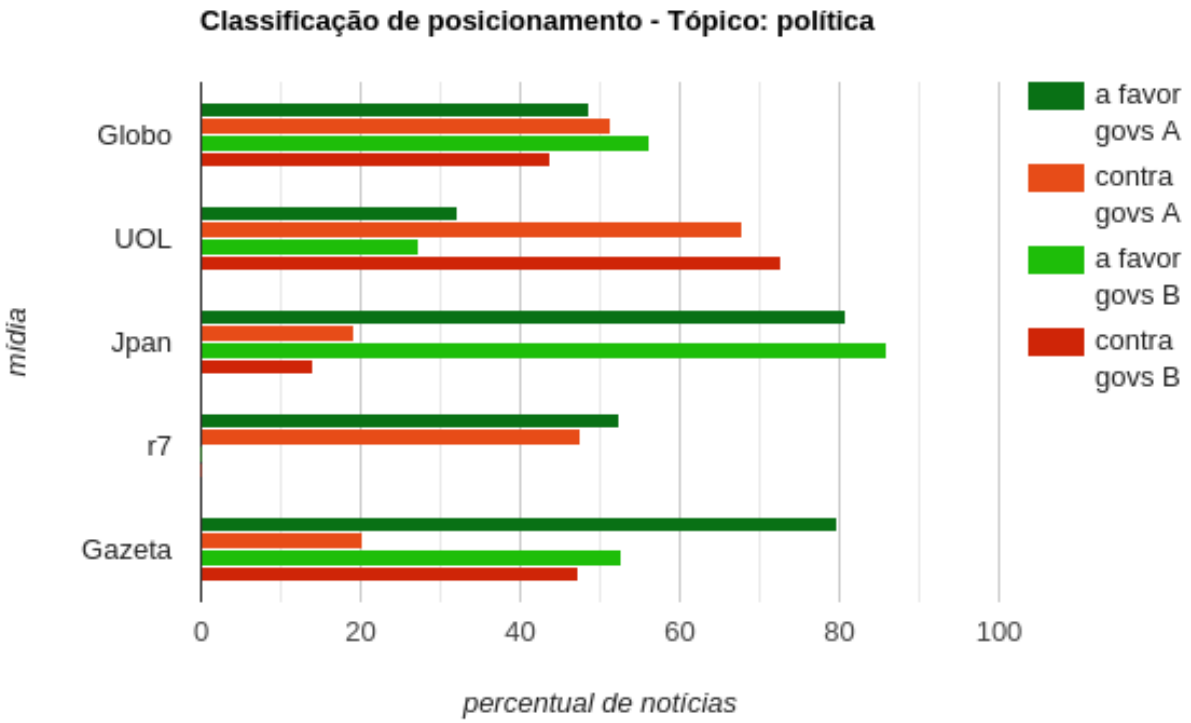
O UOL apresentou menor desbalanceamento, mantendo um equilíbrio entre as duas classes nos dois períodos, com leve crescimento de notícias desfavoráveis aos governos.

Tabela 6.8 – Posicionamento sazonal ao Tópico Política

			Governos A		**Governos B**	
Tópico	**Mídia**	**Posicionamento**	**Total**	**(%)**	**Total**	**(%)**
Política	Globo	a favor	1,442	48.74	375	56.13
		contra	1,516	51.25	293	43.86
	UOL	a favor	1,471	32.15	482	27.30
		contra	3,104	67.84	1283	72.69
	JPan	a favor	444	80.72	140	85.88
		contra	106	19.27	23	14.11
	r7	a favor	302	52.43	-	-
		contra	274	47.56	-	-
	Gazeta	a favor	567	79.74	192	52.74
		contra	144	20.25	172	47.25

Fonte: Autor

Figura 6.18 – Posicionamento sazonal ao Tópico Política



Fonte: Autor

7 CONCLUSÃO

Este trabalho de pesquisa teve como principal objetivo investigar e implementar um *framework* visando a coleta, o processamento e o agrupamento dos textos de notícias produzidos por mídias jornalísticas digitais no Brasil nos últimos três anos, evidenciando os principais temas de interesse das mesmas. Dessa forma, os resultados obtidos propiciaram uma interpretação dos respectivos interesses editoriais.

No âmbito teórico, realizou-se uma revisão bibliográfica acerca dos recentes avanços da modelagem de tópicos, aplicada a textos de notícias, por meio do SBERT e BERTopic. Além disso, complementando a parte teórica, realizou-se uma investigação de trabalhos relevantes acerca do uso de abordagens quantitativas para o tratamento do problema do viés da mídia.

A partir dos conceitos estudados, concebeu-se o *framework*. Posteriormente, visando a sua verificação e a validação, ele foi aplicado, por meio de suas cinco etapas, em um estudo de caso envolvendo 45.036 notícias produzidas, em um período de 36 meses, por cinco mídias jornalísticas brasileiras.

Portando, como resultados globais dos experimentos, obteve-se, para cada uma das cinco mídias, os tópicos e as respectivas proporções de relevância, indicando os principais interesses jornalísticos de cada mídia. Nessa análise, observou-se a política, como sendo o tópico mais recorrente em todas as mídias, seguido por economia, loterias e natalidade.

Como resultados específicos de cada mídia, obteve-se, além dos tópicos e das respectivas proporções de relevância, dados de sazonalidade semanal. Nesta análise, observou-se que, de um modo geral, o número de notícias dentre os principais tópicos de cada mídia não têm relação com dia da semana. Com exceção do tópico denominado cotidiano, no UOL, o qual apresentou um número maior de notícias aos finais de semana.

Por fim, a etapa de classificação visou a identificação do posicionamento das mídias, dentro do tópico política, acerca das textos das notícias produzidas. Assim, nesta classificação binária, comparou-se, para cada uma das cinco mídias, a respectiva distribuição de notícias com textos a favor e contra os governos, ao longo de dois momentos no tempo, representados pelos governos atuais a anteriores.

8 TRABALHOS FUTUROS

Visando o aprimoramento e evolução deste trabalho, sugere-se, para trabalhos futuros:

1) O desenvolvimento de novos coletores *web crawlers*, visando obter mais notícias de novas mídias jornalísticas;

2) A reexecução dos coletores já desenvolvidos, de modo a obter notícias de outros períodos de tempo, aprimorando e enriquecendo, assim, o inédito *corpus* criado neste trabalho;

3) Retreinar os cinco modelos de cada para mídia, desta vez, de modo a realizar um processo de *tuning de hiperparâmetros*, visando otimizar os resultados das métricas dos respectivos treinos;

4) Otimizar o treinamento e os testes do classificador de posicionamento (baseado na rotulagem pelo método de *few-shot learning*), de modo a se avaliar métricas de classificação;

5) Aplicar a classificação de posicionamento em outros tópicos; e

6) Avaliar o classificador utilizado para determinar posicionamentos em base anotada a ser gerada.

REFERÊNCIAS

AIRES, V. P. S. Detecção de viés ideológico de portais de notícias na web. 2020.

BARNIDGE, M. et al. Comparative Corrective Action: Perceived Media Bias and Political Action in 17 Countries. **International Journal of Public Opinion Research**, v. 32, n. 4, p. 732–749, 11 2019. ISSN 1471-6909. Available from Internet: <https://doi.org/10.1093/ijpor/edz043>.

BERNHARDT, D.; KRASA, S.; POLBORN, M. Political polarization and the electoral effects of media bias. **Journal of Public Economics**, v. 92, p. 1092–1104, 03 2008.

BROMLEY J., G. I. L. Y. S. E.; SHAH, R. Signature verification using a "siamese" time delay neural network. **In Advances in neural information processing systems**, p. 737–74, 1994.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J. et al. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 160–172. ISBN 978-3-642-37456-2.

CHURCHILL, L. S. R. The evolution of topic modeling. **ACM Computing Surveys**, v. 54, p. 1–35, 2022.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

FERT, Z. S. **Comparision of Topic Modeling Algorithms on News Articles**. Dissertation (Master) — Tilburg University, School of Humanities and Digital Sciences - Department of Cognitive Science & Artificial Intelligence, 2022.

GREENE, C. Effects of news media bias and social media algorithms on political polarization. 2019.

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

HERMAN, E.; CHOMSKY, N. **Manufacturing Consent: The Political Economy of the Mass Media**. Random House, 2010. ISBN 9781407054056. Available from Internet: <https://books.google.com.br/books?id=Kv_-bvCqgrEC>.

HINNEBURG, A.; RöDER, M. Exploring the space of topic coherence measures. In: SEARCH, I. P. of the eighth ACM international conference on W.; MINING data (Ed.). [S.l.]: ACM international, 2015. p. 399–408.

HUGGINGFACE. 2023. Https://huggingface.co/. Accessed: 2023-11-02.

JACCARD, P. Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisines. **Bulletin de la Société Vaudoise des Sciences Naturelles**, v. 25, n. 2, p. 241–272, 1901.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd edition draft). ed. [S.l.]: Prentice-Hall, 2021.

KHURANA ADITYA KOLI, K. K. D.; SINGH, S. Natural language processing: state of the art, current trends and challenges. 2019.

MCINNES, L.; HEALY, J.; MELVILLE, J. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2020.

MELO, A. S. d. C. **Análise de Viés na Cobertura da Imprensa baseada em Conteúdo Textual de Notícias**. Thesis (PhD) — Universidade Federal de Campina Grande Centro de Engenharia Elétrica e Informática Coordenação de Pós-Graduação em Ciência da Computação, 2020.

MOSER, G. V. B. et al. Análise de similaridade entre tf-idf e modelos contextualizados de linguagem baseados em tokens. Universidade Federal de Santa Catarina - UFSC, 2022.

NEWMAN, N. et al. Reuters institute digital news report 2021. **Reuters Institute for the study of Journalism**, 2021.

R, H. A. The political economy of mass consumption. In: HISTORY, J. of U. (Ed.). [S.l.]: Journal of Urban History, 2006. p. 32(4):607–618.

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

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: INUI, K. et al. (Ed.). **EMNLP/IJCNLP (1)**. [S.l.]: Association for Computational Linguistics, 2019. p. 3980–3990. ISBN 978-1-950737-90-1.

REIMERS, N.; GUREVYCH, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2020. Available from Internet: <https://arxiv.org/abs/2004.09813>.

ROSSUM, G. V.; JR, F. L. D. **Python**. [S.l.]: Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

SOUZA G., L. S. F. C. L. A.; NAZARé, S. Aplicação de redes neurais siamesas na autenticação de condutores. Simposio Brasileiro de Automação Inteligente - SBAI, 2019. Available from Internet: <https://proceedings.science/proceedings/100113/_papers/111100/download/fulltext_file2>.

TANDOC, Z. W. L. E. C.; LING, R. Defining "fake news". **Digital Journalism**, Routledge, v. 6, n. 2, p. 137–153, 2018. Available from Internet: <https://doi.org/10.1080/21670811.2017.1360143>.

TUNSTALL, L. et al. **Efficient Few-Shot Learning Without Prompts**. 2022.

XU ZILING LUO, H. X. M.; WANG, B. Media bias and factors affecting the impartiality of news agencies during covid-19. Behavioral Sciences, School of Electronic Information and Communications, Huahzhong University of Science and Technology, Wuhan 430074, China, Aug 2022.

YIGIT-SERT I. S. ALTINGOVDE, O. U. S. Towards detecting media bias by utilizing user comments. In: SCIENCE, A. C. on W. (Ed.). [S.l.]: Proceedings of the 8th ACM Conference on Web Science,, 2016. p. 374–375.

ZIPF, G. **The Psycho-Biology of Language: An Introduction to Dynamic Philology**. Routledge, 1935. (Cognitive psychology). ISBN 9780415209762. Available from Internet: <https://books.google.com.br/books?id=w1Z4Aq-5sWMC>.

9 APÊNDICE

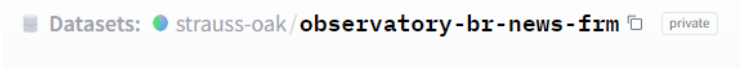
9.1 Dados Brutos Coletados

Figura 9.1 – Informações dos arquivos de dados brutos coletados e armazenados localmente

Fonte: Autor

9.2 Dados Brutos Armazenados

Figura 9.2 – Repositório dos dados brutos armazenados no Hugging Face



Fonte: Autor