UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

TBC-DTW: Medição automática de viés midiático baseada em modelagem de tópicos.

Daniel Pereira Zitei

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

Daniel Pereira Zitei

TBC-DTW: Medição automática de viés midiático baseada em modelagem de tópicos.

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ricardo Marcondes Marcacini

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Zitei, Daniel Pereira
S856m TBC-DTW: Medição automática de viés midiático baseada em modelagem de tópicos. / Daniel Pereira Zitei ; orientador Ricardo Marcondes Marcacini. – São Carlos, 2023.
61 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Marcacini, Ricardo Marcondes, orient. II. Título.

Daniel Pereira Zitei

TBC-DTW: Automatic Measurement of Media Bias Based on Topic Modeling.

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Ricardo Marcondes Marcacini

Original version

São Carlos

2023

RESUMO

Zitei, D. P. **TBC-DTW: Medição automática de viés midiático baseada em modelagem de tópicos.**. 2023. 61p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A crescente presença de notícias falsas, incompletas ou incorretas levanta preocupações no meio acadêmico. Esta pesquisa tem como objetivo desenvolver e avaliar uma abordagem não supervisionada para identificar viés na mídia em língua portuguesa. Para isso, são abordadas duas questões-chave. Primeiro, como organizar artigos de notícias em tópicos usando métodos não supervisionados, incorporando a taxonomia do International Press Telecommunications Council (IPTC) para criar uma nova abordagem de modelagem de tópicos. Em segundo lugar, como quantificar o viés entre diferentes veículos de comunicação, com base em características extraídas do conteúdo de artigos de notícias. O estudo utiliza técnicas avançadas como Sentence-BERT, BERTopic e Dynamic Topic Modeling para atingir esses objetivos. A métrica TBC-DTW (Dynamic Time Warping) é introduzida como uma abordagem inovadora para medir o viés na mídia ao longo do tempo. Os resultados indicam o potencial dessa abordagem para compreender e comparar o viés na mídia em diferentes fontes.

Palavras-chave: LLM. Ideologia. Bertopic. IPTC.

ABSTRACT

Zitei, D. P. **TBC-DTW: Automatic Measurement of Media Bias Based on Topic Modeling.**. 2023. 61p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A increasing presence of fake, incomplete, or incorrect news raises concerns in the academic community. This research aims to develop and evaluate an unsupervised approach to identify media bias in Portuguese-language media. To achieve this, two key questions are addressed. First, how to organize news articles into topics using unsupervised methods, incorporating the International Press Telecommunications Council (IPTC) taxonomy to create a new topic modeling approach. Secondly, how to quantify bias among different media outlets based on features extracted from news articles' content. The study employs advanced techniques such as Sentence-BERT, BERTopic, and Dynamic Topic Modeling to achieve these objectives. The TBC-DTW (Dynamic Time Warping) metric is introduced as an innovative approach to measure media bias over time. The results indicate the potential of this approach to understand and compare media bias across different sources.

Keywords: LLM. Ideology. Bertopic. IPTC.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Notícias falsas, incompletas ou incorretas vêm cada vez mais tomando o discurso sob diversas frentes dentro da academia. Realizando pesquisas no Google Scholar com as seguintes palavras chaves: notícias falsas jornalismo (FALSAS), notícias incompletas jornalismo (INCOMPLETAS), notícias incorretas jornalismo (INCORRETAS) temos os resultados apresentados na Tabela 1.

Tabela 1 - Quantidade de artigos catalogados no Google Schoolar

Entrada	QTD FAQM*	QTD FD2019**	QTD FD2019 / QTD FAQM
FALSAS	21.100	15.400	72,99%
INCOMPLETAS	15.700	8.320	52,99%
INCORRETAS	11.400	3.960	34,73%
TOTAL	48.200	27.680	57,42%

- * QTD FAQM - Quantidade de artigos catalogados com o filtro "a qualquer momento"
- ** QTD FD2019 - Quantidade de artigos catalogados com o filtro "desde 2019"

Pode-se observar que 57,42% de artigos científicos publicados sobre os três assuntos são dos últimos 4 anos. Essas publicações vem aumentando de produção junto com ideias de que a disseminação de notícias falsas ou do inglês *fake news* podem acabar em totalitarismos:

> "A ADMINISTRAÇÃO TRUMP também intensificou esforços para tirar de campo jogadores importantes do sistema político. Os ataques retóricos de Trump contra críticos na mídia são um exemplo disso. Suas acusações reiteradas de que espaços como o New York Times e a CNN estavam distribuindo “fake news” e conspirando contra ele soam familiares a qualquer estudante de autoritarismo."(LEVITSKY; ZIBLATT, 2018)

No Brasil a mídia nos últimos meses vem publicando sobre o assunto pois o atual presidente da república Luiz Inácio Lula da Silva está discutindo um possível projeto de lei para a regulação da mídia, (BRASIL, 2022): Lula volta a defender regulação de rádio, TV e internet e reformar “direito de resposta”, além do presidente a própria mídia por meio de seus colunistas vem demonstrando interesse sobre o assunto como (VILAçA P. E MOURA, 2022): Regulação da mídia x censura: um guia para não cair em pegadinhas e (BARBOSA, 2022): Até quando a regulação da mídia será um tabu no Brasil?

Esses são alguns exemplos de reportagens e de colunas quem vem ocupando os noticiários nos últimos meses. A mídia, segundo Stuart Hall no seu artigo "política da significação", organiza e entrega uma narrativa específica, por meio das práticas de cada veículo, eles acabam empregando sua própria concepção histórica (MARX K. E ENGELS, 2007) e suas ideologias. Existe um consenso entre os críticos de qualquer veículo de comunicação que esses lugares é onde acontecem as lutas ideológicas sobre o significado ((ALTHUSSER; BREWSTER, 2001), (GRAMSCI, 1971), (THOMPSON, 1984), (HALL, 1985)) para definir o que nós como sociedade devemos aceitar ou não, o que é o certo e o que o errado, o que é normal e o que é anômalo.

Um dos grandes estudos da área foi o livro *Comparing Media Systems* (HALLIN; MANCINI, 2004), na qual eles trazem 4 dimensões de análises dos veículos de comunicação em 18 países do ocidente sendo elas as estruturas do mercado de notícias, o nível de profissionalização do jornalismo, o papel do estado na intervenção da mídia e o paralelismo político. Os autores descrevem que cada uma dessas dimensões possuem formas de quantificar atributos na qual são úteis no processo de entendimento de como a mídia se comporta em cada país. O livro nos aponta direções importantes para o entendimento do comportamento da mídia mas pode possuir viés humano na quantificações pois todas as análises foram coletadas e inferidas pelos próprios pesquisadores.

No mundo atual, onde as informações fluem de forma instantânea e abrangente, compreender como a mídia molda nossas perspectivas e influencia nossas opiniões é uma tarefa importante, uma vez que a parcialidade na cobertura de notícias pode conduzir a divisões sociais, polarização e uma compreensão distorcida da realidade.

A utilização de métodos manuais para medir o viés na mídia, embora historicamente empregados, tem um problema “paradoxal”. Tais métodos manuais podem ter seus próprios vieses. Por exemplo, a seleção de quais tópicos devem ser utilizados na amostragem, pode ser influenciada. Além disso, tal estratégia não é escalável para lidar com o volume crescente e dinâmico das notícias.

Artigos recentes trazendo técnicas de aprendizagem de máquina e técnicas de extração de dados em larga escala vem nos ajudando na identificação do paralelismo político, uma dessas formas foi descrita em *Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale* (RIBEIRO *et al.*, 2018), na qual, através da iteração nas redes sociais Twitter e Facebook dos usuários que consomem esses veículos de mídia podemos quantificar esse paralelismo político, desta forma os autores inferem que o viés do jornal pode ser dado pela forma de interação dos usuário e não necessariamente observando como os textos do jornal são escritos.

Em abordagens mais recentes podemos ver artigos como o *Measuring Media Bias via Masked Language Modeling* (GUO; MA; VOSOUGHI, 2022) em que possuindo *datasets* sobre notícias de diversos jornais dos Estados Unidos da América e os tópicos por meio

de bigramas gerados pelos autores, podemos usar o algoritmo de aprendizagem profunda (*Masked Language Modeling*) para identificar a quantidade de viés usando a distância entre cada veículo de mídia sobre determinado tópico.

Tabela 2 - Ranking de quantidade de jornais em circulação no Brasil

Jornal	Quantidade de Jornais em circulação
O Globo	60.777
Estadão	60.446
Super Notícia	48.215
Folha	48.084
Zero Hora	41.425
Extra	35.105
Meia Hora	27.905
Valor	14.541
Estado de Minas	11.128
Correio Brasiliense	11.044
O Tempo	9.816

A proposta geral deste projeto é a investigação de técnicas de aprendizado de máquina não supervisionadas para identificar tópicos em diferentes meios de comunicação e, em seguida, quantificar o viés e diferença em relação aos meios de comunicação. O viés é quantificado de diferentes formas, por exemplo, identificando diferenças de conteúdo, vocabulário e expressões de vários meios para um mesmo tópico ou evento. As abordagens existentes exploram especialistas de domínio ou modelos de aprendizado de máquina supervisionados que, por sua vez, também são criticados por incorporarem viés na análise.

De maneira alternativa, neste projeto, serão investigadas abordagens alternativas para a modelagem de tópicos em cada meio de comunicação. Especificamente, propõe-se um modelo de referência para a organização de tópicos fornecido pelo *International Press Telecommunications Council* (IPTC). Posteriormente, serão explorados modelos que identificam diferentes distribuições, bem como análises qualitativas das palavras utilizadas nos tópicos de cada um dos três jornais impressos de maior circulação no país, conforme dados do Instituto Verificador de Comunicação (ver Tabela 1). Isso proporcionará informações matemáticas sobre o paralelismo político presente nos jornais brasileiros. Para este estudo, foi dada prioridade aos jornais O Globo, Estadão e Folha.

Neste projeto são investigados métodos para entender e quantificar o viés midiático através da análise de tópicos e da evolução temporal. Primeiro, adotar uma estratégia computacional minimiza a subjetividade de algumas decisões e reduz o viés pessoal na análise. Além disso, o uso de modelos de linguagem treinados em grandes conjuntos de dados permite que a identificação de tópicos e a análise de sua evolução sejam realizadas de

maneira consistente e automatizada. Outro fator relevante é a escalabilidade, permitindo a análise de um grande volume de informações. Assim, é apresentada a seguinte questão de pesquisa: como podemos quantificar e comparar o viés entre diferentes meios de comunicação, de forma automática, a fim de compreender melhor como as narrativas divergem entre diferentes mídias?

Para responder essa questão de pesquisa, é proposto um método, denominado de TBC-DTW, baseado em modelagem de tópicos e sua evolução temporal. O método desenvolvido possui uma etapa para coleta de dados de diferentes fontes de notícias, como portais de notícias, utilizando mecanismos de web scraping. Em seguida, há a etapa de pré-processamento, visando obter uma representação estruturada. A próxima etapa envolve a modelagem de tópicos baseado em modelos de linguagem, bem como sua análise temporal para caracterizar como certos tópicos variam ao longo do tempo. Uma contribuição desse projeto é a proposta de uma medida de comparação entre tópicos, que considera tanto a distribuições de palavras-chave de tópicos semelhantes, quanto sua distribuição temporal.

1.1 Objetivos e Questões de Pesquisa

O objetivo geral deste projeto é desenvolver e avaliar uma abordagem não supervisionada para identificação de viés em meios de comunicação na língua portuguesa. Para atingir o objetivo geral, são levantadas as seguintes questões de pesquisa:

- **Como organizar de forma não supervisionada um conjunto de notícias em tópicos?**

  Embora existam diferentes métodos de modelagem de tópicos que atuam de forma não supervisionada, é reconhecido que as variações de seus parâmetros podem ocasionar grandes diferenças na geração dos tópicos. Para tal, é proposto restringir o modelo de geração de tópicos a partir de um modelo de referência conhecido como IPTC (International Press Telecommunications Council), que apresenta uma taxonomia de aproximadamente 1300 temas para notícias ao redor do mundo. Além de fornecer um vocabulário controlado para os tópicos, também contém uma organização hierarquizada para analisar em diferentes níveis de granularidade.

  No entanto, o IPTC não foi proposto no contexto de modelagem de tópicos. Assim, um desafio de pesquisa deste projeto é adaptar um método de modelagem de tópicos para incorporar o conhecimento da taxonomia do IPTC. Espera-se que ao responder essa questão de pesquisa, seja disponibilizado um novo método para modelagem de tópicos em notícias, sem a necessidade de anotação por parte de especialista, mas respeitando a taxonomia do modelo do IPTC.

- **Como quantificar viés entre diferentes meios de comunicação?**

A quantificação de viés é um desafio em aberto e pouco explorado por meio de métodos não supervisionados, principalmente na língua portuguesa. Nesse projeto, espera-se extrair características que cada meio de comunicação utiliza para comunicar uma notícia em um determinado tópico. A extração dessas características será baseada por meio de modelos de linguagem. A ideia geral é quantificar a relação entre palavras utilizadas dentro de um mesmo tópico por diferentes meios de comunicação, bem como explorar características estatísticas dessas palavras. Assim, é uma estratégia de quantificação baseada em conteúdo e que não depende de especialistas para anotação de dados.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Diferentes estudos anteriores mostram evidências significativas de que há um viés nos meios de comunicação, especialmente partidarismo, que pode influenciar a opinião pública (JR; SHAH, 2003; KOHUT *et al.*, 2011; GUO; MA; VOSOUGHI, 2022). Nesses estudos, afirma-se que o viés de meios de comunicação geralmente ocorre por meio de filtragem de problemas e assuntos (BUDAK; GOEL; RAO, 2016), bem como o enquadramento e forma de apresentar tais assuntos. A maioria dos trabalhos existentes, no entanto, tenta quantificar tal viés usando estratégias qualitativas e muitas vezes manuais, o que dificulta inclusive a reprodução dos resultados obtidos por esses métodos (GUO; MA; VOSOUGHI, 2022).

É importante destacar métodos de quantificação de viés por meio de conteúdo, que comparam diferentes meios de comunicação por meio da forma de escrita e palavras escolhidas para reportar os mesmos eventos ou tópicos (LIM; JATOWT; YOSHIKAWA, 2018; SPINDE *et al.*, 2021). São métodos que dependem de humanos para rotulação prévia de notícias consideradas com algum tipo de viés. Embora usar dados anotados facilite representar o problema para identificação e quantificação de viés como uma tarefa de aprendizado de máquina, a própria rotulação humana, no entanto, adiciona um viés em relação ao treinamento desses modelos (FÄRBER *et al.*, 2020).

Uma alternativa recente é o uso de modelos de linguagem pré-treinados a partir de grandes bases textuais para apoiar a quantificação de viés em meios de comunicação, conforme discutido em (GUO; MA; VOSOUGHI, 2022). Nesse caso, os modelos são utilizados para inferir a escolha de palavras que cada meio de comunicação tende a utilizar em cada contexto (BHARGAVA; NG, 2022). Embora essa estratégia tenha vantagens em relação às anteriores em relação à menor subjetividade por não depender de dados rotulados, ainda há um desafio sobre comparar diferentes meios de comunicação em relação a um mesmo contexto ou evento. Os autores em Guo, Ma and Vosoughi (2022) definiram alguns tópicos de interesse, mas a escolha manual desses tópicos também pode ser um ponto para inclusão de viés na quantificação.

Uma maneira de lidar com esse problema é inicialmente organizar todos os textos dos meios de comunicação a serem comparados em um mesmo conjunto de tópicos. Essa organização deve ser preferencialmente não supervisionada e ser a mais abrangente possível para reduzir o viés. Uma estratégia popular é por meio da modelagem de tópicos que visa descobrir temas comuns ou tópicos latentes em bases textuais (VAYANSKY; KUMAR, 2020).

Os modelos mais clássicos, como LDA (BLEI; NG; JORDAN, 2003) e NMF

(FÉVOTTE; IDIER, 2011), utilizam uma representação baseada em bag-of-words que são reconhecidos por desconsiderar ordem e relações semânticas entre palavras (AGGARWAL; AGGARWAL, 2015). Como consequência, também falham em identificar contexto de palavras em frases. Os modelos mais recentes também já utilizam representações baseadas em *word embeddings* aprendidos em modelos neurais de linguagem, como o BERT (DEVLIN *et al.*, 2019) e suas variações mais recentes (GRUETZEMACHER; PARADICE, 2022). Tais modelos possuem a capacidade de representar textos em espaço vetorial capaz de computar similaridade textual de maneira mais efetiva.

De forma geral, conforme discutido em Sia, Dalmia and Mielke (2020) e Grootendorst (2022), os documentos são inicialmente representados no espaço de *embeddings* por meio de um modelo de linguagem pré-treinado (e.g. BERT). Em seguida, um método de agrupamento é utilizado para identificar grupos de documentos similares. Cada grupo é um candidato a um tópico e as palavras mais frequentes de cada grupo (e que não aparecem frequentemente em outros grupos) são selecionadas para determinar o tópico representado pelo grupo.

Embora essa estratégia de extração de tópicos possa ser uma solução viável para alinhar assuntos de diferentes meios de comunicação para, em seguida, quantificar viés em cada tópico, vale destacar que tal estratégia é muito sensível à escolha de parâmetros do método de agrupamento e escolha do modelo de linguagem. Pequenas perturbações nos parâmetros provocam grandes diferenças nos tópicos obtidos, sendo difíceis de definirem e interpretar (ABDELRAZEK *et al.*, 2022).

Considerando esses desafios, neste projeto será explorado alternativas para geração de tópicos por meio de um modelo de referência já reconhecido na mídia, denominado IPTC (*International Press Telecommunications Council*)[1]. Esse modelo de referência contém um padrão para organização de conteúdo de notícias, incluindo eventos e informações bem organizadas sobre pessoas e organizações. Dessa forma, espera-se desenvolver um modelo tópico orientado ao IPTC e com foco na quantificação de viés entre diferentes meios de comunicação, por meio das diferenças de vocabulário e na forma de abordar assuntos em cada tópico.

2.1 Sentece-Bert

A primeira fase, após a extração de dados, consiste na aplicação do algoritmo Sentence-BERT. Em 2018, baseando-se nessa arquitetura, Devlin e sua equipe desenvolveram o BERT (Bidirectional Encoder Representations from Transformers), um modelo de linguagem pré-treinado que estabeleceu recordes de estado da arte (SOTA) para várias tarefas de Processamento de Linguagem Natural (PLN), incluindo o benchmark de Similaridade Textual Semântica (STS) de Cer e colaboradores em 2017. O S-BERT (Sentence-BERT)

[1] IPTC: <https://www.iptc.org/>

é uma modificação da rede BERT pré-treinada que utiliza estruturas de rede siamesa e tripla para derivar representações de sentenças semanticamente significativas, que podem ser calculadas usando a similaridade de cosseno.

No passado, os métodos de incorporação de sentenças neurais começavam com inicialização aleatória. No entanto, o S-BERT utiliza modelos pré-treinados, como BERT e RoBERTa, e ajusta-os para produzir incrustações de sentenças de tamanho fixo. Isso é alcançado por meio de uma estrutura de rede siamesa, conforme ilustrado na Figura 1.

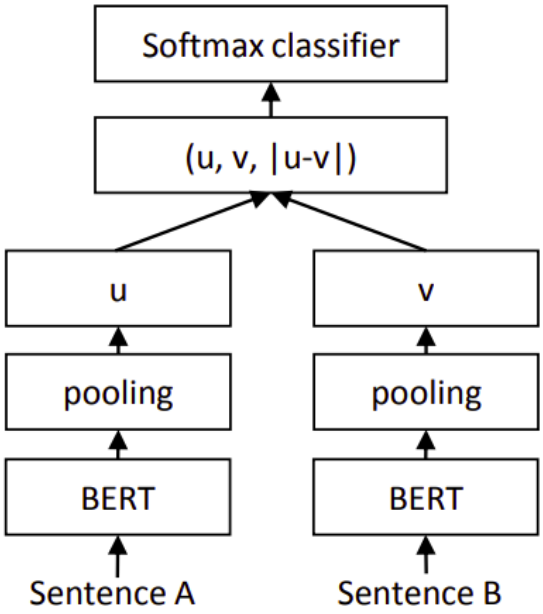


Figura 1 – Representação da Rede Siamesa.

[2]

Na rede siamesa, BERT ou RoBERTa podem ser empregados como modelos de linguagem pré-treinados. A estratégia de agrupamento padrão envolve o cálculo da média de todos os vetores de saída, onde u e v representam as incrustações de sentenças. Antes de passar pelo classificador softmax, a etapa de concatenação é realizada na forma de $(u, v, |u - v|)$, que foi determinada como a mais adequada para o ajuste fino. A função de objetivo de classificação é otimizada por meio da perda de entropia cruzada e pode ser representada pela seguinte equação:

$$W_t \in R^{3n \times k}, \quad o = softmax(W_T(u, v, |u - v|)) \tag{2.1}$$

Após o treinamento, o S-BERT emprega uma função de objetivo regressiva para a inferência dentro de uma rede siamesa, semelhante àquela utilizada para o ajuste fino.

[2] Cópia de de Figura presente no artigo (REIMERS; GUREVYCH, 2019)

Como ilustrado na Figura 2, a similaridade de cosseno entre duas incrustações de sentenças, representadas como u e v, é calculada como um escore variando de -1 a 1.

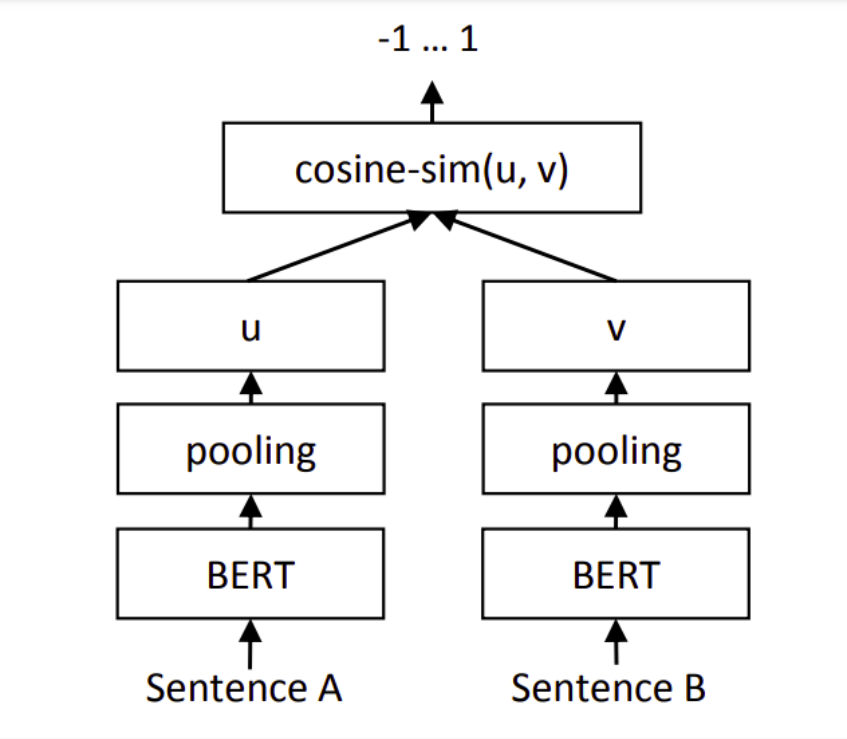


Figura 2 – Arquitetura SBERT com função de objetivo de classificação.

[3]

2.2 Modelagem de Tópicos

Para alcançar isso, inicia-se por ajustar o BERTopic como se não houvesse um aspecto temporal nos dados. Isso resulta na criação de um modelo de tópicos geral que captura a estrutura global dos principais tópicos, que podem variar ao longo de diferentes intervalos de tempo. Para cada tópico e intervalo de tempo, calcula-se a representação c-TF-IDF (Term Frequency-Inverse Document Frequency personalizado).

[3] Cópia de de Figura presente no artigo (REIMERS; GUREVYCH, 2019)

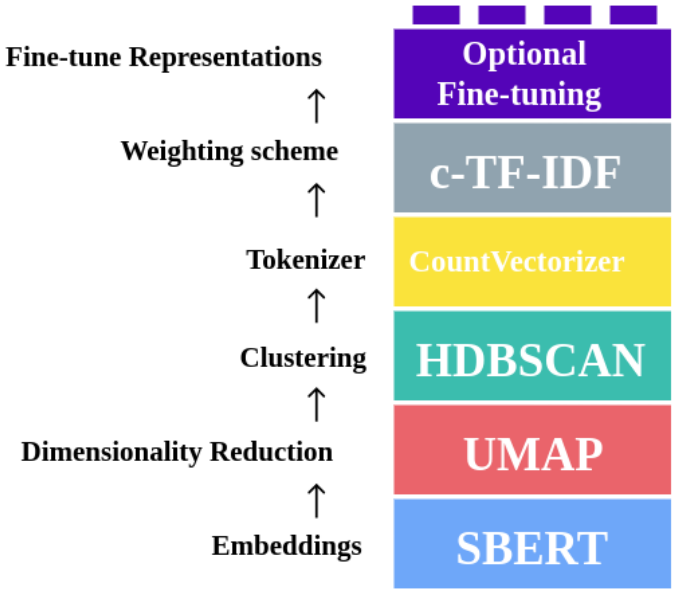


Figura 3 – Funcionamento do Bertopic - Retirado de (GROOTENDORST, 2022)

Essa abordagem nos permite obter uma representação específica dos tópicos em cada intervalo de tempo sem a necessidade de criar agrupamentos a partir das incrustações, uma vez que eles já foram gerados previamente. Além disso, no contexto da análise de tópicos dinâmicos, é importante mencionar alguns dos métodos e técnicas utilizados:

1. S-BERT (Sentence-BERT): Como mencionado anteriormente, o S-BERT é um componente essencial que gera representações semânticas de sentenças, permitindo uma análise mais profunda dos textos jornalísticos.

2. UMAP (Uniform Manifold Approximation and Projection) é uma técnica de redução de dimensionalidade que pode ser aplicada no contexto do BERTopic para visualizar e analisar a distribuição dos tópicos de forma mais eficaz. Ele ajuda a identificar padrões e relacionamentos entre os tópicos.

3. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) é um algoritmo de clustering que pode ser aplicado ao BERTopic para agrupar tópicos semelhantes. Isso ajuda na organização e interpretação dos tópicos extraídos dos textos jornalísticos.

4. CountVectorizing uma técnica que pode ser usada em conjunto com o BERTopic para transformar os textos em representações vetoriais, levando em consideração a frequência das palavras. Isso pode ser útil na análise de tópicos.

5. C-TF-IDF (Custom Term Frequency-Inverse Document Frequency) é uma adaptação do TF-IDF que pode ser aplicada no BERTopic para calcular a importância de termos em relação aos tópicos em um contexto específico.

2.3 Medição de Viés na Mídia

Iniciativas anteriores na literatura se basearam em abordagens predominantemente manuais ou semi-automatizadas para avaliar o viés na mídia. No entanto, nosso interesse reside em adotar uma abordagem automatizada.

Outros métodos se apoiam em dicionários ou recursos léxicos como em (LAZARIDOU; KRESTEL, 2016) e (HAMBORG; DONNAY; GIPP, 2019), o que constitui uma limitação para nossa pesquisa. Essas abordagens costumam ser específicas para determinados contextos (por exemplo, política).

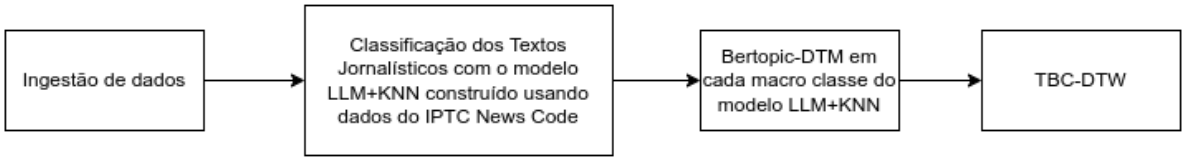
Além disso, os métodos existentes utilizam representações de texto que podem não ser suficientemente flexíveis para capturar as nuances inerentes a diferentes tópicos e suas mudanças ao longo do tempo. Nesse sentido, a utilização de modelos de linguagem, como o BERT, se mostra fundamental para superar essas limitações. Vale ressaltar que a análise da evolução temporal dos tópicos é de extrema importância, uma dimensão que muitas iniciativas anteriores negligenciam.

Um ponto crucial a ser considerado é que muitos métodos presumem a disponibilidade prévia de uma base de dados. No entanto, nos dias de hoje, as mídias e portais de notícias frequentemente impõem restrições à acessibilidade de seus dados históricos. Isso inclui mecanismos como paywalls ou captchas, que exigem a assinatura paga para acesso completo. Portanto, é altamente recomendável que abordagens para medir o viés na mídia incorporem técnicas de raspagem de dados (web scraping). No entanto, é importante notar que, até o momento, nenhum dos estudos existentes integrou de maneira eficaz essas técnicas em suas metodologias.

3 DESENVOLVENDO TRABALHO

A Figura 4 ilustra o método de análise envolvendo todas as etapas, incluindo a ingestão dos dados, criação de um modelo de classificação de textos, predição dos textos ingeridos e medição do viés, na qual serão descritas com mais detalhes nos capítulos seguintes.

Figura 4 – Esteira de desenvolvimento do trabalho



3.1 Ingestão de Dados

Nesta etapa, será descrito como os dados são coletados e preparados para análise. Isso pode envolver a obtenção dos conjuntos de dados correspondentes e a limpeza dos dados, removendo informações irrelevantes ou ruidosas.

A ingestão de dados no método é divido em duas etapas, a etapa de captura dos dados de jornais para fazer a análise de viés e os dados para criar o modelo de classificação de notícias.

3.1.1 Jornais

Para iniciar o processo de ingestão de notícias, foi realizado um levantamento dos oito principais jornais do país, conforme apresentado na Tabela 2. Para realizar o teste do método foi realizada a busca no intervalo de tempo de 01/02/2018 até 31/12/2022. Cada veículo possui suas próprias formas de exibir notícias em HTML e pode bloquear requisições HTTP diretas. Para contornar esses obstáculos, o projeto utiliza a biblioteca Selenium, que permite realizar chamadas por meio de um navegador, possibilitando varrer os jornais e coletar as notícias de forma automatizada.

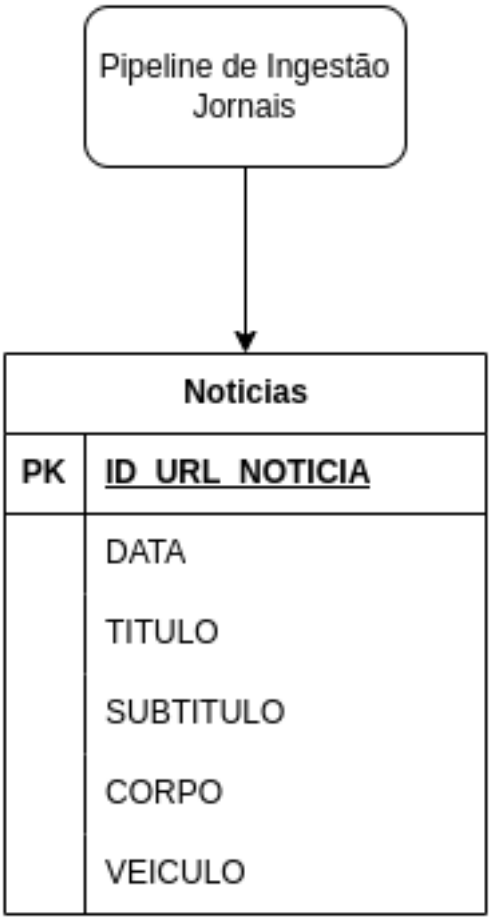
A biblioteca Selenium é uma ferramenta poderosa que permite a automação de interações em páginas web, como preenchimento de formulários, cliques em botões e rolagem de página. Com ela, podemos simular o comportamento de um usuário real ao navegar pelos jornais online. Dessa forma, conseguimos contornar bloqueios e acessar o conteúdo desejado para coleta das notícias.

Ao utilizar o Selenium, podemos desenvolver scripts para percorrer as páginas dos jornais, encontrar os elementos que contêm as notícias, extrair seu conteúdo e armazená-lo para posterior processamento. É importante ressaltar que, como cada jornal possui uma estrutura HTML diferente, será necessário desenvolver um script específico para cada um, adaptando-se às particularidades de cada site.

Inicialmente, aproveitando a capacidade do Selenium de interagir com o navegador, é possível fazer chamadas às páginas dos jornais para buscar os identificadores (IDs) das notícias do dia. Esses IDs servirão como referência para localizar e extrair os detalhes das notícias posteriormente.

Após o armazenamento dos IDs das notícias, inicia-se a segunda etapa, que consiste na busca dos detalhes de cada notícia. Utilizando os IDs previamente obtidos, os scripts específicos para cada jornal são responsáveis por percorrer novamente as páginas, localizar os elementos que contêm os detalhes desejados (como título, subtitulo, data de publicação e texto completo, vide Figura 5) e extrair essas informações. É possível também capturar outros elementos multimídia, como imagens ou vídeos associados às notícias, se necessário.

Figura 5 – Ingestão de Jornais



Os detalhes das notícias extraídas podem ser armazenados em um formato adequado ao projeto, como um banco de dados, arquivos estruturados ou qualquer outro meio de armazenamento apropriado. A fim de padronizar a forma de busca das notícias nos jornais, a Figura 5 representa qual o formato de cada objeto retirado das notícias. O *schema* pode ser interpretado da seguinte forma:

- ID_URL_NOTICIA: url do site da notícia;
- DATA: data da notícia;
- TITULO: título da notícia;
- SUBTITULO: subtítulo da notícia;
- CORPO: corpo da notícia;
- VEICULO: nome do veículo da notícia;

3.1.2 IPTC

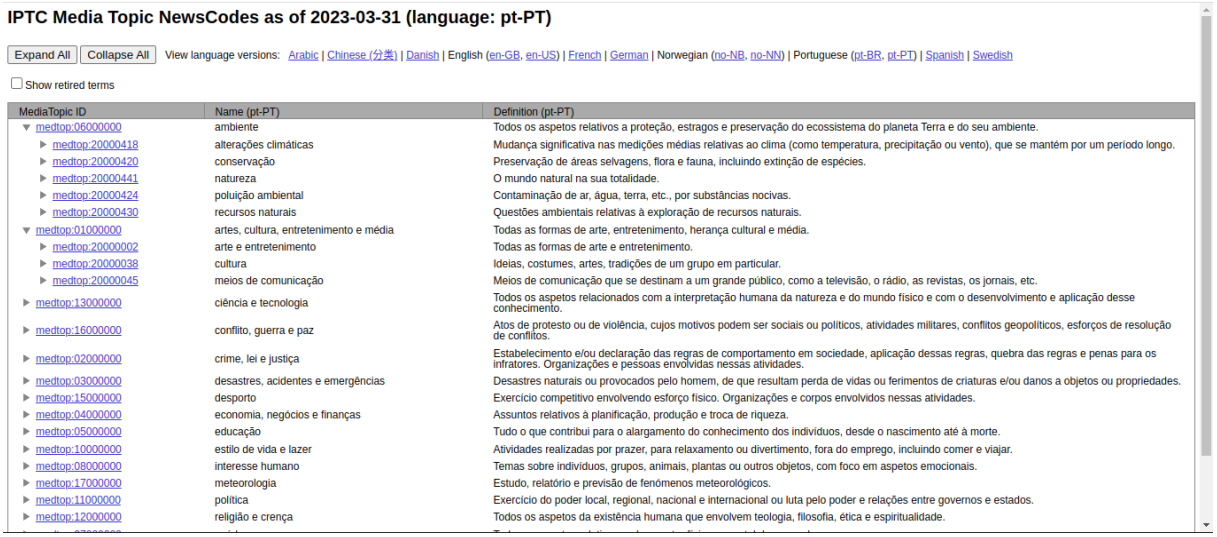
Para capturar os dados para criar o modelos de classificação foi utilizado o *IPTC Media Topic NewsCodes*. *IPTC Media Topic NewsCodes*, também conhecido como *IPTC NewsCodes*, é um sistema de classificação e categorização usado para rotular e organizar notícias e conteúdos jornalísticos. Desenvolvido pela *International Press Telecommunications Council* (IPTC), esse sistema permite que as notícias sejam categorizadas de forma padronizada e consistente em diferentes plataformas e organizações de mídia.

Os *NewsCodes* do IPTC são uma lista hierárquica de categorias e subcategorias que abrangem uma ampla gama de tópicos, temas e assuntos. Cada código representa uma categoria específica, como política, esportes, entretenimento, saúde, economia, ciência, tecnologia, meio ambiente, entre outros. Essa estrutura hierárquica permite que as notícias sejam classificadas em níveis gerais e específicos, fornecendo uma granularidade adequada para a categorização. O IPTC diponibiliza a versão desta lista em vários idiomas como ilustra a Figura 6.

Segundo o IPTC, o NewsCodes, é um recurso valioso para a indústria de mídia, permitindo uma organização eficiente e uma melhor compreensão dos tópicos e temas abordados nas notícias. Sua adoção generalizada promove a consistência na categorização de conteúdo, facilita a colaboração e o compartilhamento de notícias entre diferentes plataformas e contribui para uma melhor experiência de consumo de informações para os leitores e espectadores.

Da mesma forma que as notícias, os dados capturados no site do IPTC são organizados em uma tabela com o seguinte formato:

Figura 6 – NewsCodes IPTC



IPTC Media Topic NewsCodes as of 2023-03-31 (language: pt-PT)

Expand All | Collapse All | View language versions: Arabic | Chinese (分类) | Danish | English (en-GB, en-US) | French | German | Norwegian (no-NB, no-NN) | Portuguese (pt-BR, pt-PT) | Spanish | Swedish

Show retired terms

MediaTopic ID	Name (pt-PT)	Definition (pt-PT)
medtop:06000000	ambiente	Todos os aspetos relativos a proteção, estragos e preservação do ecossistema do planeta Terra e do seu ambiente.
medtop:20000418	alterações climáticas	Mudança significativa nas medições médias relativas ao clima (como temperatura, precipitação ou vento), que se mantém por um período longo.
medtop:20000420	conservação	Preservação de áreas selvagens, flora e fauna, incluindo extinção de espécies.
medtop:20000441	natureza	O mundo natural na sua totalidade.
medtop:20000424	poluição ambiental	Contaminação de ar, água, terra, etc., por substâncias nocivas.
medtop:20000430	recursos naturais	Questões ambientais relativas à exploração de recursos naturais.
medtop:01000000	artes, cultura, entretenimento e média	Todas as formas de arte, entretenimento, herança cultural e média.
medtop:20000002	arte e entretenimento	Todas as formas de arte e entretenimento.
medtop:20000038	cultura	Ideias, costumes, artes, tradições de um grupo em particular.
medtop:20000045	meios de comunicação	Meios de comunicação que se destinam a um grande público, como a televisão, o rádio, as revistas, os jornais, etc.
medtop:13000000	ciência e tecnologia	Todos os aspetos relacionados com a interpretação humana da natureza e do mundo físico e com o desenvolvimento e aplicação desse conhecimento.
medtop:16000000	conflito, guerra e paz	Atos de protesto ou de violência, cujos motivos podem ser sociais ou políticos, atividades militares, conflitos geopolíticos, esforços de resolução de conflitos.
medtop:02000000	crime, lei e justiça	Estabelecimento e/ou declaração das regras de comportamento em sociedade, aplicação dessas regras, quebra das regras e penas para os infratores. Organizações e pessoas envolvidas nessas atividades.
medtop:03000000	desastres, acidentes e emergências	Desastres naturais ou provocados pelo homem, de que resultam perda de vidas ou ferimentos de criaturas e/ou danos a objetos ou propriedades.
medtop:15000000	desporto	Exercício competitivo envolvendo esforço físico. Organizações e corpos envolvidos nessas atividades.
medtop:04000000	economia, negócios e finanças	Assuntos relativos à planificação, produção e troca de riqueza.
medtop:05000000	educação	Tudo o que contribui para o alargamento do conhecimento dos indivíduos, desde o nascimento até à morte.
medtop:10000000	estilo de vida e lazer	Atividades realizadas por prazer, para relaxamento ou divertimento, fora do emprego, incluindo comer e viajar.
medtop:08000000	interesse humano	Temas sobre indivíduos, grupos, animais, plantas ou outros objetos, com foco em aspetos emocionais.
medtop:17000000	meteorologia	Estudo, relatório e previsão de fenómenos meteorológicos.
medtop:11000000	política	Exercício do poder local, regional, nacional e internacional ou luta pelo poder e relações entre governos e estados.
medtop:12000000	religião e crença	Todos os aspetos da existência humana que envolvem teologia, filosofia, ética e espiritualidade.

- CHAVE_IPTC: chave primária do IPTC;
- TEMA: nome do tema dado pelo IPTC;
- DESCRICAO: descrição do tema feita pelo IPTC;

3.2 Weak Model

O primeiro passo em nossa abordagem é treinar o Modelo Fraco (WM) com base no algoritmo dos k-vizinhos mais próximos (k-NN). Dado um conjunto de dados pré-processado, composto por amostras de texto $\{x_1, x_2, ..., x_n\}$ e suas respectivas etiquetas de classe $\{y_1, y_2, ..., y_n\}$, nosso objetivo é usar uma função de incorporação $f : x \rightarrow R^d$ que mapeia cada texto x_i para um vetor d-dimensional que captura sua informação semântica. A função de incorporação pode ser matematicamente definida como:

$$e_i = f(x_i), \quad parai = 1, 2, ..., n \tag{3.1}$$

Onde $e_i \in R^d$ representa o vetor de incorporação para o texto x_i. Essa função de incorporação utiliza a potência de um Modelo de Linguagem de Grande Escala (LLM) para gerar as incorporações semânticas.

O Modelo Fraco é construído utilizando essas incorporações. Quando uma nova amostra de texto x é introduzida, o sistema calcula sua incorporação e_x e encontra os k vizinhos mais próximos no espaço de incorporação. Isso pode ser representado como:

$$top - kvizinhosmaispróximosx = argmin_{x_i}\, d(e_x, e_i) \tag{3.2}$$

Onde $d(\cdot, \cdot)$ representa uma métrica de distância, como distância euclidiana ou cosseno, utilizada para medir a similaridade entre as incorporações.

Durante o treinamento, os parâmetros do Modelo Fraco são otimizados para aprimorar seu desempenho na identificação das k classes mais prováveis para cada texto. Isso é feito maximizando a probabilidade das etiquetas de classe corretas estarem dentro dos k vizinhos mais próximos. A otimização visa encontrar os parâmetros θ ótimos, resolvendo o seguinte problema de otimização:

$$\hat{\theta} = textargmax_{\theta} \sum_{i=1}^{n} I(y_i \in Top - kneighborsof x_i) \tag{3.3}$$

Onde $I(\cdot)$ é uma função indicadora que retorna 1 se a condição for verdadeira e 0 caso contrário. Esse objetivo garante que o Modelo Fraco foque na identificação das classes corretas dentro dos k vizinhos mais próximos. Esse processo de classificação dos k vizinhos mais próximos serve como uma etapa preliminar de filtragem, reduzindo o espaço de busca para as etapas subsequentes e melhorando a eficiência computacional.

3.3 Geração de Prompt para o Modelo de Linguagem de Grande Escala (LLM - Large Language Model)

Nesta etapa, aproveita-se a saída do Modelo Fraco (WM) para gerar um prompt específico para cada texto, que servirá como entrada para o Modelo de Linguagem de Grande Escala (LLM) na classificação final. O prompt é cuidadosamente construído para incorporar informações relevantes das k classes identificadas pelo WM. Isso pode incluir etiquetas de classe, descrições textuais ou outros recursos informativos contextualmente relevantes que auxiliem o LLM na geração de previsões precisas. A geração do prompt é realizada por meio de uma abordagem de aprendizado de prompt "few-shot". Isso envolve o ajuste fino do LLM utilizando uma quantidade limitada de dados rotulados, especificamente as amostras de texto associadas às k classes principais de acordo com a saída do WM. Matematicamente, representa-se o aprendizado de prompt "few-shot"da seguinte forma. Seja Dtrain o conjunto de treinamento composto por amostras de texto rotuladas e suas respectivas etiquetas de classe. Dada a saída do WM que identifica as k principais classes para um texto x, denotadas como Ctop-k, extraímos as amostras rotuladas correspondentes de Dtrain:

$$D_{few-shot} = \{(x_i, y_i) \,|\, (x_i, y_i) \in D_{train}, and y_i \in C_{top-k}\} \tag{3.4}$$

Utiliza-se então Dfew-shot para o ajuste fino do LLM, permitindo que ele gere previsões mais precisas com base no prompt fornecido. Matematicamente, denota-se o prompt como P e a etiqueta de classe alvo como y. O objetivo é gerar um prompt que maximize a probabilidade condicional da etiqueta de classe alvo dado o prompt, representada como $P(y|P)$. A tarefa do LLM é gerar a próxima palavra condicionada ao prompt. Isso pode ser formulado como:

$$P(y|P) = P(y|Prompt) = \prod_{t=1}^{T} P(y_t|y_{t-1}, Prompt) \tag{3.5}$$

onde y_t representa a palavra prevista no passo de tempo t, e o "Prompt"representa o prompt fornecido ao LLM. A probabilidade de cada palavra y_t é condicionada às palavras anteriores y_{t-1} e ao prompt

No exemplo acima, onde o Modelo Fraco (WM) identifica três classes principais para um determinado texto:

Texto de entrada: "Um terremoto devastador atingiu uma pequena cidade, causando danos generalizados e perda de vidas". Usando o WM (primeira etapa), obtemos as dez classes mais prováveis com base na saída do WM. Neste caso, supondo que k=10. As classes principais com base na saída do WM são as seguintes:

1. Terremotos
2. Desastres e Acidentes
3. Planejamento de Emergência
4. Desastres Naturais
5. Atividade Sísmica
6. Geologia
7. Crise e Gerenciamento de Emergências
8. Ciências da Terra
9. Acidentes e Desastres Ambientais
10. Incidentes de Emergência

Agora, pode-se construir o prompt incorporando essas informações:

Classificar o seguinte texto com uma das seguintes categorias:

1. Terremotos
2. Desastres e Acidentes
3. Planejamento de Emergência
4. Desastres Naturais
5. Atividade Sísmica
6. Geologia
7. Crise e Gerenciamento de Emergências
8. Ciências da Terra
9. Acidentes e Desastres Ambientais
10. Incidentes de Emergência

Texto: "Um terremoto devastador atingiu uma pequena cidade, causando danos generalizados e perda de vidas." Resposta:

Neste exemplo, o prompt indica claramente a tarefa de classificar o texto fornecido relacionado a um desastre natural recente em uma das categorias fornecidas, permitindo que o Modelo de Linguagem de Grande Escala (LLM) se concentre apenas nas classes relevantes da tarefa de classificação. Além disso, ao estruturar o prompt dessa forma, possibilita-se que o LLM gere uma saída estruturada, fornecendo uma etiqueta prevista que pode ser usada nas etapas subsequentes do nosso método.

3.3.1 Aplicando em Dataset Conhecido

Este dataset consiste em um conjunto de dados classificado e rotulado com base nas categorias do DBPEDIA, um projeto que visa coletar informações estruturadas sobre entidades do mundo real e categorizá-las de acordo com diversas classes. O dataset contém 21.900 textos que foram associados a categorias específicas do DBPEDIA.

O conjunto de categorias neste dataset abrange uma extensa variedade de tópicos, capturando a riqueza e diversidade de entidades e conceitos do mundo real. Entre as categorias representadas neste conjunto de dados, encontram-se, mas não se limitam a, as seguintes.

- Categorias de Entretenimento: Como "Ator Adulto,Álbum,Mangá", e "Filme de Animação de Hollywood".
- Categorias Esportivas: Tais como "Jogador de Futebol Americano", "Equipe de Basquete,Atleta de Ciclismo", e "Evento de Mixed Martial Arts".
- Categorias Políticas e Sociais: Como "Presidente", "Filósofo", "Poeta", "Partido Político", e "Suprema Corte dos Estados Unidos - Caso".
- Categorias Geográficas e Naturais: Incluindo "Vulcão", "Montanha", "Rio", "Planeta", "Vila", e "Galeria".
- Categorias Culturais: Tais como "Artista de Música Clássica", "Compositor de Música Clássica", "Escritor de Teatro", e "Cervejaria".

Métrica	**Valor**
Máximo de ganho (llm+knn $\geq$ knn)	53.33%
Média de ganho (llm+knn $\geq$ knn)	12.49%
Contagem de classes em que llm+knn $\geq$ knn	168
Máximo de perda (llm+knn $\leq$ knn)	-51.35%
Média de perda (llm+knn $\leq$ knn)	-9.10%
Contagem de classes em que llm+knn $\leq$ knn	66
Média da diferença entre ganho e perda	3.39%
Diferença de precisão	5.79%
Precisão llm+knn	86.28%
Precisão knn	80.49%

Tabela 1 – Resumo das Métricas

Os resultados obtidos, na tabela 1, a partir da análise das métricas de desempenho dos modelos de classificação LLM+KNN (*Linguistic Model+K-Nearest Neighbors*) e KNN (*K-Nearest Neighbors*) revelam importantes informações sobre a eficácia e vantagens de utilizar a abordagem conjunta desses métodos para prever a classe correta em tarefas de classificação.

Uma das observações mais notáveis é o aumento significativo no ganho de desempenho quando compara-se o modelo LLM+KNN ao KNN. O valor máximo de ganho de precisão (max gain llm+knn >= knn) é de aproximadamente 53.33%, o que indica que, em muitos casos, o LLM+KNN é capaz de superar substancialmente o KNN na precisão das previsões. Além disso, a média desse ganho (mean gain llm >= knn) é cerca de 12.49%, demonstrando que essa melhoria não é um caso isolado, mas sim uma tendência consistente.

Outro ponto positivo a ser destacado é a quantidade significativamente maior de classes em que o LLM+KNN supera o KNN (count number class win llm+knn >= knn:

168). Isso significa que o modelo LLM+KNN é mais capaz de identificar corretamente a classe de exemplos de dados em uma variedade de cenários, o que é fundamental em aplicações práticas de classificação.

Quando avalia-se o desempenho do KNN em relação ao LLM+KNN, observa-se que, embora em algumas situações o KNN possa ser menos preciso, a diferença não é tão acentuada quanto no caso oposto. O valor máximo de perda de precisão (max lost llm+knn <= knn) é de -51.35%, o que indica que o KNN pode ter um desempenho significativamente pior em comparação com o LLM em algumas classes, mas a média dessa perda (mean lost llm+knn <= knn) é de cerca de -9.10%, sugerindo que essas situações são menos frequentes.

A classe "PublicTransitSystem", que teve o maior valor de perda (max lost llm+knn <= knn: -51.35%), parece ser uma das áreas onde o modelo LLM+KNN cometeu erros substanciais em relação ao KNN. Ao analisar alguns exemplos de textos classificados como "PublicTransitSystem"pelo modelo LLM+KNN e discutir minha opinião sobre os erros cometidos:

1. Texto: *"Krasnoyarsk Railway is a subsidiary of the Russian Railways headquartered in Krasnoyarsk and serving the south of Siberia."LLM-prediction: "RailwayLine"*

Opinião: O modelo LLM+KNN cometeu um erro ao classificar este texto como "RailwayLine"em vez de "PublicTransitSystem". Embora o texto mencione uma ferrovia, a ênfase está no fato de ser uma parte essencial do sistema de transporte público na região de Krasnoyarsk.

2. Texto: *"The South Eastern Railway (SER) is one of the seventeen railway zones in India and Part of Eastern Railways. It is headquartered at Garden Reach, Kolkata, West Bengal, India. It comprises four divisions..."LLM-prediction: "RailwayLine"*

Opinião: Novamente, o modelo LLM+KNN errou ao classificar este texto como "RailwayLine". O texto faz referência a uma importante zona ferroviária na Índia, mas essa informação está relacionada ao sistema de trânsito público, o que justificaria a classificação como "PublicTransitSystem".

3. Texto: *"The Louisville and Interurban Railroad (L&I) was an interurban line that operated in and around Louisville, Kentucky during the first half of the 20th century..."LLM-prediction: "RailwayLine"*

Opinião: Neste caso, o modelo LLM+KNN novamente errou ao classificar o texto como "RailwayLine"O texto descreve uma linha interurbana que serviu Louisville, Kentucky, e, portanto, se encaixaria melhor na classe "PublicTransitSystem".

Em todos esses exemplos, o modelo LLM+KNN parece ter se concentrado demasiadamente na presença de termos relacionados a ferrovias e linhas ferroviárias, ignorando o

contexto mais amplo em que essas informações estavam inseridas. Esses erros podem ser atribuídos à natureza complexa da tarefa de classificação textual e à necessidade de um contexto mais abrangente para a compreensão completa do significado dos textos.

Em geral, esses erros ressaltam a importância de um treinamento rigoroso e de um conjunto de dados diversificado para aprimorar a capacidade do modelo LLM+KNN de compreender contextos sutis e classificar corretamente textos em categorias específicas, como "PublicTransitSystem". É importante observar que, apesar desses erros, o modelo LLM+KNN ainda apresentou um desempenho superior em muitos casos, como indicado pelos resultados gerais.

Além disso, ao considerar a média da diferença entre ganho e perda (mean diference gain-lost), observa-se que, em média, a abordagem conjunta do LLM e KNN resulta em um ganho líquido de cerca de 3.39%. Isso significa que, em geral, a combinação desses modelos melhora o desempenho de classificação.

É importante mencionar que, em relação à métrica de acurácia, o modelo LLM+KNN também supera o KNN, com uma diferença positiva de aproximadamente 5.79% . A alta acurácia do LLM+KNN (accuracy llm+knn: 86.28%) em comparação com a do KNN (accuracy knn: 80.49%) destaca a capacidade do LLM+KNN de fazer previsões precisas e confiáveis.

Em resumo, os resultados destacam as vantagens do método de classificação que utiliza o LLM+KNN em conjunto com o KNN. Essa abordagem conjunta proporciona um ganho substancial de precisão em comparação com o uso isolado do KNN, tornando-a uma escolha atraente para tarefas de classificação onde a precisão é crucial. A combinação de habilidades de processamento linguístico do LLM+KNN com a capacidade do KNN de considerar vizinhos mais próximos resulta em um desempenho geral superior na tarefa de prever a classe correta.

3.4 Medição de Viés na Mídia usando Tópicos e Evolução Temporal

Desde o advento da prensa móvel, instrumento criado pelo alemão Johannes Gutenberg, a burguesia faz uso desta tecnologia como forma de propagação da sua ideologia, sendo o primeiro evento divisor de águas, a reforma protestante contra a ideologia propagada pela igreja católica, mostrando o poder de rompimento das ideias. Os primeiros a organizar uma obra fazendo uma crítica a esse sistema são Karl Marx e Frederick Engels em "A Ideologia Alemã".

Karl Marx e Friedrich Engels, ao desenvolverem suas teorias, destacaram a importância da ideologia na manutenção do sistema burguês. Eles argumentaram que a classe dominante usa a ideologia para legitimar e perpetuar seu poder, moldando a consciência das massas. A obra "A Ideologia Alemã"(MARX K. E ENGELS, 2007) aborda como a

ideologia da classe dominante permeia todos os aspectos da vida social, incluindo a cultura, a religião e a política.

Além disso, é relevante considerar o papel da ideologia na propagação dos meios de comunicação de massa, como rádio e televisão, pelo Estado, à luz dos conceitos gramscianos. Antonio Gramsci, que viveu durante a ditadura de Mussolini, desenvolveu a ideia de "hegemonia cultural", que se refere à dominação cultural exercida pela classe dominante. Ele argumentava que o Estado não apenas impõe sua vontade pela força, mas também através da cultura e da educação.

Um exemplo contemporâneo desse fenômeno pode ser observado na cobertura da operação Lava Jato, como documentado em artigos do The Intercept Brasil ((INTERCEPT, 2023a) e (INTERCEPT, 2023b)). A influência da ideologia é evidente na maneira como certos veículos de mídia retratam os eventos e atores envolvidos na operação, refletindo suas posições ideológicas e políticas.

Por fim, é importante compreender os movimentos dos veículos de mídia para entender como o Estado de Direito burguês suporta certas rupturas, como crimes e questões de gênero e etnia. O algoritmo proposto, que combina dois modelos de séries temporais: DTM (Dynamic Topic Modeling) para analisar todas as notícias relacionadas a um determinado tópico e, em seguida, aplica a matriz DTW (Dynamic Time Warping) para comparar a similaridade das linhas do tempo geradas pela DTM, pode ser um passo inicial para identificar o alinhamento ideológico mencionado por Gramsci e Marx. Isso permitiria uma análise mais aprofundada das influências ideológicas na mídia e em como elas afetam a percepção pública de questões cruciais.

3.4.1 Modelagem de Tópicos

Ao trabalhar com os termos atribuídos a cada texto jornalístico, realiza=se uma seleção no macro tópico do IPTC, como o exemplo 11000000, que se refere à categoria de política. Em seguida, aplica-se o treinamento de um modelo de tópicos, como o BERTopic. O BERTopic é uma técnica que nos permite realizar a modelagem de tópicos de forma dinâmica, calculando a representação de tópicos em cada intervalo de tempo, sem a necessidade de executar o modelo completo várias vezes. Para alcançar isso, inicia-se por ajustar o BERTopic como se não houvesse um aspecto temporal nos dados. Isso resulta na criação de um modelo de tópicos geral que captura a estrutura global dos principais tópicos, que podem variar ao longo de diferentes intervalos de tempo. Para cada tópico e intervalo de tempo, calcula-se a representação c-TF-IDF (Term Frequency-Inverse Document Frequency personalizado). Essa abordagem nos permite obter uma representação específica dos tópicos em cada intervalo de tempo sem a necessidade de criar agrupamentos a partir das incrustações, uma vez que eles já foram gerados previamente.

Além disso, no contexto da análise de tópicos dinâmicos, é importante mencionar

a utilização da equação matemática do DTM (Dynamic Topic Modeling), que permite modelar a evolução temporal dos tópicos nos textos jornalísticos.

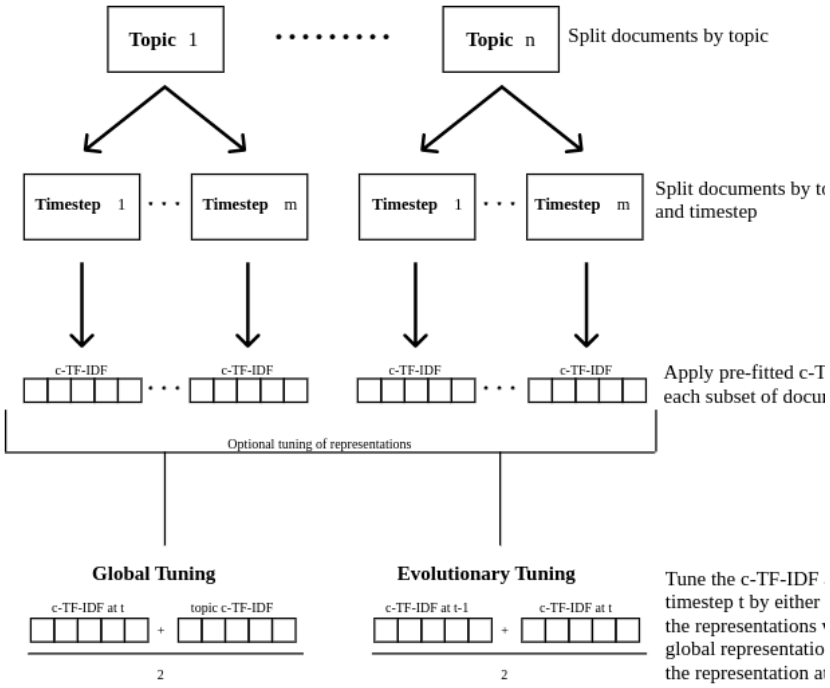


Figura 7 – BERTOPIC DTM - Imagem retirada da documentação da biblioteca

A obtenção das Dynamic Topic Models (DTMs) para os textos jornalísticos em um tópico específico é apenas o primeiro passo em uma análise abrangente e poderosa. Uma vez que temos essas DTMs, podemos explorar diversas dimensões e aplicar técnicas avançadas para extrair informações valiosas e insights significativos.

3.5 TBC-DTW

Nesta seção, discutiremos os resultados do modelo BERTopic-DTM, que produziu dois conjuntos de dados temporais representados por $df1$ e $df2$. O modelo BERTopic-DTM é uma abordagem avançada que combina tópicos gerados pelo BERTopic com modelagem de séries temporais (DTM - Dynamic Topic Modeling).

Os conjuntos $df1$ e $df2$ contêm informações sobre a dinâmica de tópicos em um contexto temporal. $df1$ e $df2$ podem ser interpretados como resultados da aplicação do modelo BERTopic-DTM em duas fontes diferentes de dados, como dois jornais, blogs ou sites.

A análise dos resultados inclui a aplicação de métricas de similaridade, como BLEU, para avaliar o quão semelhantes são os tópicos em $df1$ e $df2$. Além disso, é realizada a comparação das séries temporais geradas a partir dos tópicos para entender a evolução dos tópicos ao longo do tempo.

A partir dos conjuntos de dados $df1$ e $df2$, são calculadas métricas de distância, como a métrica DTW (Dynamic Time Warping), para avaliar a dissimilaridade entre as séries temporais. A métrica TBC-DTW, que combina a similaridade BLEU e a distância DTW, é aplicada para obter uma métrica global que quantifica a relação entre as duas fontes de dados.

A equação para a métrica TBC-DTW é definida como:

$$TBC{-}DTW(journal\_1, journal\_2) = \frac{1}{|T_1|} \sum_{t_1 \in T_1} \frac{1}{|T_2(n)|} \sum_{t_2 \in T_2(t_1)} D(F(df_1, t_1), F(df_2, t_2)) \cdot B(t_1, t_2, n) \tag{3.6}$$

Onde T_1 representa o conjunto de tópicos em $df1$, $T_2(t_1)$ representa o conjunto de tópicos em $df2$ selecionados com base nos top_n maiores valores de similaridade BLEU com t_1, e D é a métrica de distância DTW.

O algoritmo de alinhamento temporal dinâmico DTW é usado para comparar e alinhar sequências temporais que podem ter diferentes comprimentos, ritmos ou distorções. Ele encontra um mapeamento ótimo entre os pontos correspondentes das duas sequências, considerando a similaridade entre eles e minimizando a distância acumulada.

Neste pré-projeto, é explorado o uso do algoritmo de alinhamento temporal dinâmico DTW para medir o alinhamento dos tópicos do modelo dinâmico de tópicos DTM entre diferentes jornais, a fim de avaliar o grau de correspondência entre eles em tópicos específicos. Utilizando o poderoso algoritmo BERTopic, é possível extrair uma lista de palavras-chave relevantes juntamente com suas frequências ao longo do tempo.

Dada a hipótese que esteja sendo analisado o tópico de saúde em um jornal específico, como o Estadão. O algoritmo BERTopic nos fornecerá uma lista de palavras-chave que são frequentes nesse tópico ao longo do tempo. Em seguida, compara-se essa lista de palavras-chave com os tópicos relacionados à saúde em outros jornais. Ao realizar essa comparação pode-se calcular uma medida de similaridade ou alinhamento médio. Isso é feito analisando todas as matrizes de palavras-chave relacionadas ao tópico de saúde em diferentes períodos de tempo e calculando uma média ponderada das correspondências encontradas. O mesmo vale para qualquer tópico do IPTC.

As etapas anteriores forneceram os passos principais para a coleta de textos de várias mídias, pré-processamento e categorização em tópicos de interesse. Também foi apresentado o conceito de DTW (*Dynamic Time Warping*) para analisar séries temporais. Um diferencial do método proposto neste trabalho é adaptar o DTW para identificação de viés entre diferentes mídias sociais, aqui denominado TBC-DTW (*Topic-based Bias Comparison with Dynamic Time Warping*).

De forma geral, a proposta envolve comparar tópicos e sua evolução temporal a

partir de bases de notícias de duas mídias, com o uso do IPTC e BERTopic. Na série temporal de um tópico, cada observação representa a frequência das palavras do tópico em um intervalo de tempo específico.

Uma etapa importante do DTW é o cálculo da matriz de custo. No TBC-DTW a matriz de custo é calculada levando em consideração não apenas a distância entre os pontos das séries temporais, mas também a dissimilaridade das palavras de cada tópico em cada período de tempo. Obtida a matriz de custo, o processo de alinhamento tradicional do DTW é aplicado, em que é calculada a matriz de distâncias acumuladas e em seguida, o caminho de menor custo nesta matriz.

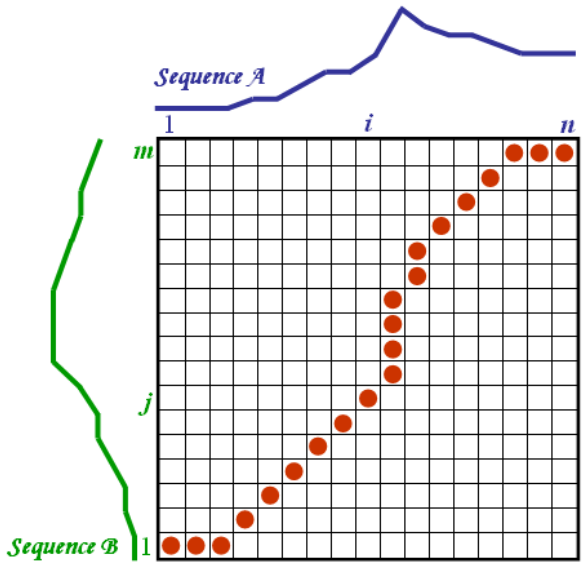


Figura 8 – Matriz DTW - Retirada de (SAKOE; CHIBA, 1978)

O valor obtido pelo TBC-DTW é uma medida de distância de alinhamento entre as séries temporais das mídias para os tópicos selecionados. Uma distância menor indica um viés mais semelhante entre as mídias, enquanto uma distância maior indica diferenças de viés.

As vantagens da proposta TBC-DTW envolvem tanto uma (1) técnica de comparação objetiva de viés, quanto a (2) tolerância a variações de tempo. Nesse último caso, é possível entender como as mídias reagem a eventos específicos, sem a necessidade de um alinhamento perfeito entre as duas mídias para compará-las.

Os resultados esperados pelo método proposto são, de certa forma, ambiciosos. A ideia é apresentar uma dos primeiros métodos objetivos para quantificar o quanto a análise da cobertura jornalística pode ser imparcial sobre as informações divulgadas, quando comparadas entre seus pares.

Essa estratégia também fornece ferramentas para avaliar se vozes e opiniões diversas

estão sendo representadas, assim visa promover um discurso mais inclusivo e plural. Por fim, o método é um avanço para apoiar tomada de decisões informadas, pois ao compreender o viés entre mídias, as pessoas podem tomar decisões informadas considerando como interpretar e ponderar as informações recebidas.

É importante ressaltar também que o uso do algoritmo TBC-DTW proporciona uma análise mais refinada e granular em comparação com abordagens tradicionais de análise de tópicos. Ao levar em consideração a evolução temporal dos tópicos e a correspondência entre eles, podemos obter visões mais detalhadas sobre como os jornais abordam e discutem questões ao longo do tempo. Essas informações são fundamentais para compreender a dinâmica da cobertura jornalística e a evolução das narrativas.

Portanto, a aplicação do algoritmo TBC-DTW na análise de alinhamento ideológico na cobertura jornalística é uma forma de examinar e compreender as complexas interações entre a mídia, a ideologia e a sociedade, proporcionando visões mais profundas sobre como as informações são apresentadas, interpretadas e influenciam a percepção pública, em consonância com as perspectivas de Althusser e Gramsci.

4 AVALIAÇÃO EXPERIMENTAL

Neste capítulo, é apresentado os resultados de nossa pesquisa, que teve como objetivo analisar o comportamento de diferentes algoritmos de clusterização e a semelhança entre eles nos macro tópicos do IPTC News Topics. Foram coletadas amostras de três jornais (O Globo, Folha de S.Paulo e Estadão) no período de 01 de janeiro de 2018 a 01 de junho de 2022. As amostras coletadas de cada jornal possuem os seguintes tamanhos:

Tabela 2 – Amostras de Documentos dos Jornais

Jornal	Tamanho de Amostra
O Globo	183,035
Folha de S.Paulo	128,396
Estadão	175,028

A tabela a seguir apresenta os dados coletados de várias fontes de notícias e suas categorias, juntamente com o número de ocorrências em cada categoria.

Fonte	**Categoria**	**Número de Ocorrências**
O Globo	Saúde	9827
Estadão	Saúde	10886
Folha-SP	Saúde	10232
O Globo	Meio Ambiente	2161
Estadão	Meio Ambiente	2009
Folha-SP	Meio Ambiente	1353
O Globo	Artes, Cultura, Entretenimento e Mídia	21837
Estadão	Artes, Cultura, Entretenimento e Mídia	16044
Folha-SP	Artes, Cultura, Entretenimento e Mídia	10977
O Globo	Ciência e Tecnologia	2231
Estadão	Ciência e Tecnologia	2333
Folha-SP	Ciência e Tecnologia	2228
O Globo	Conflito, Guerra e Paz	2990
Estadão	Conflito, Guerra e Paz	3182
Folha-SP	Conflito, Guerra e Paz	2705
O Globo	Crime, Lei e Justiça	16737
Estadão	Crime, Lei e Justiça	14628
Folha-SP	Crime, Lei e Justiça	12809
O Globo	Desastre, Acidente e Incidente de Emergência	2962
Estadão	Desastre, Acidente e Incidente de Emergência	3019
Folha-SP	Desastre, Acidente e Incidente de Emergência	2417
O Globo	Esportes	15590
Estadão	Esportes	14079
Folha-SP	Esportes	9080
O Globo	Economia, Negócios e Finanças	24790
Estadão	Economia, Negócios e Finanças	31717
Folha-SP	Economia, Negócios e Finanças	16900
O Globo	Educação	1630
Estadão	Educação	1550
Folha-SP	Educação	1116
O Globo	Estilo de Vida e Lazer	9071

Fonte	**Categoria**	**Número de Ocorrências**
Estadão	Estilo de Vida e Lazer	7597
Folha-SP	Estilo de Vida e Lazer	3598
O Globo	Interesse Humano	3219
Estadão	Interesse Humano	2290
Folha-SP	Interesse Humano	1746
O Globo	Clima	37
Estadão	Clima	49
Folha-SP	Clima	37
O Globo	Política	36481
Estadão	Política	38654
Folha-SP	Política	31929
O Globo	Religião	4827
Estadão	Religião	3174
Folha-SP	Religião	2741
O Globo	Sociedade	8104
Estadão	Sociedade	5789
Folha-SP	Sociedade	5818
O Globo	Trabalho	2784
Estadão	Trabalho	2790
Folha-SP	Trabalho	2071

A tabela apresenta a distribuição de notícias em diversas categorias, provenientes de três fontes de notícias distintas. Notavelmente, surgem variações consideráveis na quantidade de ocorrências em cada categoria, o que sinaliza divergências significativas nos enfoques e ênfases adotados por cada fonte de notícias. Essas diferenças podem ser atribuídas a uma série de fatores, como as abordagens jornalísticas adotadas, o público-alvo específico de cada veículo ou até mesmo as fontes de informação utilizadas.

Adicionalmente, merece destaque que a categoria "Política"é a mais recorrentemente abordada por todas as fontes de notícias, sendo seguida por "Economia, Negócios e Finanças", "Arte, Cultura, Entretenimento e Mídia", "Crime, lei e Justiça"e "Esportes", sendo "Saúde"atrás de todos os outros citados com destaque a Folha de São Paulo por "Saúde"estar a frente de "Esportes". Por outro lado, categorias como "Clima"e "Educação"apresentam um número consideravelmente menor de ocorrências, sugerindo uma menor cobertura jornalística nesses tópicos específicos.

Os testes foram feitos com diferentes parâmetros para dois modelos de agrupamento: UMAP e HDBSCAN. Os parâmetros dos testes foram os seguintes:

Tabela 4 – Parâmetros dos Testes

Teste	**UMAP**	**HDBSCAN**	**TOP_N**
1	n neighbors=5, n components=5, min samples=5	min cluster size=5, min samples=5	top_n=1
2	n neighbors=25, n components=25	min cluster size=25, min samples=25	top_n=1
3	n neighbors=50, n components=50	min cluster size=50, min samples=50	top_n=1

Para avaliar a similaridade entre os algoritmos, introduzimos a métrica TBC-DTW e a aplica-se a pares de jornais (journal 1, journal 2) em cada macro tópico do IPTC News Topics. As distribuições resultantes estão ilustradas nas Figuras 9, 10 e 11. É notável que essas distribuições exibem um padrão próximo ao de uma distribuição normal, sem a presença de valores atípicos. Essa uniformidade nas distribuições em cada um dos tópicos sugere uma homogeneidade na cobertura de notícias. Esse padrão pode ser um indício da existência de uma mídia hegemônica em relação aos tópicos analisados.

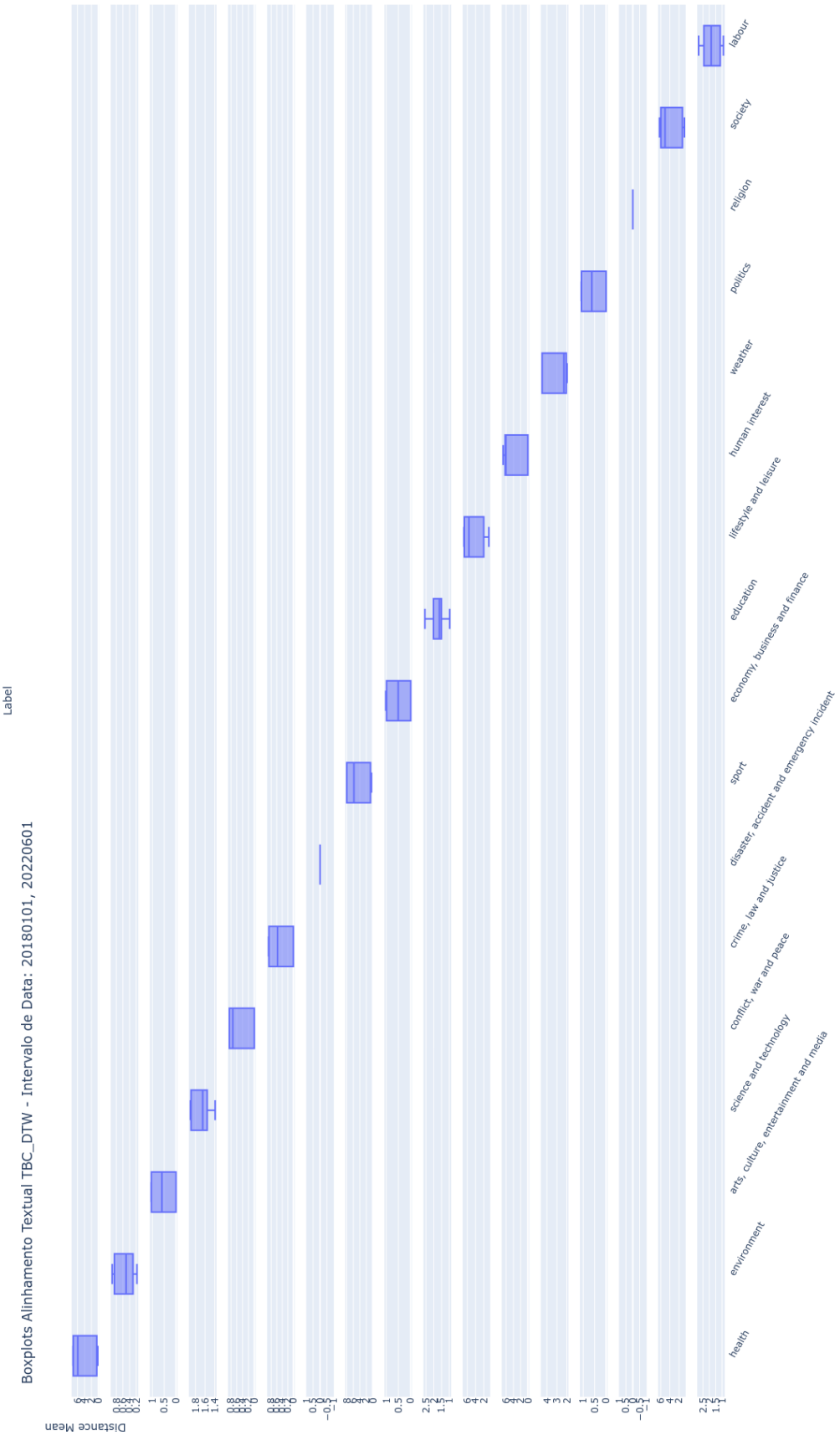


Figura 9 – Distribuição dos Tópicos entre os jornais Teste 1 - Elaborada pelo autor

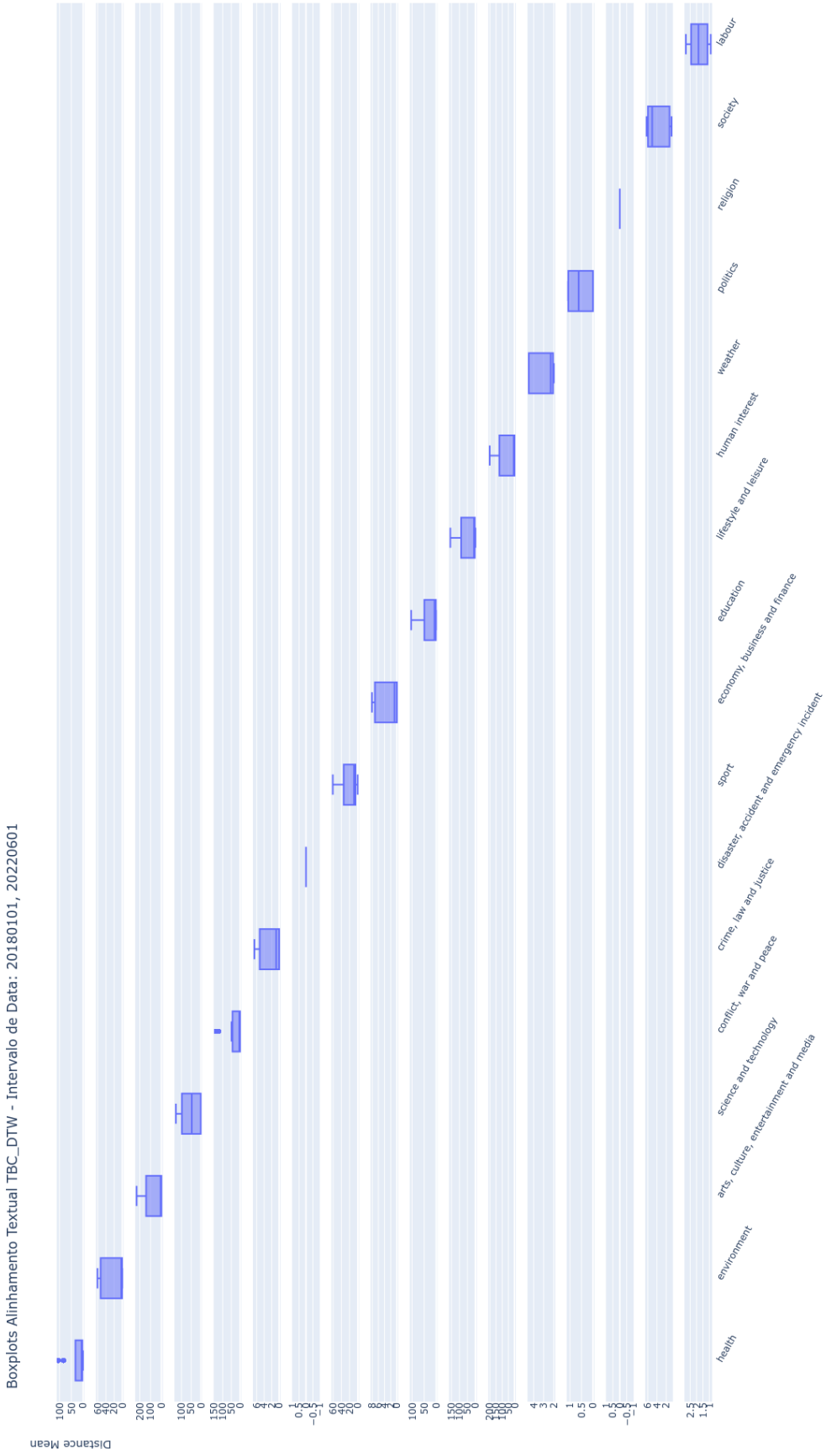


Figura 10 – Distribuição dos Tópicos entre os jornais Teste 2 - Elaborada pelo autor

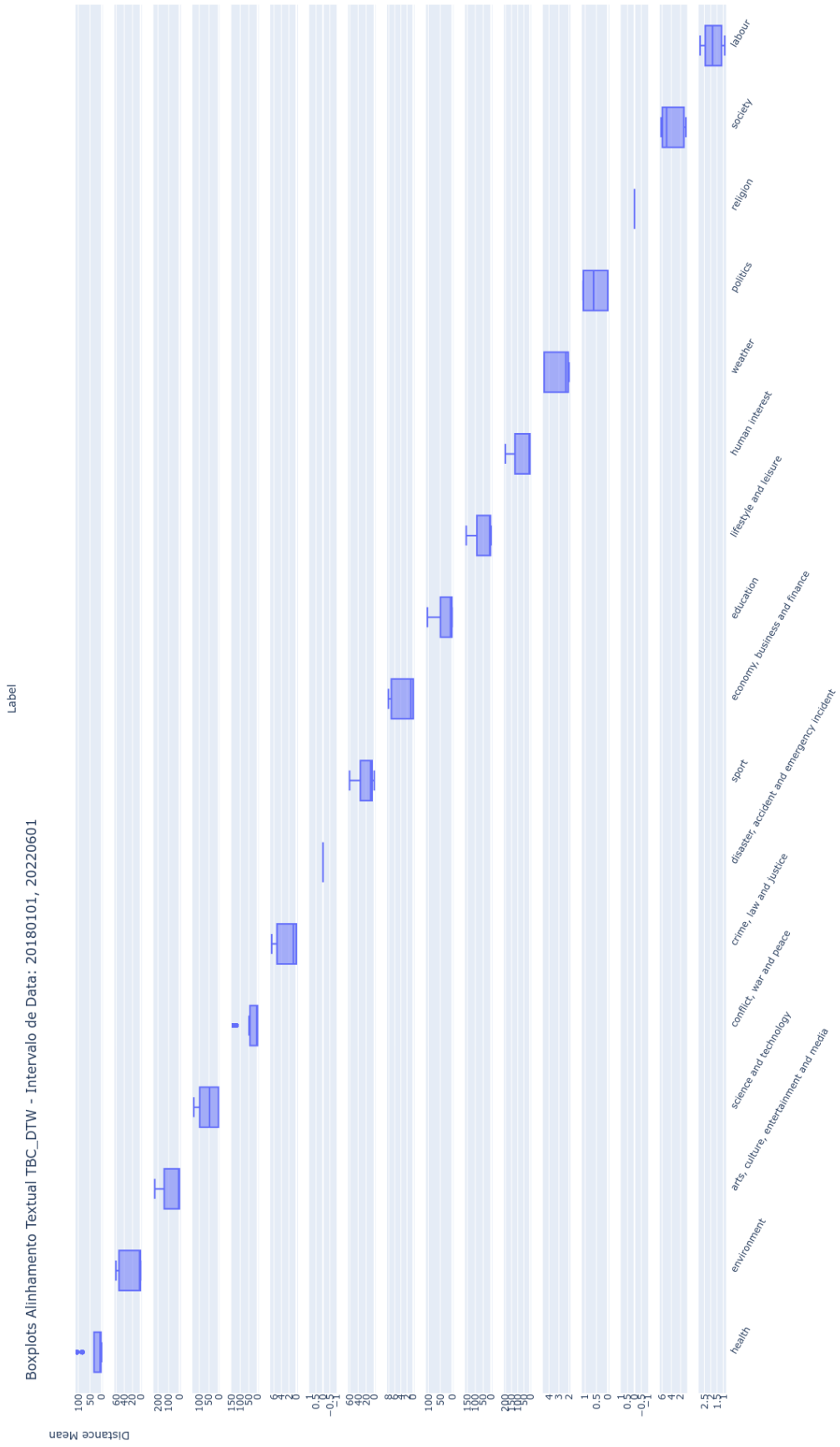


Figura 11 – Distribuição dos Tópicos entre os jornais Teste 3 - Elaborada pelo autor

Além disso, ao realizar análises das Dynamic Time Models (DTMs) com filtragem por palavra-chave e por tópico do IPTC macro tópicos, incluindo temas como "política - amazônia,política - lula,política - bolsonaro,política - jato,política - covid,"e "crime, lei e justiça - marielle."Pode-se notar que todas as curvas de DTMs apresentam comportamentos semelhantes, ressaltando a importância de uma mídia regional. Em particular, destaca-se o gráfico relacionado a "marielle,"que foi abordado exclusivamente pelo jornal O Globo.

Resultados do algoritmo Bertopic-DTM com os parâmetros de teste 3 da Tabela 4:

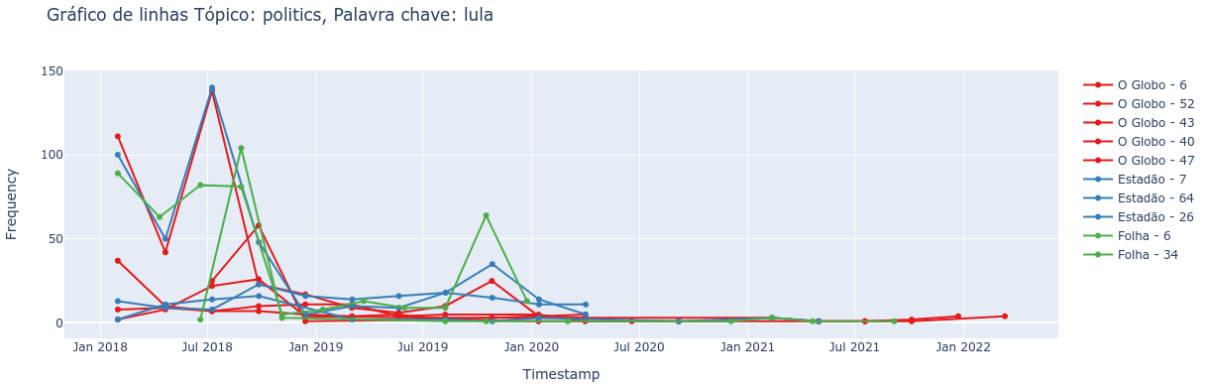


Figura 12 – Gráfico do termo pesquisado: lula - Elaborada pelo autor

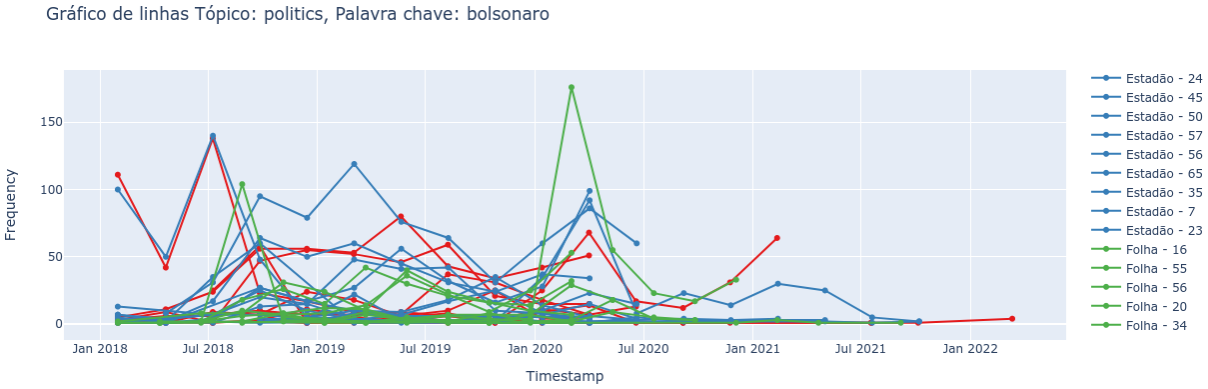


Figura 13 – Gráfico do termo pesquisado: bolsonaro - Elaborada pelo autor

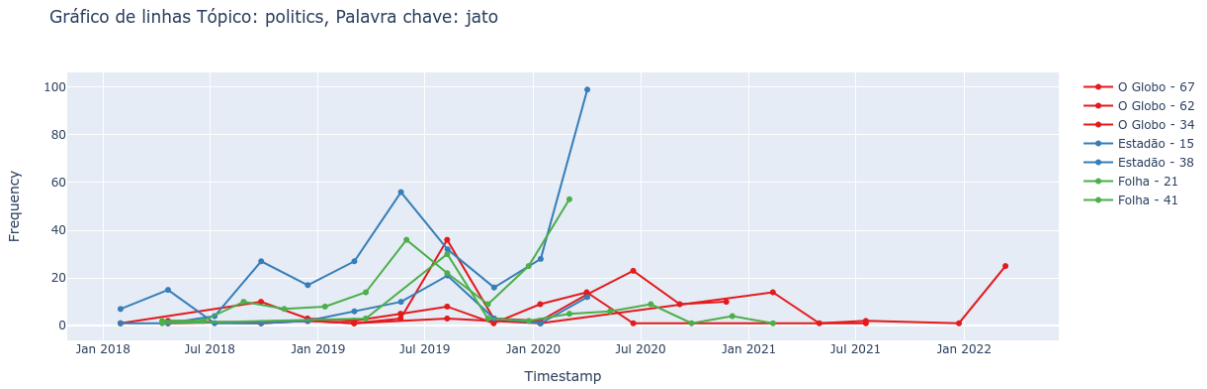


Figura 14 – Gráfico do termo pesquisado: jato - Elaborada pelo autor

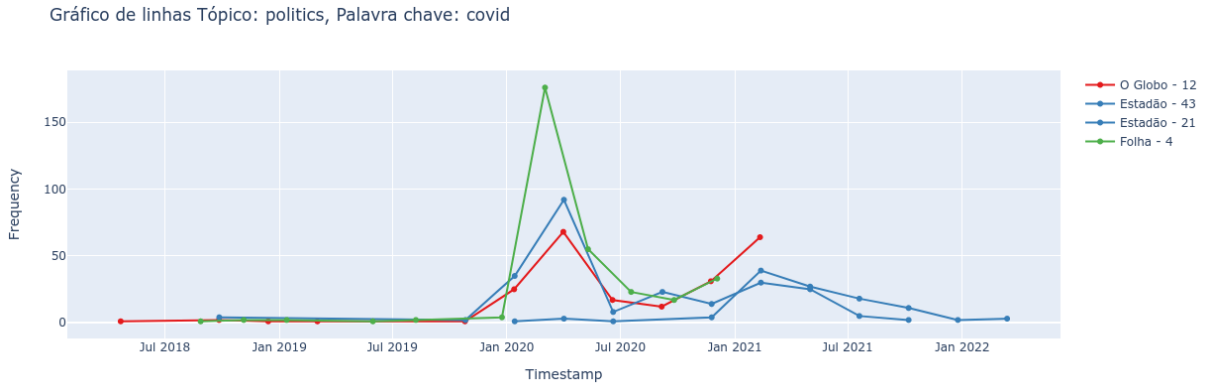


Figura 15 – Gráfico do termo pesquisado: covid - Elaborada pelo autor

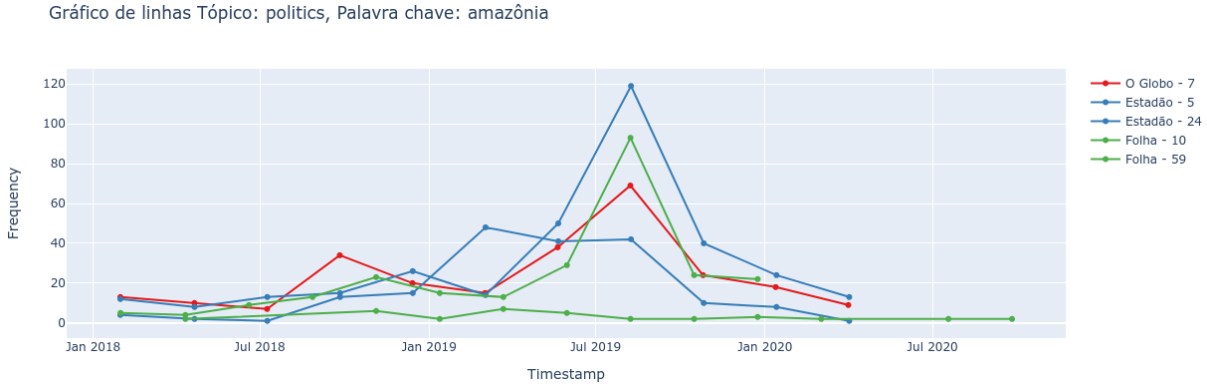


Figura 16 – Gráfico do termo pesquisado: amazônia - Elaborada pelo autor

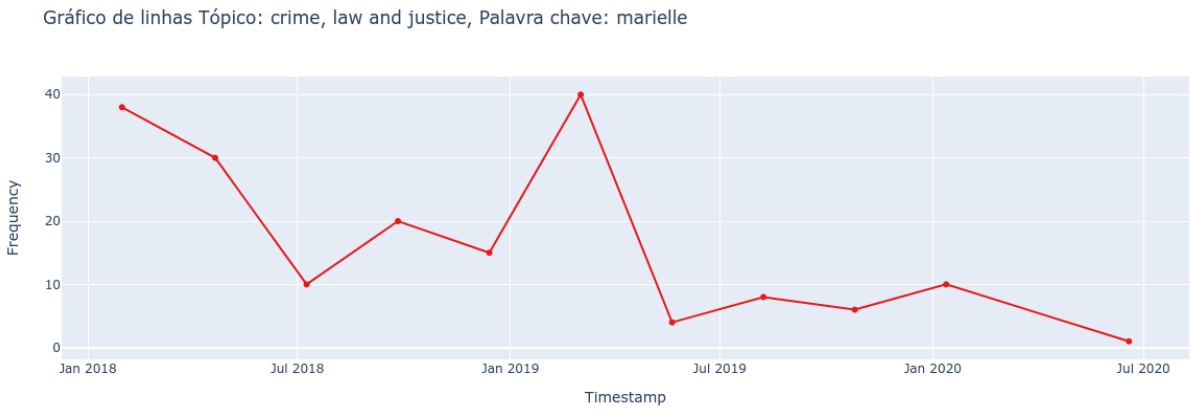


Figura 17 – Gráfico do termo pesquisado: Marielle - Elaborada pelo autor

Após a apresentação das DTMs, pode-se explorar as nuvens de palavras relacionadas ao tópico "covid", considerando a cobertura dos três jornais sobre a gestão do Ministério da Saúde sob ordens do ex-presidente Jair Messias Bolsonaro durante a pandemia e a CPI que investigou casos de superfaturamento na compra de vacinas.

Nuvem de palavras com os parâmetros de teste 3 e relativo a evolução temporal da Figura 15:

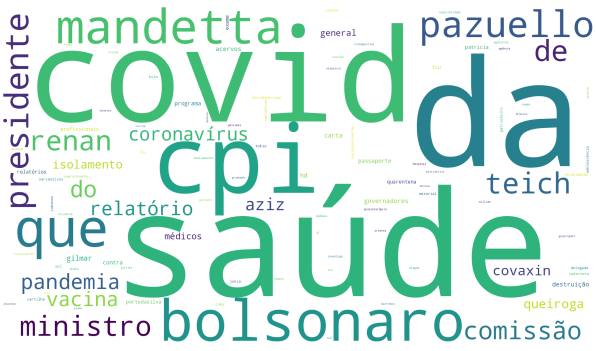


Figura 18 – Folha-SP - Nuvem de Palavras

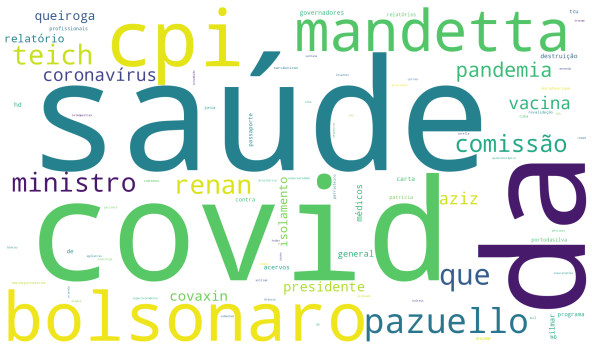


Figura 19 – Estadão - Nuvem de Palavras

Figura 20 – Estadão - Nuvem de Palavras

Figura 21 – O Globo - Nuvem de Palavras

Estes resultados fornecem análises valiosas sobre a semelhança entre diferentes algoritmos de clusterização e a abordagem de temas relevantes pelos jornais. Eles também destacam a importância da análise regional na cobertura de eventos específicos, como o caso "Marielle", e na intensificação na cobertura da pandemia de Covid-19.

5 CONSIDERAÇÕES FINAIS

Como conclusão final, é fundamental ressaltar que os resultados obtidos neste estudo fornecem uma visão valiosa da clusterização de notícias e da semelhança entre diferentes algoritmos, especialmente em jornais impressos como O Globo, Folha de S.Paulo e Estadão. No entanto, para uma compreensão completa da eficácia do algoritmo em diferentes formas de mídia, é necessário aprofundar a análise em diversos portais de notícias e tipos de mídia, como jornais televisivos, canais de notícias no YouTube, rádios, dentre outras fontes de informação.

A questão central que se coloca é se o algoritmo, que demonstrou ser eficaz na identificação da homogeneidade entre jornais impressos, manterá o mesmo desempenho quando aplicado a outros tipos de mídia, como o jornal Brasil de Fato e Jovem Pan. Para medir o impacto desse efeito heterogêneo, é importante considerar algumas estratégias:

1. **Coleta de Dados Ampliada:** Expansão da coleta de dados para incluir diversas fontes de mídia, abrangendo diferentes tipos de mídia, plataformas e formatos. Isso permitirá a comparação direta entre jornais impressos, jornais televisivos, canais de YouTube, rádios e outras formas de mídia.

2. **Desenvolvimento de Métricas Adicionais:** Desenvolvimento de métricas adicionais ou adaptação das métricas existentes para refletir as especificidades de cada tipo de mídia. Cada plataforma pode ter características únicas que afetam a forma como as notícias são apresentadas e agrupadas.

3. **Análise de Dados Comparativa:** Realização de análises comparativas entre os diferentes tipos de mídia para avaliar a homogeneidade ou heterogeneidade dos grupos. Isso pode envolver a aplicação do algoritmo em diferentes conjuntos de dados e a comparação dos resultados.

4. **Avaliação de Contexto e Conteúdo:** Além da análise de grupos, é importante avaliar o contexto e o conteúdo das notícias em diferentes tipos de mídia. Isso pode revelar como as notícias são abordadas, contextualizadas e apresentadas de maneira geral.

5. **Envolver Especialistas:** Envolver especialistas em comunicação e mídia para avaliar os resultados e fornecer análises qualitativas sobre o impacto do algoritmo em diferentes formas de mídia.

Medir o impacto desse efeito heterogêneo é essencial para compreender a generalização do algoritmo e sua utilidade em uma variedade de cenários de mídia. Essa pesquisa

mais ampla pode ajudar a adaptar e aprimorar o algoritmo, tornando-o uma ferramenta poderosa para a análise de notícias em diversos contextos midiáticos.

REFERÊNCIAS

ABDELRAZEK, A. *et al.* Topic modeling algorithms and applications: A survey. **Information Systems**, Elsevier, p. 102131, 2022.

AGGARWAL, C. C.; AGGARWAL, C. C. **Mining text data**. [*S.l.: s.n.*]: Springer, 2015.

ALTHUSSER, L.; BREWSTER, B. **Lenin and philosophy, and other essays**. [*S.l.: s.n.*]: Monthly Review Press, 2001.

BARBOSA, B. **Até quando a regulação da mídia será um tabu no Brasil?** 2022. <https://www.poder360.com.br/opiniao/ate-quando-a-regulacao-da-midia-sera-um-tabu-no-brasil/>. [Online; acessado 19-mar-2023].

BHARGAVA, P.; NG, V. Commonsense knowledge reasoning and generation with pre-trained language models: a survey. *In*: **Proceedings of the AAAI Conference on Artificial Intelligence**. [*S.l.: s.n.*], 2022. v. 36, n. 11, p. 12317–12325.

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

BRASIL, C. **Lula volta a defender regulação de rádio, TV e internet e reformar "direito de resposta"**. 2022. <https://www.cnnbrasil.com.br/politica/lula-volta-a-defender-regulacao-de-radio-tv-e-internet-e-reformar-direito-de-resposta/>. [Online; acessado 19-mar-2023].

BUDAK, C.; GOEL, S.; RAO, J. M. Fair and balanced? quantifying media bias through crowdsourced content analysis. **Public Opinion Quarterly**, Oxford University Press US, v. 80, n. S1, p. 250–271, 2016.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. *In*: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. [*S.l.: s.n.*], 2019. p. 4171–4186.

FÄRBER, M. *et al.* A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. *In*: **Proceedings of the 29th ACM International Conference on Information & Knowledge Management**. [*S.l.: s.n.*], 2020. p. 3007–3014.

FÉVOTTE, C.; IDIER, J. Algorithms for nonnegative matrix factorization with the β-divergence. **Neural computation**, MIT Press, v. 23, n. 9, p. 2421–2456, 2011.

GRAMSCI, A. **Selections from the prison notebooks of Antonio Gramsci**. [*S.l.: s.n.*]: Lawrence and Wishart, 1971.

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

GRUETZEMACHER, R.; PARADICE, D. Deep transfer learning & beyond: Transformer language models in information systems research. **ACM Computing Surveys (CSUR)**, ACM New York, NY, v. 54, n. 10s, p. 1–35, 2022.

GUO, X.; MA, W.; VOSOUGHI, S. Measuring media bias via masked language modeling. *In*: **Proceedings of the International AAAI Conference on Web and Social Media**. [*S.l.: s.n.*], 2022. v. 16, p. 1404–1408.

HALL, S. Signification, representation, ideology: Althusser and the post-structuralist debates. **Critical Studies in Mass Communication**, Informa UK Limited, v. 2, n. 2, p. 91–114, jun. 1985. Available at: <https://doi.org/10.1080/15295038509360070>.

HALLIN, D. C.; MANCINI, P. **A Comparing Media Systems: Three Models of Media and Politics**. [*S.l.: s.n.*]: Cambridge University Press, 2004.

HAMBORG, F.; DONNAY, K.; GIPP, B. Automated identification of media bias in news articles: an interdisciplinary literature review. **International Journal on Digital Libraries**, v. 20, p. 391–415, 2019.

INTERCEPT, T. **MERVAL PEREIRA MENTE NO GLOBO SOBRE A VAZA JATO. ESTA É NOSSA RESPOSTA.** 2023. <https://www.intercept.com.br/2023/08/03/merval-pereira-mente-no-globo-sobre-a-vaza-jato-esta-e-nossa-resposta/>. [Online; acessado 04-nov-2023].

INTERCEPT, T. **RICARDO NUNES JÁ GASTOU QUASE R**$3MILHÕESCOMPROPAGANDA$
[Online; acessado 04-nov-2023].

JR, W. P. E.; SHAH, D. V. The impact of individual and interpersonal factors on perceived news media bias. **Political psychology**, Wiley Online Library, v. 24, n. 1, p. 101–117, 2003.

KOHUT, A. *et al.* Press widely critisized, but trusted more than other information sources. **Pew Research Center**, 2011.

LAZARIDOU, K.; KRESTEL, R. Identifying political bias in news articles. **Bulletin of the IEEE TCDL**, 2016. Available at: <https://bulletin.jcdl.org/Bulletin/v12n2/papers/lazaridou.pdf>.

LEVITSKY, S.; ZIBLATT, D. Como as democracias morrem. *In*: . [*S.l.: s.n.*]: Zahar, 2018. p. 201.

LIM, S.; JATOWT, A.; YOSHIKAWA, M. Understanding characteristics of biased sentences in news articles. *In*: **CIKM workshops**. [*S.l.: s.n.*], 2018.

MARX K. E ENGELS, F. A ideologia alemã. *In*: . [*S.l.: s.n.*]: Boitempo Editorial, 2007. p. 32–44.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, August 27 2019. Available at: <https://arxiv.org/pdf/1908.10084.pdf>.

RIBEIRO, F. *et al.* Media bias monitor: Quantifying biases of social media news outlets at large-scale. **Proceedings of the International AAAI Conference on Web and Social Media**, Association for the Advancement of Artificial Intelligence (AAAI), v. 12, n. 1, jun. 2018. Available at: <https://doi.org/10.1609/icwsm.v12i1.15025>.

SAKOE, H.; CHIBA, S. Dynamic programming algorithm optimization for spoken word recognition. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 26, p. 43–49, 1978.

SIA, S.; DALMIA, A.; MIELKE, S. J. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *In*: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [*S.l.: s.n.*], 2020. p. 1728–1736.

SPINDE, T. *et al.* How can the perception of media bias in news articles be objectively measured? best practices and recommendations using user studies. *In*: **Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)**. [*S.l.: s.n.*], 2021.

THOMPSON, J. B. **Studies in the theory of ideology.** [*S.l.: s.n.*]: Polity Press, 1984.

VAYANSKY, I.; KUMAR, S. A. A review of topic modeling methods. **Information Systems**, Elsevier, v. 94, p. 101582, 2020.

VILAçA P. E MOURA, I. **Regulação da mídia x censura: um guia para não cair em pegadinhas**. 2022. <https://www.cartacapital.com.br/blogs/intervozes/regulacao-da-midia-x-censura>. [Online; acessado 19-mar-2023].