Uso de modelagem de tópicos para agrupamento de notícias: uma abordagem usando BERTopic

Pedro Henrique Pereira[1], Ticiana Linhares Coelho da Silva[1]

[1] Universidade Federal do Ceará (UFC) –Fortaleza, CE –Brasil

pedrohenripereira@gmail.com, ticianalc@insightlab.ufc.br

Abstract. *Today there is a large amount of news circulating in the media and grouping them by subjects or topics takes a lot of time. In this work, the topic modeling technique was used, through BERTopic, to group a set of news items under topics that reflect their subjects. The headlines of a set of news in English, from January to September 2022, were used. From the modeling of topics, analyzes were made and it was noticed that BERTopic can both represent the most reported topics throughout the year, as well as capture punctual events in a given period of the year. The modeling was also sensitive to variations in the context of the news.*

Resumo. *Hoje há uma grande quantidade de notícias em circulação nas mídias e agrupá-las por assuntos ou tópicos demanda muito tempo. Neste trabalho foi utilizada a técnica de modelagem de tópicos, por meio do BERTopic, para agrupar um conjunto de notícias sob tópicos que reflitam os assuntos delas. Foram utilizadas as headlines de um conjunto de notícias em inglês, de janeiro a setembro de 2022. A partir da modelagem de tópicos foram feitas análises e percebeu-se que o BERTopic tanto consegue representar os tópicos mais noticiados ao longo do ano, como também capturar eventos pontuais em um dado período do ano. A modelagem também se mostrou sensível a variações de contexto das notícias.*

1. Introdução

O advento da Web 2.0 evidenciou uma transformação nas formas de relacionamento e de interação da sociedade, influenciando também a forma como se veicula e se consome notícias. Diante da abundância de informações a que somos submetidos diariamente, muitas delas mediadas por algoritmos, torna-se essencial a identificação, a organização e a categorização dos assuntos [Arroyo-Vázquez 2014]. Esse tipo de tarefa repetitiva apresenta esforço proporcional à quantidade de dados a serem analisados e está sujeita à falha humana. Assim, surge a necessidade de automatizá-la, integral ou parcialmente, e um dos meios de fazer isso é com técnicas ligadas à área de Ciência de Dados.

Nesse contexto, uma abordagem para descobrir informações latentes em coleções de documentos é a modelagem de tópicos [Blei et al. 2003], onde um tópico representa um conjunto de palavras que descreve um assunto e os documentos são uma mistura de tópicos. Os tópicos são descobertos com base na co-ocorrência de palavras no conjunto de documentos.

Uma ferramenta que pode ser empregada nessa tarefa de modelagem de tópicos é o BERTopic, que engloba algoritmos para busca automática de tópicos representativos

em uma coleção de documentos, assumindo que documentos semanticamente semelhantes estejam em um mesmo tópico [Amorim et al. 2022].

Este trabalho busca avaliar a eficácia da ferramenta BERTopic para a tarefa de agrupar textos semanticamente afins, por meio da modelagem de tópicos. Assim, espera-se identificar, agrupados sob um mesmo tópico, notícias relacionadas por meio do assunto e também observar possíveis interações entre notícias de tópicos diferentes.

2. Fundamentação teórica

Neste trabalho foi usada a ferramenta BERTopic, desenvolvida por Maarten Grootendorst (2022), que a define como "uma ferramenta para a técnica de modelagem de tópicos que utiliza arquitetura de *transformers* e a medida c-TF-IDF para criar *clusters* densos, permitindo tópicos facilmente interpretáveis, mantendo palavras importantes nas descrições do tópico". O BERTopic prevê 3 etapas macro: (i) *embedding* de documentos, (ii) clusterização de documentos e (iii) representação de tópicos.

3. Metodologia

O presente trabalho busca avaliar o desempenho da ferramenta de modelagem de tópicos BERTopic sobre um conjunto de textos curtos oriundos dos *headlines* de notícias de jornais. O experimento para este trabalho foi pensando em 3 etapas: (i) a coleta e preparação dos dados, (ii) a modelagem dos tópicos com o BERTopic e (iii) a análise dos tópicos.

Após o estudo do referencial teórico e de alguns trabalhos relacionados foram elaboradas três perguntas de pesquisas, a saber: (i) os tópicos gerados se alinham com as categorias atribuídas previamente pelo autor do conjunto de dados?, (ii) a variação da janela temporal dos dados influencia na formação dos tópicos? e (iii) possíveis relações no mundo real entre as notícias podem ser percebidas nos tópicos formados?

Na última etapa da metodologia proposta, na seção de Resultados, os tópicos são analisados, a fim de responder ou reestruturar tais perguntas de pesquisa.

3.1. Coleta e preparação dos dados

Para o escopo deste trabalho, foi utilizado o conjunto de *headlines* de notícias coletadas por [Misra 2022], veiculadas ao longo do ano de 2022, no site HuffPost, sendo a última publicação com data de 23/09/2022, perfazendo um total de 1.426 notícias, em inglês, que foram usadas na íntegra, sem etapa de pré-processamento. Junto às *headlines* também estava informações sobre a qual categoria cada notícia estava associada no *site*. O maior texto de *headline* tinha 103 caracteres e 18 palavras e o menor apresentava 25 caracteres e 5 palavras. A escolha por essa janela temporal foi para tornar viável a análise comparativa manual com os fatos veiculados ao longo do ano de 2022.

Para analisarmos como o tamanho da janela temporal pode influenciar na modelagem dos tópicos, optou-se por aplicar a modelagem de tópicos em dois arranjos diferentes dos dados: (i) um primeiro arranjo (A1) contendo notícias dos noves primeiros meses do ano de 2022, isto é, de janeiro a setembro e (ii) um segundo arranjo

(A2), onde as notícias serão agrupadas por trimestre: A2-trim-1 (janeiro, fevereiro e março), A2-trim-2 (abril, maio e junho) e A2-trim-3 (julho, agosto e setembro).

3.2. Modelagem dos tópicos

Por uma questão de tempo e escopo deste trabalho, foram utilizados os valores padrões sugeridos pelo BERTopic. Após a criação do modelo, a modelagem é executada recebendo as *headlines* como entrada, organizados nos conjuntos anual (A1) e trimestrais (A2-trim-1, A2-trim-2 e A2-trim-3).

4. Resultados

4.1. Primeira Análise: Alinhamento entre tópicos e categorias

Os tópicos gerados apresentam relação com as categorias atribuídas a cada notícia pelo criador do conjunto de dados. Pode-se perceber isso ao fazer uma comparação com algumas amostras da saída do experimento.

Ao verificar o conteúdo e a categoria dos documentos mais representativos do tópico 1, por exemplo, percebe-se as notícias ligadas à guerra na Ucrânia rotuladas com categorias '*World News*' e '*Politics*', conforme Figura 1.

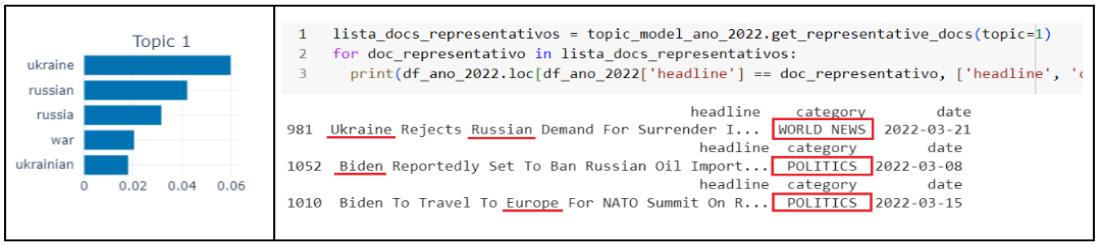

Figura 1. Conjunto de documentos mais representativos do tópico 1 da análise anual.

4.2. Segunda Análise: Comparação de tópicos em diferentes janelas temporais

Os assuntos que foram veiculados ao longo de todo o ano de 2022 geraram tópicos tanto quando se olha a modelagem anual (A1), quanto na modelagem trimestral (A2).

Isso pode ser notado nos gráficos de barra dos dois arranjos de dados, quando se observa o tópico ligado à Guerra da Ucrânia, que teve início em fevereiro de 2022 e perdura até o presente momento, junho de 2023. Na Figura 2, pode-se verificar a ocorrência desse tópico tanto quando se faz a modelagem sobre os dados de todo o ano de 2022 como quando se aplica a modelagem a cada trimestre de 2022.

Figura 2. Tópicos ligados à Guerra da Ucrânia, nas diversas janelas temporais

Percebe-se que o assunto central do tópico permanece tanto na janela anual quanto na janela trimestral. Mas as palavras que formam o tópico vão mudando conforme se muda a janela temporal. O termo Mariupol não aparece na janela anual, mas aparece na janela correspondente ao segundo trimestre de 2022, quando houve um ataque à cidade russa de Mariupol, em 08/05/2022.

4.3. Terceira Análise: Tópicos refletem outras notícias do mundo real

A visualização de documentos proposta pelo BERTopic auxilia a perceber as relações entre tópicos e como isso é um reflexo dos acontecimentos do mundo real.

Isso é notado na visualização de documentos, no trimestre 3, ao se analisar a proximidade entre os termos 'Biden' e 'Trump', na Figura 3. Em geral, são termos que aparecem nos mesmos tipos de notícias e até mesmo juntos em muitas dessas notícias, dada a rivalidade política entre os dois. No entanto, percebe-se o termo 'Biden' compondo o tópico ligado à Ucrânia e distante do tópico que traz o nome do 'Trump'.

Isso ocorreu porque, em setembro de 2022, o presidente norte-americano Joe Biden discursou na Assembleia Geral da ONU criticando fortemente o ataque russo à Ucrânia. Assim, percebe-se que para essa janela temporal apareceram mais notícias ligando o Biden à Ucrânia do que notícias relacionadas ao Trump.

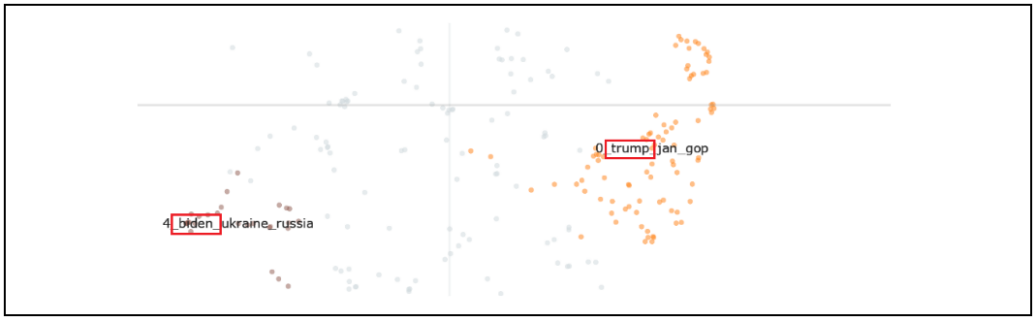

Figura 3. Visualização de documentos do trimestre 3 pelo BERTopic

5. Conclusões

Como conclusões deste trabalho, pode-se citar: (i) a metodologia proposta mostrou que o BERTopic consegue dar apoio ao processo de modelagem de tópicos, tanto na extração dos tópicos e na visualização dos resultados, de diferentes formas e (ii) os resultados mostraram que variações na janela temporal das notícias influenciaram diretamente o processo de modelagem de tópicos, uma vez que a janela temporal está ligada à quantidade de notícias de um dado assunto. Isso irá influenciar o cálculo de TF-IDF e, por conseguinte, dos tópicos gerados.

Uma contribuição é a análise na qual pode-se notar que os tópicos gerados conseguem ser mais específicos sobre os assuntos dos documentos quando comparados às categorias atribuídas pelo autor do conjunto de dados.

Um trabalho futuro seria avaliar o uso do BERTopic quando os tamanhos dos documentos apresentam grande variação de tamanhos entre si. Além do BERTopic, poderia-se também usar outras ferramentas que implementam algoritmos já difundidos em aplicações com modelagem de tópicos e *embeddings*, como, por exemplo, o Top2Vec [Angelov 2020].

Referências

Amorim, A; Murrugarra-Llerena, N.; Silva, V.; Oliveira, D.; Paes, A. (2022). "Modelagem de Tópicos em Textos Curtos: uma Avaliação Experimental". In: Anais do XXXVII Simpósio Brasileiro de Bancos de Dados. SBC, 2022. p. 254-266.

Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Arroyo-Vázquez, N. (2014). El content curator. Guía básica para el nuevo profesional de internet. Javier Guallar, Javier Leiva-Aguilera. Barcelona: Editorial UOC, 2013.(El profesional de la información: 24). ISBN 978-84-9064-018-0. Revista Española de Documentación Científica, v. 37, n. 2, p. e051-e051.

Blei, D. M.; NG, A.Y.; Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, v. 3, n. Jan, p. 993-1022.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Misra, R. (2022). News category dataset. arXiv preprint arXiv:2209.11429.