ALTES: uma Ferramenta de Rotulação Automática de Tópicos a partir de Fontes Externas*

Annie Amorim[1], Nils Murrugarra-Llerena[2], Vítor Silva, Daniel de Oliveira[1], Aline Paes[1]

[1]Instituto de Computação, Universidade Federal Fluminense (UFF), Niterói, RJ, Brasil

annieamorim@id.uff.br, {danielcmo,alinepaes}@ic.uff.br

[2]Weber State University, Ogden, Utah, Estados Unidos

nmurrugarrallerena@weber.edu

Abstract. *Interpreting the content of a large number of stored documents is challenging. Topic modeling is an unsupervised machine learning technique that supports this interpretation by identifying groups of words related to the same subject into sets of documents. However, interpreting the generated topics can be complex due to the lack of a straightforward semantic context in the grouped words. To address this challenge, the paper presents the ALTES labeling tool, which supports the interpretation of topics generated by the topic modeling technique through enrichment with data from external sources. ALTES finds words related to the terms that compose the topics and establishes associations between ideas or concepts that are not initially evident in the identified topics.*

Resumo. *Interpretar o conteúdo de uma grande quantidade de documentos é um desafio. A modelagem de tópicos é uma técnica não-supervisionada de Aprendizado de Máquina que apoia essa interpretação por meio do agrupamento de palavras relacionadas a um mesmo assunto em conjuntos de documentos. No entanto, a interpretação dos tópicos gerados pode ser complexa, uma vez que o contexto semântico que as une pode não estar evidente. Para enfrentar esse desafio, o artigo apresenta a ferramenta ALTES, que apoia a interpretação dos tópicos gerados pela técnica de modelagem de tópicos por meio da rotulação com dados de fontes externas. A ALTES encontra palavras relacionadas aos termos que compõem os tópicos e estabelece associações entre ideias ou conceitos não evidentes inicialmente nos tópicos identificados.*

1. Introdução

À medida que a quantidade de informações geradas por ferramentas e dispositivos digitais aumenta, a tarefa de buscar ou interpretar documentos que contenham uma certa informação se torna cada vez mais desafiadora [Allahyari et al. 2017]. Para facilitar a interpretação do conteúdo de tais documentos, foi proposta a técnica denominada Modelagem de Tópicos (MT) [Blei et al. 2010]. A MT, uma técnica de aprendizado de máquina não-supervisionado,

*Vídeo de demonstração da ferramenta: https://youtu.be/uS0aCqveJ9w. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores gostariam ainda de agradecer ao CNPq (*grant* 311898/2021-1) e FAPERJ (*grant* E-26/202.806/2019) pelo apoio financeiro.

apresenta-se como uma solução promissora, pois é capaz de identificar tópicos latentes em documentos por meio do agrupamento de palavras relacionadas a um mesmo contexto. Esses tópicos latentes geralmente podem ser entendidos como distribuições de probabilidades das palavras que ocorrem frequentemente juntas nos documentos. Por sua vez, os documentos podem ser interpretados como distribuições de probabilidades sobre tópicos. Uma prática comum para identificar cada tópico é representá-lo por meio das N palavras categorizadas como tendo maiores valores de probabilidade de pertencerem àquele tópico. Essas palavras ajudam a elucidar o sentido do tópico [Allahyari et al. 2017].

A interpretação dos tópicos, contudo, pode se mostrar complexa devido à ausência de contexto semântico das palavras agrupadas, já que, por vezes, o agrupamento de palavras pode não possuir uma associação com uma interpretação intuitiva. Nesse sentido, os *rótulos* atuam como etiquetas associadas a um conjunto de palavras que traduzem a semântica de cada tópico, tornando o assunto tratado mais compreensível e evidente [Lau et al. 2011]. A rotulação simplifica o entendimento e facilita a identificação do tópico em questão, eliminando a necessidade de interpretar uma lista de termos, uma vez que permite relacioná-los a um único conceito [Bhatia et al. 2016]. Na tarefa de rotulação de tópicos, é fundamental selecionar rótulos concisos, informativos e expressivos, que consigam representar de forma precisa o tema do tópico em questão. A consistência entre o conteúdo dos textos e os rótulos escolhidos é um fator crítico para evitar ambiguidades e imprecisões de interpretação. A rotulação de tópicos é uma técnica amplamente aplicável em diversas áreas, *e.g.*, análise de redes sociais, análise de conteúdo de mídias, análise de dados de pesquisas científicas e análise de *feedbacks* de usuários [Amorim et al. 2022]. Em cada uma dessas áreas, a rotulação de tópicos pode ajudar a identificar os temas mais relevantes e recorrentes, permitindo uma melhor compreensão e análise dos dados em questão e, consequentemente, aprimorando os processos e tomadas de decisão [Lau et al. 2011].

No entanto, a escolha de rótulos associados às palavras que compõem os tópicos é um processo complexo. Uma estratégia comum, apesar de conceitualmente simples, é empregar os N termos mais relevantes do tópico, organizados em uma lista, para assumirem o papel do rótulo. Porém, esses rótulos, também conhecidos como "saco de palavras", podem ser desafiadores para a interpretação [Kozbagarov et al. 2021]. Diversos tópicos requerem um conhecimento especializado e muitas vezes não intuitivo. Como exemplo, consideremos um conjunto de documentos relacionados a uma doença [Baratieri et al. 2021]. Muitos dos termos identificados podem ser de difícil entendimento para aqueles que não possuem familiaridade com o jargão médico. Adicionalmente, quando a tarefa de rotulagem é delegada a um indivíduo, ela fica sujeita à subjetividade, já que a interpretação das palavras possui uma carga cognitiva significativa, ocasionando na dificuldade de reprodutibilidade. Assim, a automatização do processo de rotulagem de tópicos torna-se um desafio de grande relevância, considerando a necessidade de interpretar semanticamente os tópicos para uma identificação mais precisa dos temas em análise.

Para abordar este desafio, este artigo apresenta a ferramenta denominada `ALTES` (*Automatic Labeling of Topics with External Sources*)[1], para rotulagem automática de tópicos que auxilia na interpretação de tópicos gerados por meio de técnicas de modelagem de tópicos tradicionais. A `ALTES` faz uso de fontes externas, *e.g.*, *Wikipedia*[2], para

[1] `https://github.com/UFFeScience/ALTES.git`
[2] `https://en.wikipedia.org/wiki/English_Wikipedia`

identificar palavras que possam estar relacionadas aos termos que compõem cada tópico. A estratégia de empregar fontes externas está intrinsecamente associada à oportunidade de ampliar a quantidade de recursos disponíveis para a criação de rótulos. Dessa forma, a ALTES estabelece conexões entre ideias ou conceitos que não são inicialmente evidentes nos tópicos identificados, aprimorando assim a interpretação destes. A ALTES executa essa função por meio da construção de rótulos tendo como base a similaridade entre os vetores de palavras coletadas de fontes externas e as palavras que se mostram mais relevantes dentro do tópico identificado. Este artigo de demonstração se encontra estruturado em três seções além desta Introdução. Na Seção 2, será apresentada a ferramenta ALTES. A Seção 3 será dedicada à discussão sobre a demonstração do funcionamento da ferramenta. Por fim, a Seção 4 conclui o artigo.

2. A Ferramenta ALTES

A ferramenta ALTES tem como finalidade gerar, de maneira automática, rótulos para cada tópico produzido pelas técnicas de modelagem de tópicos. Estes rótulos são criados a partir de palavras coletadas de fontes externas. Os rótulos gerados devem ser semanticamente relevantes para o tópico que representam e facilmente compreensíveis para o usuário, *i.e.*, cada rótulo deve possibilitar a identificação do assunto abordado em cada tópico de forma precisa e sem ambiguidades. Os rótulos são constituídos por combinações de palavras ou expressões que não necessariamente figuram entre os termos principais dos tópicos, ampliando assim a compreensão do tema abordado. O *workflow* de execução da ALTES é composto das seguintes macroatividades (Figura 1): (I) Coleta dos Dados, (II) Construção dos Candidatos, (III) Vetorização e (IV) Seleção dos Rótulos. Os rótulos identificados nesta tarefa são nomeados rótulos preditos dos tópicos.

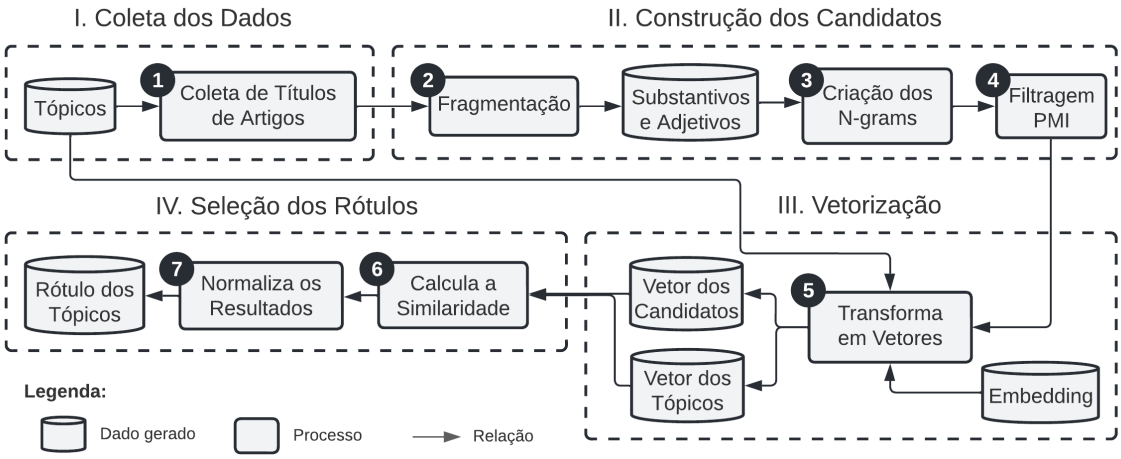

Figura 1. Rotulação automática dos tópicos por meio da ALTES.

Uma vez que um conjunto de tópicos foi gerado a partir de uma técnica de modelagem de tópicos, a macroatividade (I) envolve a coleta de títulos de artigos da *Wikipédia*. As N palavras mais relevantes (com as maiores probabilidades) de cada tópico são usadas na busca de artigos que as contenham tanto no título quanto no texto. No entanto, apenas são coletados os títulos de artigos que apresentem duas ou mais palavras que não estejam entre as palavras mais relevantes, garantindo que os tópicos não sejam rotulados com as mesmas palavras encontradas pela ferramenta de MT. Na macroatividade (II), são formados os

possíveis candidatos a rótulo para cada tópico. Este processo é feito a partir da sequência de K palavras (*n-grams*) compostas por substantivos e adjetivos extraídos dos títulos coletados. Uma observação fundamental é que, com o objetivo de assegurar que os rótulos adicionem nova informação aos tópicos gerados, as palavras N palavras mais relevantes utilizadas na busca são excluídas dos títulos coletados. Isso significa que as palavras importantes do tópico não são incluídas na criação dos potenciais candidatos a rótulos. Finalmente, são selecionados os *top* X rótulos candidatos cujo valor da PMI [Lau et al. 2011] é superior a um determinado limiar, calculado a partir das palavras no rótulo.

A macroatividade (III) consiste em vetorizar as palavras dos tópicos e dos candidatos a rótulo utilizando *embeddings* de palavras para comparação de similaridade. Para tanto, calcula-se a média dos vetores de cada palavra, para cada conjunto. Após vetorizar, a macro-atividade (IV) avalia a similaridade entre os candidatos a rótulo e as palavras mais relevantes do tópico. Essa avaliação é realizada por meio do cálculo normalizado da média ponderada das seguintes métricas: similaridade do cosseno, distância Euclidiana e distância Angular, com a similaridade do cosseno tendo um peso maior. Para escolher o melhor rótulo para um tópico específico, os candidatos a rótulo são comparados a cada uma das palavras no conjunto das palavras relevantes do tópico e o rótulo é selecionado com base na média ponderada das médias calculadas para cada palavra do tópico. O rótulo de cada tópico é definido conforme a sequência de palavras que apresenta maior similaridade com as palavras mais relevantes identificadas para o tópico. Todas as macroatividades da `ALTES` foram implementadas em Python e disponibilizadas em um *Jupyter Notebook* que pode ser acessado em `https://github.com/UFFeScience/ALTES`, de forma a facilitar o uso e a reprodutibilidade.

3. Demonstração

A demonstração da ferramenta `ALTES` é conduzida por meio de um estudo de caso que se baseia em um *dataset* de *tweets*, coletados nos dias 19 e 20 de fevereiro de 2022. Este *dataset*, o qual é composto de 26.863 *tweets*, foi usado como conjunto de treinamento do BERTopic, sendo uma técnica de modelagem de tópicos que utiliza transformadores e c-TF-IDF para criar *clusters* densos, permitindo a criação de tópicos potencialmente interpretáveis, ao mesmo tempo que mantém palavras importantes nas descrições dos tópicos. Para facilitar a compreensão e a visualização, um tópico específico é selecionado como exemplo no processo de rotulação automática. A Figura 2 ilustra as dez palavras mais relevantes do Tópico 4 identificado pelo BERTopic, utilizadas na busca de artigos na *Wikipedia*. Os títulos dos artigos que contêm as dez palavras mais relevantes são coletados por meio da API da *Wikipedia*[3] (conforme apresentado na Figura 3).

A partir das palavras obtidas nos títulos, formam-se os possíveis candidatos a rótulo. A similaridade de cada candidato com as palavras mais relevantes é avaliada, e aquele que apresenta a melhor média das métricas é definido como o rótulo do tópico. Para o Tópico 4, foram avaliados 36 candidatos, com os três principais sendo destacados na Figura 4. Assim, para o Tópico 4, o rótulo atribuído é "*Russo Ukrainian Nuclear*". A palavra "*Russo*" correlaciona-se diretamente com "*Russia*" e "*Russian*", e "*Ukrainian*" relaciona-se diretamente com "*Ukraine*". A palavra "*Nuclear*" pode ser interpretada como parte do contexto geopolítico mais amplo envolvendo a Rússia e a Ucrânia, especialmente considerando

[3] `https://en.wikipedia.org/w/api.php`

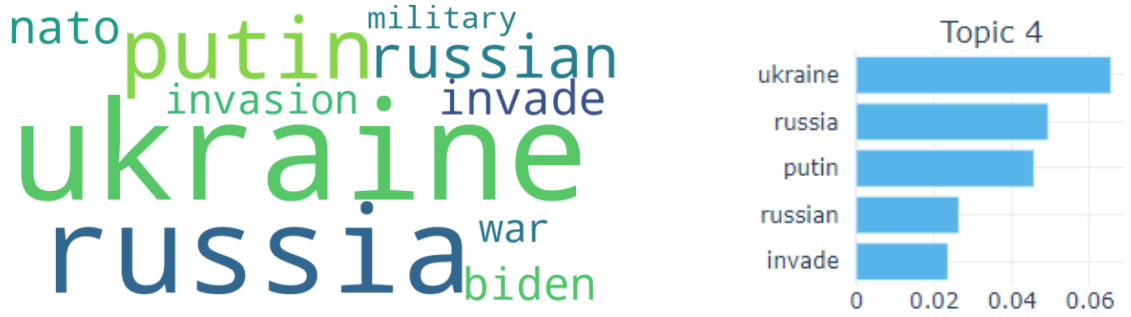

Figura 2. Tópico gerado pelo BERTopic

questões militares e de segurança, sugeridas pelas palavras "*Invade*", "*Nato*", "*Biden*", "*Invasion*", "*War*" e "*Military*". Portanto, em um contexto geral, o rótulo pode ser visto como relevante para as palavras que compõem o tópico. É importante destacar que a normalização também pode ser aplicada às palavras obtidas a partir dos títulos dos artigos, podendo resultar em termos mais distintos das palavras mais relevantes.

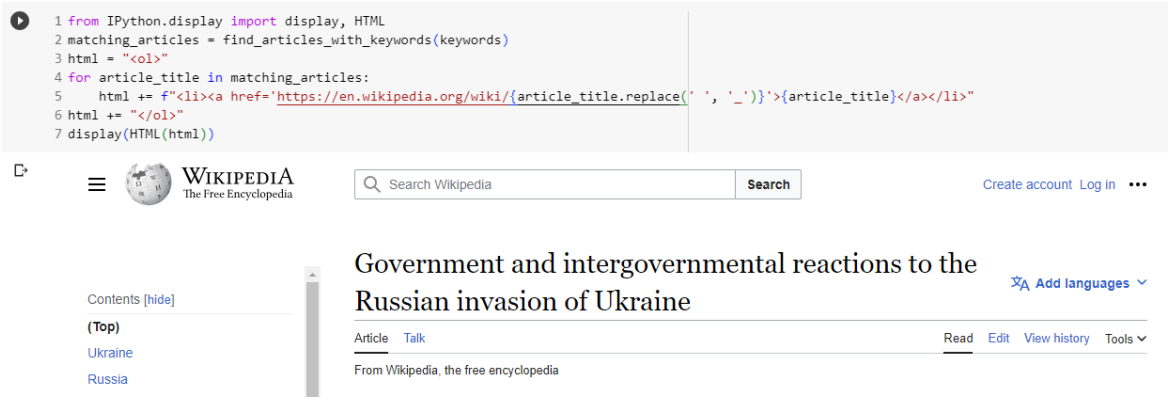

Figura 3. Exemplo de título de artigo coletado

É relevante frisar que, embora a ferramenta tenha sido exemplificada com entradas de textos curtos, a `ALTES` pode ser empregada em diversos contextos de modelagem de tópico. Ela não necessita de conexão com algum algoritmo de MT específico, apenas solicita como entrada os tópicos gerados. Adicionalmente, a ferramenta pode ser vinculada a outras fontes externas além da *Wikipedia*. Apesar da demonstração ter sido planejada utilizando *tweets* já coletados, os usuários são encorajados a trazer seus próprios documentos (sejam eles *tweets* ou não) para uso durante o SBBD.

	Label	Cosine	Angular	Euclidean	Mean
29	russo ukrainian nuclear	0.355065	0.617008	0.970621	0.884275
28	aid russo ukrainian	0.279502	0.590659	1.148971	0.859379
14	ukrainian nuclear threat	0.309477	0.600444	1.000609	0.843161

Figura 4. Os três principais candidatos à rótulo

4. Conclusões e Trabalhos Futuros

A ferramenta ALTES tem como objetivo automatizar a geração de rótulos de tópicos. A capacidade da ALTES de integrar-se a fontes de informação externas, como a *Wikipedia*, permite a geração de rótulos semanticamente ricos que auxiliam na interpretação dos tópicos identificados por algoritmos de MT. A ALTES marca um avanço na maneira de lidar com os desafios da interpretação de tópicos, oferecendo uma camada adicional de contextualização e compreensão. Esta perspectiva adicional é de grande valor para uma gama de aplicações, desde o estudo de interações em textos curtos de redes sociais até a análise de documentos longos. Com respeito aos trabalhos futuros, há planos para explorar mais fontes de dados externas para ampliar o leque de possíveis rótulos gerados pela ALTES. Há também o objetivo de expandir a ferramenta para outros idiomas. Em resumo, a ALTES representa um passo em direção a uma interpretação mais intuitiva dos tópicos gerados por meio da modelagem de tópicos.

As pesquisas mais recentes destacam o uso extensivo de modelos de linguagem, como, por exemplo, os LLMs (*Large Language Models*), para designar rótulos e compor descrições de tópicos gerados a partir da modelagem de tópicos [Praveen 2023]. No entanto, a implantação desses LLMs acarreta um consumo considerável de recursos computacionais. Em um cenário de menor escala, a ALTES se evidencia como uma alternativa aos LLMs, por não necessitar de um investimento tão elevado em termos de recursos computacionais.

Referências

Allahyari, M., Pouriyeh, S., Kochut, K. J., and Arabnia, H. R. (2017). A knowledge-based topic modeling approach for automatic topic labeling. *International Journal of Advanced Computer Science and Applications*, 8:335–349.

Amorim, A., Murrugarra-Llerena, N., Silva, V., de Oliveira, D., and Paes, A. (2022). Modelagem de tópicos em textos curtos: uma avaliaçao experimental. In *SBBD*, pages 254–266.

Baratieri, T., Lentsck, M. H., Peres, C. K., and de Brito Pitilin, É. (2021). Modelagem de tópicos de pesquisa sobre o novo coronavírus: aplicação do latent dirichlet allocation. *Ciência, Cuidado e Saúde*.

Bhatia, S., Lau, J. H., and Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. *CoRR*, abs/1612.05340.

Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65.

Kozbagarov, O., Mussabayev, R., and Mladenović, N. (2021). A new sentence-based interpretative topic modeling and automatic topic labeling. *Symmetry*, 13:837.

Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1536–1545, Portland, Oregon, USA.

Praveen, SV e Vajrobol, V. (2023). O chatgpt pode ser confiável para consultoria? desvendando as percepções do médico usando técnicas de aprendizagem profunda. *Anais de Engenharia Biomédica*, pages 1–4.