UNIVERSIDADE DO ESTADO DA BAHIA (UNEB)
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA, CAMPUS I
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

JEOSTON ARAUJO DA CRUZ JÚNIOR

UM PIPELINE COM BERTOPIC E GPT-4 PARA ANÁLISE DE PUBLICAÇÕES CIENTÍFICAS: UM EXPERIMENTO PRÁTICO APLICADO À PLATAFORMA DO SIMCC

SALVADOR
2024

JEOSTON ARAUJO DA CRUZ JÚNIOR

UM PIPELINE COM BERTOPIC E GPT-4 PARA ANÁLISE DE PUBLICAÇÕES CIENTÍFICAS: UM EXPERIMENTO PRÁTICO APLICADO À PLATAFORMA DO SIMCC

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do Departamento de Ciências Exatas e da Terra (DCET) - Campus I, da Universidade do Estado da Bahia (UNEB), como requisito à obtenção do grau de bacharel em Sistemas de Informação.
Área de concentração: Ciências da Computação.

Orientador: Prof. Dr. Eduardo Manuel de Freitas Jorge

SALVADOR
2024

TERMO DE ANUÊNCIA DO ORIENTADOR

Declaro para os devidos fins que li e revisei este trabalho e atesto sua qualidade como resultado final desta monografia. Confirmo que o referencial teórico apresentado é completo e suficiente para fundamentar os objetivos propostos e que a metodologia científica utilizada e os resultados finais são consistentes e com qualidade suficiente para submissão à banca examinadora final do Trabalho de Conclusão de Curso II do curso de Bacharelado em Sistemas de Informação.

Prof. Dr. Eduardo Manuel de Freitas Jorge
Orientador

JEOSTON ARAUJO DA CRUZ JÚNIOR

UM PIPELINE COM BERTOPIC E GPT-4 PARA ANÁLISE DE PUBLICAÇÕES CIENTÍFICAS: UM EXPERIMENTO PRÁTICO APLICADO À PLATAFORMA DO SIMCC

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do Departamento de Ciências Exatas e da Terra (DCET) - Campus I, da Universidade do Estado da Bahia (UNEB), como requisito à obtenção do grau de bacharel em Sistemas de Informação. **Área de concentração:** Ciências da Computação.

Aprovada em: .

BANCA EXAMINADORA

Prof. Dr. Eduardo Manuel de Freitas Jorge
Orientador

Prof. Dr. Nome Completo da Pessoa
Examinador interno (DCET-I/UNEB)

Prof. Dra. Nome Completo da Pessoa
Examinador interno (DCET-I/UNEB)

AGRADECIMENTOS

Dedico este trabalho à minha amada Vovó, Cenice, cuja alma resiliente enfrentou os desafios da vida com força e graça incomparáveis. Sua memória permanece viva em mim, provando que aqueles que amamos nunca partem verdadeiramente enquanto os mantemos em nossos corações (In memoriam).

À minha mãe, Elisandra, que me deu a vida e, mesmo com suas asas cortadas pela vida, nunca hesitou em me ajudar a voar. Sua coragem e sacrifício me ensinaram o verdadeiro significado do amor incondicional e da força silenciosa.

Ao meu amor, Beatriz, que nos dias de tempestade foi o meu sol, iluminando meus caminhos e aquecendo minha alma com esperança e carinho.

A presença dessas três mulheres eternas foi o que me deu forças para seguir em frente, mesmo nos momentos mais difíceis. Aprendi que o amor é um ato de vontade. Foi através desse amor que vocês me sustentaram, me inspiraram e me ensinaram a acreditar em mim mesmo. Este trabalho é tão meu quanto de vocês, pois cada conquista minha carrega o peso do incentivo e do sacrifício de cada uma de vocês. Obrigado por nunca desistirem de mim, mesmo quando eu quase desisti.

"Não importa o quanto a vida possa parecer difícil, há sempre algo que você pode fazer para ter sucesso."
(Stephen Hawking)

RESUMO

A análise de grandes volumes de publicações científicas apresenta desafios complexos, principalmente na organização e categorização de padrões temáticos. Em resposta a esse cenário, este estudo propõe o desenvolvimento de um pipeline que combina o BERTopic e o GPT-4 para a análise de publicações científicas na plataforma SIMCC. O BERTopic é empregado para a modelagem de tópicos através do uso de embeddings contextuais, da redução de dimensionalidade com UMAP e do agrupamento com HDBSCAN. Paralelamente, o GPT-4 é utilizado para enriquecer semanticamente os clusters de tópicos identificados, gerando rótulos descritivos e precisos que complementam a modelagem. A base de dados do projeto provém do SIMCC, uma plataforma da Secretaria Estadual de Ciência, Tecnologia e Inovação da Bahia que centraliza e organiza dados de produção acadêmica de profissionais vinculados a instituições de ensino e pesquisa do estado, integrando informações de diversas fontes como Currículos Lattes, Sucupira e OpenAlex. O sistema oferece funcionalidades para o gerenciamento do conhecimento acadêmico. A integração desse pipeline à base de dados do SIMCC visa facilitar a análise e a visualização das publicações por meio de um modelo de mapeamento visual, semelhante ao WizMap, que organiza os tópicos em clusters. Essa abordagem busca aprimorar a categorização temática, contribuindo para uma compreensão mais estruturada e detalhada do acervo científico disponível na plataforma.

Palavras-chave: Processamento de Linguagem Natural; Inteligência Artificial; Modelagem de Tópicos; BERTopic; GPT-4; SIMCC; Análise de Publicações Científicas.

ABSTRACT

The analysis of large volumes of scientific publications presents complex challenges, mainly in the organization and categorization of thematic patterns. In response to this scenario, this study proposes the development of a pipeline that combines BERTopic and GPT-4 for the analysis of scientific publications on the SIMCC platform. BERTopic is used for topic modeling through the use of contextual embeddings, dimensionality reduction with UMAP, and clustering with HDBSCAN. Simultaneously, GPT-4 is utilized to semantically enrich the identified topic clusters, generating descriptive and precise labels that complement the modeling. The project's database comes from SIMCC, a platform from the State Secretariat for Science, Technology, and Innovation of Bahia, which centralizes and organizes academic production data from professionals affiliated with teaching and research institutions in the state. The system integrates information from various sources such as Lattes Curricula, Sucupira, and OpenAlex. The integration of this pipeline into the SIMCC database aims to facilitate the analysis and visualization of publications through a visual mapping model, similar to WizMap, which organizes topics into clusters. This approach seeks to improve thematic categorization, contributing to a more structured and detailed understanding of the available scientific collection.

Key-words: Natural Language Processing; Artificial Intelligence; Topic Modeling; BERTopic; GPT-4; SIMCC; Scientific Publications Analysis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O cenário da pesquisa científica global tem testemunhado um crescimento exponencial nas últimas décadas, resultando em um vasto volume de dados que desafia os métodos tradicionais de organização e análise. Para navegar nessa imensidão de informações, pesquisadores confiam em plataformas de busca, como Web of Science, Scopus e IEEE Xplore, utilizando principalmente palavras-chave. Contudo, essa abordagem de recuperação de informações é limitada pela ambiguidade e pela diversidade do léxico científico, o que frequentemente resulta em buscas que não retornam a completude esperada e na dificuldade de identificar tendências emergentes na literatura (Galli *et al.*, 2024). A complexidade inerente a esses acervos de dados e a necessidade de uma análise mais profunda têm impulsionado o desenvolvimento e a aplicação de técnicas avançadas de Processamento de Linguagem Natural (PNL) (Datchanamoorthy; S; B, 2023).

Desta forma, o avanço das ferramentas associadas à ciência da informação, inteligência artificial e linguística computacional posiciona essas áreas como fundamentais na construção de soluções para a gestão do conhecimento acadêmico. Estudos como os de (Mohammadi; Karami, 2020), que analisaram tendências de pesquisa em big data por meio de mineração de texto, e (Xie *et al.*, 2020), que exploraram tópicos monolíngues e multilíngues com LDA e embeddings de *BERT*, destacam a relevância da integração de técnicas de modelagem de tópicos com modelos baseados em transformadores.

Os Transformadores introduzidos por (Vaswani *et al.*, 2017), com seu mecanismo de autoatenção (*self-attention*), revolucionaram o campo da PNL ao permitir que modelos como o BERT capturassem relações contextuais em textos com alta eficiência (Devlin *et al.*, 2019). A partir dessa base, surgiram os embeddings, representações numéricas que codificam o significado semântico de palavras e frases, superando as limitações de modelos tradicionais de bag-of-words e de modelagem de tópicos como o LDA (Galli *et al.*, 2024).

Nesse contexto, o BERTopic surge como uma abordagem moderna que se diferencia por utilizar os embeddings contextuais de modelos como o BERT para a modelagem de tópicos (Grootendorst, 2022). Esta técnica permite identificar tópicos de forma dinâmica e mais coesa, superando as deficiências de modelos tradicionais em capturar nuances semânticas e lidar com a complexidade de textos interdisciplinares.

Apesar do avanço, a aplicação de novas técnicas em larga escala e a adaptação a bases de dados complexas ainda enfrentam desafios em escalabilidade e adaptação (Datchanamoorthy; S; B, 2023). Estudos como o de (Dillan; Fudholi, 2023) mostram que a integração de transformadores com sistemas baseados em *LLMs* (Large Language Models) como o GPT-4 pode melhorar a geração de *embeddings* e a rotulagem de tópicos. No

entanto, desafios como o ajuste de hiperparâmetros para maximizar a granularidade dos tópicos e a dependência de grandes volumes de dados rotulados permanecem obstáculos importantes (Weng; Wu; Dyer, 2022).

Este projeto de pesquisa aborda esse problema por meio do desenvolvimento de um pipeline que combina o BERTopic (Grootendorst, 2022) e o GPT-4 para a análise de publicações científicas aplicado à plataforma do SIMCC, uma ferramenta que integra e centraliza dados de produção acadêmica de profissionais vinculados a instituições de ensino e pesquisa da Bahia. A plataforma, que coleta informações de fontes como Currículos Lattes, Sucupira e OpenAlex, tem um papel fundamental na gestão do conhecimento científico regional. O BERTopic é empregado para realizar a modelagem de tópicos a partir da base de dados, utilizando embeddings contextuais para identificar padrões temáticos de forma robusta e coerente. Complementarmente, o GPT-4 é utilizado para enriquecer a rotulagem dos tópicos, gerando rótulos descritivos e precisos que facilitam a interpretação dos resultados.

A metodologia (DSR) é adotada como a estrutura principal deste estudo, orientando a criação de um artefato que visa resolver um problema prático por meio de uma solução híbrida. A pesquisa é estruturada em dois ciclos complementares: no primeiro, é desenvolvido o pipeline de modelagem de tópicos que combina o *BERTopic* e o *GPT-4* para extrair padrões temáticos da base de dados do *SIMCC* e gerar representações semânticas enriquecidas. No segundo ciclo, o pipeline é integrado à plataforma, permitindo a identificação de temas emergentes de forma visual através de um *WizMap*, um visualizador interativo e escalável para explorar grandes incorporações de aprendizado de máquina (Wang; Hohman; Chau, 2023). Este mecanismo otimiza a experiência dos usuários e contribui para a gestão estratégica da pesquisa na plataforma.

REFERÊNCIAS

DATCHANAMOORTHY, K.; S, A. M. G.; B, P. Text mining: Clustering using bert and probabilistic topic modeling. **Social Informatics Journal**, 2023. Disponível em: https://api.semanticscholar.org/CorpusID:267122800. Citado na página 12.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: https://arxiv.org/abs/1810.04805. Citado na página 12.

DILLAN, T.; FUDHOLI, D. H. Ldaviewer: An automatic language-agnostic system for discovering state-of-the-art topics in research using topic modeling, bidirectional encoder representations from transformers, and entity linking. **IEEE Access**, v. 11, p. 59142–59163, 2023. Citado na página 12.

GALLI, C. *et al.* Topic modeling for faster literature screening using transformer-based embeddings. **Metrics**, v. 1, n. 1, 2024. ISSN 3042-5042. Disponível em: https://www.mdpi.com/3042-5042/1/1/2. Citado na página 12.

GROOTENDORST, M. **BERTopic: Neural topic modeling with a class-based TF-IDF procedure**. 2022. Disponível em: https://arxiv.org/abs/2203.05794. Citado nas páginas 12 e 13.

MOHAMMADI, E.; KARAMI, A. Exploring research trends in big data across disciplines: A text mining analysis. **Journal of Information Science**, v. 48, 06 2020. Citado na página 12.

VASWANI, A. *et al.* **Attention Is All You Need**. 2017. Disponível em: https://arxiv.org/abs/1706.03762. Citado na página 12.

WANG, Z. J.; HOHMAN, F.; CHAU, D. H. **WizMap: Scalable Interactive Visualization for Exploring Large Machine Learning Embeddings**. 2023. Disponível em: https://arxiv.org/abs/2306.09328. Citado na página 13.

WENG, M.-H.; WU, S.; DYER, M. Identification and visualization of key topics in scientific publications with transformer-based language models and document clustering methods. **Applied Sciences**, v. 12, n. 21, 2022. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/12/21/11220. Citado na página 13.

XIE, Q. *et al.* Monolingual and multilingual topic analysis using lda and bert embeddings. **Journal of Informetrics**, v. 14, n. 3, p. 101055, 2020. ISSN 1751-1577. Disponível em: https://www.sciencedirect.com/science/article/pii/S1751157719305127. Citado na página 12.