



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 (UNEBA), 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

| | |
|--|----|
| Figura 1 – Transformador - modelo arquitetural. | 16 |
| Figura 2 – Diagrama esquemático detalhado da comparação de métricas de avaliação entre modelos. | 20 |
| Figura 3 – Diagrama ilustrativo do UMAP, que demonstra a relação entre os hiperparâmetros $n_neighbors$ e min_dist e a representação visual dos dados. O parâmetro $n_neighbors$ controla a balança entre a preservação da estrutura global (valores altos) e local (valores baixos), enquanto min_dist ajusta a densidade dos agrupamentos, determinando a proximidade entre os pontos no espaço de baixa dimensionalidade. Esta visualização é crucial para otimizar o algoritmo e garantir que a estrutura semântica das publicações científicas seja fielmente representada para a subsequente etapa de agrupamento. | 21 |
| Figura 4 – Figura ilustrativa de um <i>dataset</i> sintético com quatro <i>clusters</i> e ruído de fundo. A imagem demonstra o tipo de desafio que o algoritmo HDBSCAN é capaz de superar, como a identificação de agrupamentos de densidades e formas variadas, além de tratar <i>outliers</i> de forma eficiente. Este comportamento é ideal para a análise de publicações científicas, onde a distribuição dos tópicos tende a ser heterogênea e não segue padrões geométricos rígidos. | 22 |
| Figura 5 – Adaptação da Design Science Research para este projeto. | 32 |
| Figura 6 – Diagrama esquemático detalhado da ferramenta Wizmap para visualização de dados. | 42 |

LISTA DE TABELAS

Tabela 1 – Quadro Resumo: Comparativo de Trabalhos Correlatos 29

LISTA DE QUADROS

SUMÁRIO

| | | |
|-------|--|----|
| 1 | INTRODUÇÃO | 12 |
| 2 | REFERENCIAL TEÓRICO | 14 |
| 2.1 | Ciência da Informação e Análise de Publicações Científicas . . | 14 |
| 2.2 | Transformadores e Embeddings no Contexto do PLN | 15 |
| 2.3 | Abordagens Tradicionais de Modelagem de Tópicos | 17 |
| 2.4 | BERTopic: Uma Abordagem Moderna | 19 |
| 2.5 | Modelos de Linguagem de Grande Escala (LLMs) | 22 |
| 3 | TRABALHOS CORRELATOS | 26 |
| 3.1 | Síntese Comparativa dos Trabalhos Correlatos | 28 |
| 4 | METODOLOGIA | 30 |
| 4.1 | Identificação do Problema e Definição de Objetivos | 30 |
| 4.2 | Desenvolvimento do Artefato | 30 |
| 4.3 | Avaliação e Validação | 31 |
| 4.4 | Apresentação dos Resultados e Comunicação | 31 |
| 5 | PROJETO DE DESENVOLVIMENTO | 33 |
| 5.1 | Tecnologias Utilizadas | 34 |
| 5.1.1 | <i>Base Tecnológica da Plataforma SIMCC</i> | 34 |
| 5.1.2 | <i>Pipeline de Modelagem e Análise de Tópicos</i> | 34 |
| 5.2 | Arquitetura da solução | 35 |
| 5.2.1 | <i>Coleta e Pré-processamento dos Dados</i> | 35 |
| 5.2.2 | <i>Geração de Embeddings Contextuais</i> | 36 |
| 5.2.3 | <i>Redução de Dimensionalidade</i> | 37 |
| 5.2.4 | <i>Agrupamento de Tópicos (Clustering)</i> | 38 |
| 5.2.5 | <i>Rotulagem de Tópicos</i> | 40 |
| 5.2.6 | <i>Visualização Interativa (WizMap)</i> | 41 |
| 6 | RESULTADOS ESPERADOS | 43 |
| | REFERÊNCIAS | 45 |

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 (PLN) (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 PLN 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.

2 REFERENCIAL TEÓRICO

O referencial teórico deste estudo abordará diversos aspectos cruciais relacionados à Ciência da Informação, Análise de Publicações Científicas, Processamento de Linguagem Natural (PLN), Modelagem de Tópicos, e Modelos de Linguagem de Grande Escala (LLMs).

2.1 Ciência da Informação e Análise de Publicações Científicas

A explosão da produção científica global nas últimas décadas, impulsionada pela maior acessibilidade à tecnologia e pela colaboração interdisciplinar, delineia um cenário desafiador para a área da Ciência da Informação. Como destacam Kim, Kogler e Maliphol (2024), o volume crescente de publicações dificulta a atualização contínua de pesquisadores e a identificação de áreas emergentes do conhecimento. Nesse contexto, estratégias tradicionais de busca baseadas em palavras-chave mostram-se limitadas, uma vez que desconsideram a complexidade semântica do léxico científico. Esse fator resulta não apenas na omissão de trabalhos relevantes, mas também na dificuldade de mapear de forma consistente o progresso em determinados campos.

Um aspecto que amplia essa complexidade é a diversidade linguística no ambiente científico. Segundo Xie *et al.* (2020), embora o inglês desempenhe papel predominante na comunicação acadêmica, uma parcela significativa da produção ocorre em outros idiomas. Os autores argumentam que metodologias convencionais baseadas em citações revelam-se insuficientes para a análise multilíngue, visto que publicações em inglês raramente fazem referência a pesquisas em outras línguas. Essa limitação restringe a circulação global do conhecimento, reduzindo a visibilidade e o impacto de estudos relevantes. Plataformas de indexação consolidadas, como Scopus e Web of Science, tendem a privilegiar artigos publicados em inglês, contribuindo para a sub-representação de pesquisas em outros idiomas. Além disso, abordagens tradicionais de mineração de dados e categorização apresentam dificuldades em alinhar conceitos e terminologias em diferentes línguas, o que frequentemente resulta em tópicos fragmentados e de menor valor analítico.

A maioria dos estudos até agora sobre análise de tópicos tem sido baseada em publicações em inglês e tem dependido fortemente da análise de evolução de tópicos baseada em citações (Xie *et al.* (2020, Traduzido)).

Diante desse cenário, técnicas contemporâneas de *Topic Modeling*, em especial aquelas fundamentadas em *embeddings*, têm sido investigadas como alternativas promissoras. De acordo com Galli *et al.* (2024), a utilização de representações densas derivadas de modelos como o BERT potencializa a análise de grandes volumes textuais, permitindo

capturar aspectos semânticos que vão além da simples coincidência lexical. Essa capacidade favorece a identificação de padrões temáticos em documentos que não compartilham necessariamente o mesmo vocabulário. Nesse sentido, métodos como o *BERTopic*, que constituem a primeira etapa do pipeline proposto neste trabalho, oferecem uma estrutura metodológica adequada para a extração de tópicos a partir de representações vetoriais densas dos textos científicos heterogêneos.

A aplicação dessas ferramentas em plataformas como o SIMCC, que contêm publicações em diversos idiomas — com destaque para o Português —, torna-se particularmente relevante. O pipeline delineado nesta pesquisa propõe uma abordagem híbrida que busca não apenas organizar o conhecimento de maneira mais sistemática, mas também contribuir para uma análise mais equitativa da produção científica, valorizando trabalhos independentemente do idioma em que foram originalmente publicados.

2.2 Transformadores e Embeddings no Contexto do PLN

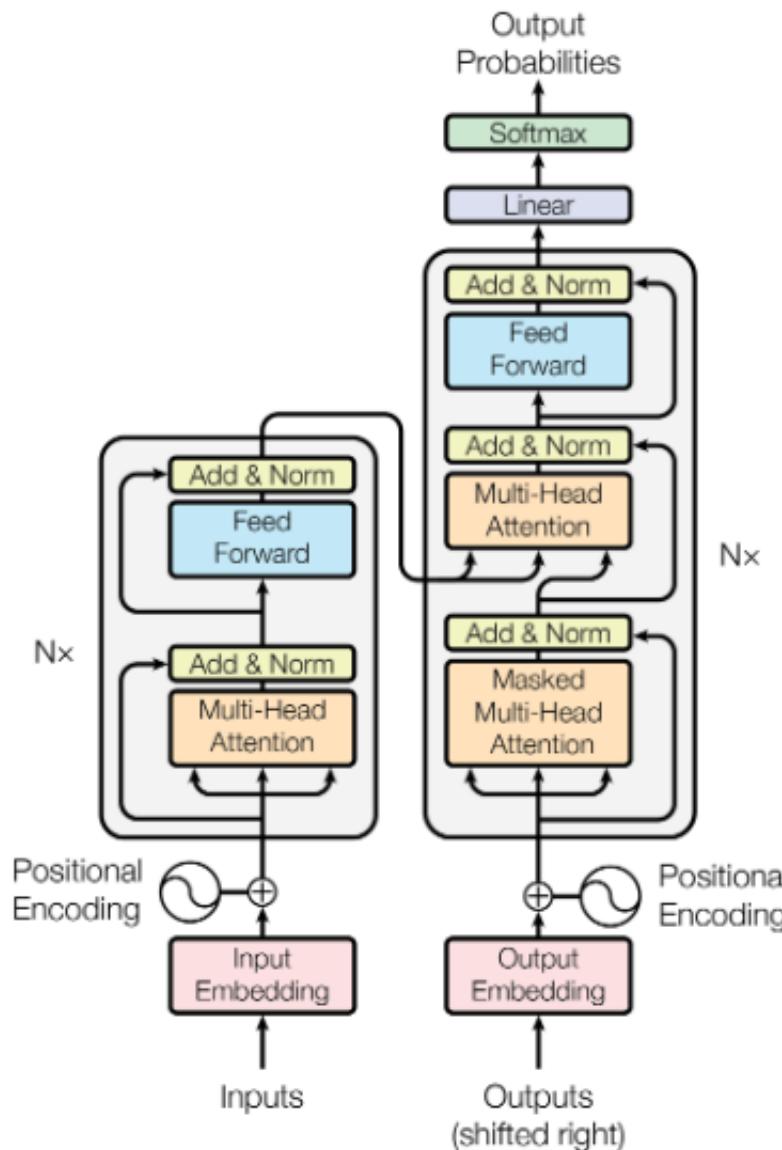
O avanço no campo do PLN tem sido marcado pela busca por representações vetoriais que capturem não apenas informações sintáticas, mas também aspectos semânticos e contextuais dos textos. As primeiras abordagens, como o *Word2Vec* de Mikolov *et al.* (2013) e o *GloVe* de Pennington, Socher e Manning (2014), consolidaram a noção de *embeddings*, isto é, vetores em espaços de alta dimensionalidade capazes de representar o significado aproximado de uma palavra. Esses modelos, embora inovadores em seu período, apresentavam a limitação de atribuir um único vetor fixo a cada termo, independentemente do contexto de ocorrência. Por exemplo, a palavra “banco” pode referir-se a uma instituição financeira ou a um assento, dependendo do contexto. Tal restrição, usualmente referida como o problema da *ambiguity of word meaning*, compromete a precisão em tarefas que exigem desambiguação semântica.

Um componente essencial para alcançar a compreensão semântica são os *embeddings* — representações numéricas que codificam o significado de palavras ou mesmo frases — que são essenciais na PLN para capturar relacionamentos complexos entre palavras e frases usando arquiteturas especiais conhecidas como transformadores (Galli *et al.* (2024, Tradução nossa)).

A verdadeira virada de paradigma ocorreu com a introdução do modelo *Transformer*, proposto por Vaswani *et al.* (2017) no artigo seminal *Attention Is All You Need*. Essa arquitetura rompeu com o paradigma das arquiteturas de redes recorrentes (RNNs) e convolucionais, fundamentando-se inteiramente no mecanismo de atenção (*attention mechanism*), permitindo que o modelo ponderasse a importância de diferentes palavras em uma sequência. Através dele, o modelo atribui pesos diferenciados a tokens em uma sequência, permitindo processar de forma simultânea e bidirecional a totalidade do contexto

textual. Essa propriedade conferiu aos *Transformer-based* models a capacidade de gerar representações contextuais, um avanço significativo em relação às técnicas anteriores.

Figura 1 – Transformador - modelo arquitetural.



Fonte: Vaswani *et al.* (2017, p. 24, Tradução nossa)

Sobre essa base arquitetônica foram desenvolvidos os LLMs pré-treinados, entre os quais se destaca o BERT (*Bidirectional Encoder Representations from Transformers*), introduzido por Devlin *et al.* (2019). Diferentemente de abordagens anteriores, como o *GPT-1* de Radford e Narasimhan (2018), que utilizava um treinamento unidirecional, o BERT foi projetado com pré-treinamento bidirecional, possibilitando a modelagem simultânea do contexto à esquerda e à direita de cada token. Essa característica permite a geração de representações semânticas profundas e contextualmente dependentes, adequadas para

tarefas de classificação, extração de relações e análise semântica. Em termos conceituais, essa bidirecionalidade constitui um elemento central para este trabalho, uma vez que fornece vetores de alta qualidade para unidades textuais de diferentes granularidades (palavras, sentenças e documentos).

No âmbito da modelagem de tópicos, os *embeddings* derivados do BERT são utilizados em variantes como o *Sentence-BERT* (SBERT), proposto por Reimers e Gurevych (2019), cujo objetivo é otimizar a geração de *sentence embeddings*. Tais representações são fundamentais para o funcionamento do *BERTopic*, introduzido por Grootendorst (2022), uma vez que o algoritmo se apoia em medidas de similaridade semântica para agrupar documentos. Essa abordagem, diferentemente de técnicas tradicionais baseadas em frequência de termos — como LDA e PLSA —, permite a organização de corpora heterogêneos a partir de relações de significado. Em cenários multilingues, como no caso da plataforma SIMCC, a utilização de modelos como o *paraphrase-multilingual-minilm-l12-v2* é particularmente relevante, visto que tais modelos produzem *embeddings* semanticamente consistentes mesmo em diferentes idiomas. Estudos como o de Weng, Wu e Dyer (2022) reforçam a pertinência dessa estratégia ao demonstrarem que representações baseadas em transformadores, quando associadas a métodos de agrupamento, revelam-se eficazes para a detecção e visualização de tópicos em coleções científicas.

Ainda que ferramentas como o *BERTopic* se mostrem adequadas para a identificação inicial de tópicos, uma limitação frequentemente relatada diz respeito à interpretabilidade dos rótulos gerados, que tendem a ser genéricos ou de difícil compreensão. Nesse ponto, a integração com modelos de geração de linguagem natural mais recentes, como o GPT-4, torna-se pertinente. Ao empregar sua capacidade de compreensão contextual e síntese textual, o GPT-4 pode ser utilizado para refinar e enriquecer a rotulagem dos tópicos identificados, além de produzir sumarizações mais coerentes e descriptivas. Essa etapa complementar insere-se como um mecanismo de aprimoramento da interpretabilidade dos resultados obtidos, contribuindo para análises mais consistentes do ponto de vista científico.

2.3 Abordagens Tradicionais de Modelagem de Tópicos

Com o crescimento exponencial de dados textuais e a consequente necessidade de organizar informação em larga escala, a modelagem de tópicos consolidou-se como uma técnica fundamental na área de PLN. Em termos gerais, trata-se de um conjunto de métodos estatísticos cujo objetivo é identificar estruturas semânticas latentes — denominadas *tópicos* — em coleções de documentos. Assim, essas técnicas permitem inferir distribuições temáticas que não são explicitamente observáveis, mas que emergem a partir de regularidades no uso do vocabulário. Essa perspectiva abriu caminho para aplicações em áreas diversas, desde ciências sociais até biomedicina (Jung *et al.* (2024)).

Entre as abordagens iniciais destacam-se três marcos históricos: a *Latent Semantic Analysis* LSA, a *Probabilistic Latent Semantic Analysis* PLSA e a *Latent Dirichlet Allocation* LDA. Esses métodos não apenas moldaram a compreensão inicial sobre a representação semântica de textos, como também estabeleceram fundamentos conceituais e metodológicos que orientaram o desenvolvimento de modelos mais avançados.

A LSA, proposta por Deerwester *et al.* (1990), parte da decomposição de matrizes termo-documento por meio da técnica de *Singular Value Decomposition* (SVD). Nesse enquadramento, documentos e termos são projetados em um espaço vetorial de dimensionalidade reduzida, o que permite atenuar ruídos lexicais e capturar relações de similaridade latentes. Apesar de sua relevância histórica, a linearidade da LSA e sua insensibilidade a variações contextuais limitam seu desempenho em cenários onde relações semânticas complexas são determinantes (George e Sumathy (2023), Xie *et al.* (2020)).

Com o intuito de superar parte dessas limitações, Hofmann (1999), Hofmann (2001) introduziram a PLSA, que reformulou a representação semântica a partir de um modelo probabilístico. Nessa abordagem, cada ocorrência de palavra em um documento é modelada como proveniente de um tópico latente, de forma que a probabilidade conjunta de palavra w e documento d é expressa como:

$$P(w, d) = \sum_{z \in Z} P(z|d) P(w|z),$$

onde z representa o conjunto de tópicos latentes. Embora tenha representado um avanço em relação à LSA, a PLSA apresenta limitações notáveis, em especial no que se refere à escalabilidade: o número de parâmetros cresce linearmente com a quantidade de documentos, o que compromete sua generalização e a torna suscetível a *overfitting* (Datchanamoorthy, S e B (2023)).

A evolução natural desse paradigma ocorreu com a formulação da LDA, proposta por Blei, Ng e Jordan (2003). Ao contrário da PLSA, a LDA incorpora uma camada Bayesiana por meio da utilização de distribuições de *Dirichlet* como *priors*. Essa estrutura permite regularizar o modelo e definir uma distribuição de tópicos não apenas a nível de documento, mas também a nível de corpus, resultando em maior robustez e interpretabilidade. A LDA parte da premissa de que cada documento é representado como uma mistura de tópicos, e cada tópico, por sua vez, é caracterizado por uma distribuição de palavras. Essa formulação tornou o modelo amplamente aplicável em diferentes domínios, como saúde pública (Mifrah e Benlahmar (2020)) e eficiência energética (Polyzos e Wang (2022)).

Apesar de sua influência, tanto a LSA quanto a PLSA e a LDA compartilham limitações estruturais. Todas operam no paradigma de *bag-of-words*, que ignora a ordem e o contexto local das palavras, o que frequentemente conduz a representações semânticas superficiais em textos técnicos ou multilíngues (George e Sumathy (2023), Xie *et al.* (2020)). Além disso, a sensibilidade da LDA à definição do número de tópicos (K) representa um

desafio adicional: valores reduzidos podem fundir tópicos distintos em um único, enquanto valores elevados podem fragmentar temas coesos em subtemas artificiais (Datchanamoorthy, S e B (2023)).

A sensibilidade do LDA ao parâmetro do número de temas (K) é uma de suas desvantagens. Encontrar o valor ideal para (K) pode ser desafiador. O modelo pode simplificar excessivamente e combinar diferentes temas em um só se (K) for configurado muito baixo. No entanto, se (K) for configurado muito alto, o modelo pode se tornar muito complexo e produzir temas errôneos (Datchanamoorthy, S e B (2023, Traduzido)).

Essas restrições evidenciam que, embora fundamentais, tais técnicas não capturam relações profundas e não lineares entre palavras e tópicos. Esse cenário motivou a emergência de abordagens modernas baseadas em *embedding* e arquiteturas de *transformer* (Vaswani *et al.* (2017), Devlin *et al.* (2019), Radford e Narasimhan (2018)), que oferecem maior sensibilidade contextual e escalabilidade para corpora heterogêneos e de grande volume.

2.4 BERTopic: Uma Abordagem Moderna

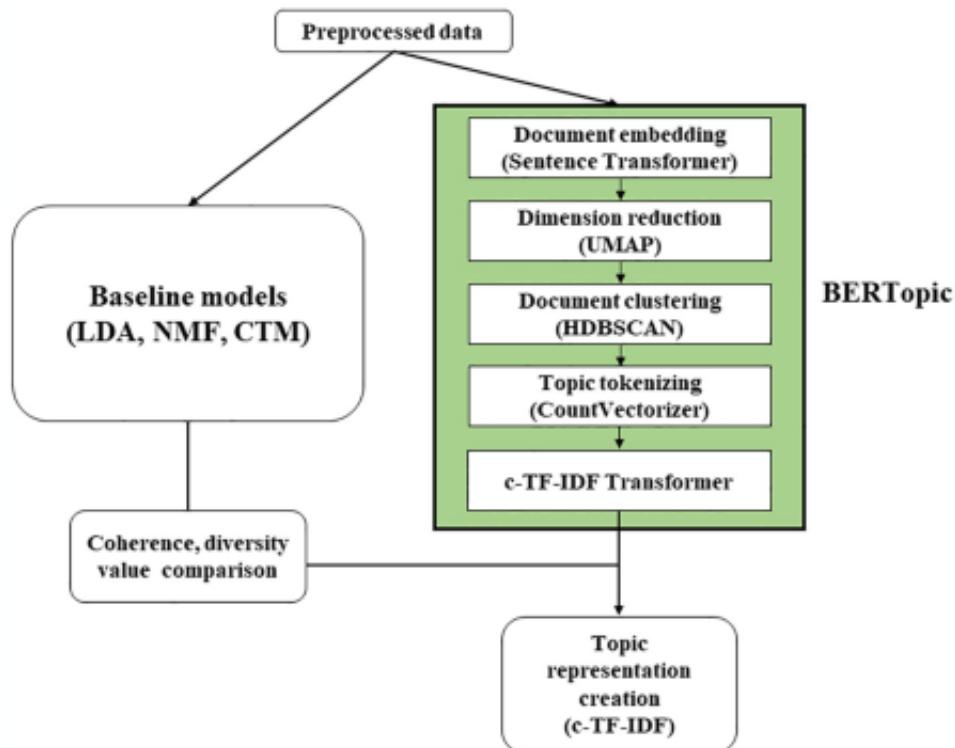
O *Bidirectional Encoder Representations from Transformers* BERT, introduzido por Devlin *et al.* (2019), marcou um avanço significativo no campo do *Natural Language Processing* NLP. Baseado na arquitetura de *Transformers* (Vaswani *et al.* (2017)), o BERT emprega o mecanismo de *self-attention* para capturar relações contextuais entre palavras em um texto. Diferentemente de abordagens anteriores, que analisavam sequências de maneira unidirecional, o BERT considera simultaneamente o contexto à esquerda e à direita de cada palavra, resultando em *embeddings* ricos e contextuais. Essa característica tornou o BERT amplamente utilizado em tarefas como classificação de texto, análise de sentimentos e resposta a perguntas.

Apesar de sua relevância, o BERT não foi projetado para tarefas de similaridade semântica entre sentenças ou documentos, pois os vetores que gera não são diretamente comparáveis em termos de proximidade semântica (Reimers e Gurevych (2019)). Essa limitação levou ao desenvolvimento do *Sentence-BERT* S-BERT, uma variante que adapta o BERT ao treinamento em redes siamesas (*Siamese Networks*) e funções de perda específicas, como *triplet loss*. O resultado é a produção de *embeddings* que são calculados por meio de técnicas como *Class-based Term Frequency-Inverse Document Frequency* (c-TF-IDF), que ajusta os pesos das palavras com base em suas frequências e relevâncias dentro de um corpus, elas podem ser comparadas de forma eficiente por meio de medidas de distância, como *cosine similarity*, viabilizando tarefas de busca semântica e agrupamento de documentos.

Sobre essa base, Grootendorst (2022) propôs o *BERTopic*, que não deve ser entendido como um único modelo, mas como um *pipeline* que integra diferentes técnicas

complementares para a modelagem de tópicos. Esse arranjo inicia-se pela geração de *embeddings* com o S-BERT, etapa que garante representações semânticas adequadas para comparação entre artigos científicos. Essa combinação permite que o BERTTopic identifique tópicos de maneira dinâmica e precisa, mesmo em grandes volumes de dados textuais diversificados (George e Sumathy (2023), Jung *et al.* (2024), Datchanamoorthy, S e B (2023)).

Figura 2 – Diagrama esquemático detalhado da comparação de métricas de avaliação entre modelos.



Fonte: (Jung *et al.* (2024, p. 7, Tradução nossa))

Como observado no diagrama comparativo entre modelos, a etapa de redução de dimensionalidade no pipeline utiliza o *Uniform Manifold Approximation and Projection* UMAP (McInnes, Healy e Melville (2018)), uma técnica que projeta vetores de alta dimensionalidade em um espaço reduzido. Essa abordagem se fundamenta em princípios teóricos de geometria Riemanniana e topologia algébrica, o que a diferencia de métodos mais antigos, como o *t-SNE* (Maaten e Hinton (2008)), e lhe confere maior escalabilidade e eficiência para a análise de grandes volumes de dados. O UMAP opera em duas fases principais: primeiro, constrói um grafo ponderado que representa a estrutura topológica dos dados em alta dimensão; em seguida, projeta esse grafo para um espaço de baixa dimensão, otimizando o layout para minimizar a entropia cruzada entre as duas representações. Essa metodologia é crucial para preservar tanto as estruturas locais quanto as globais do corpus, garantindo a coesão semântica dos dados. Ao aplicar o UMAP ao conjunto de

publicações científicas da plataforma SIMCC, é possível manter a fidelidade das relações entre os documentos, um requisito fundamental para a subsequente fase de agrupamento do BERTopic.

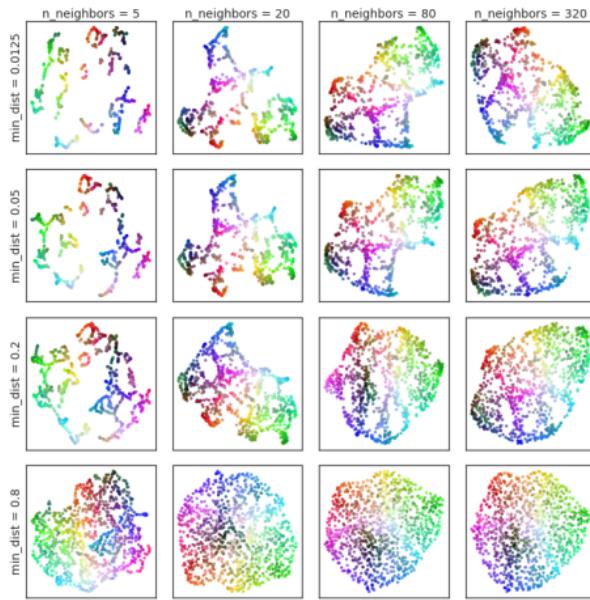


Figura 3 – Diagrama ilustrativo do UMAP, que demonstra a relação entre os hiperparâmetros *n_neighbors* e *min_dist* e a representação visual dos dados. O parâmetro *n_neighbors* controla a balança entre a preservação da estrutura global (valores altos) e local (valores baixos), enquanto *min_dist* ajusta a densidade dos agrupamentos, determinando a proximidade entre os pontos no espaço de baixa dimensionalidade. Esta visualização é crucial para otimizar o algoritmo e garantir que a estrutura semântica das publicações científicas seja fielmente representada para a subsequente etapa de agrupamento.

Fonte: (McInnes, Healy e Melville (2018, p. 24))

Com os vetores de alta dimensionalidade reduzidos pelo UMAP, a etapa subsequente é o agrupamento por meio do *Hierarchical Density-Based Spatial Clustering of Applications with Noise* HDBSCAN. Diferentemente de métodos clássicos como o K-Means, que assume *clusters* esféricos e de densidade uniforme, o HDBSCAN é um algoritmo de agrupamento baseado em densidade que não faz suposições prévias sobre a forma ou a densidade dos agrupamentos (Campello, Moulavi e Sander (2013)). Sua arquitetura hierárquica constrói uma árvore de conectividade que reflete a estrutura de densidade subjacente dos dados, permitindo a identificação de *clusters* de densidade variável. Essa capacidade é particularmente relevante para a análise de publicações científicas, onde a distribuição dos tópicos tende a ser heterogênea. O HDBSCAN também se destaca por sua robustez ao tratar documentos que não se ajustam a nenhum padrão temático, classificando-os como outliers de forma intrínseca, sem a necessidade de um passo de pós-processamento. Essa característica é especialmente relevante em contextos de produção científica, onde coexistem tanto publicações centrais com alta densidade de tópicos quanto trabalhos periféricos ou com temas emergentes. Essa abordagem garante que a sua análise não apenas identifique

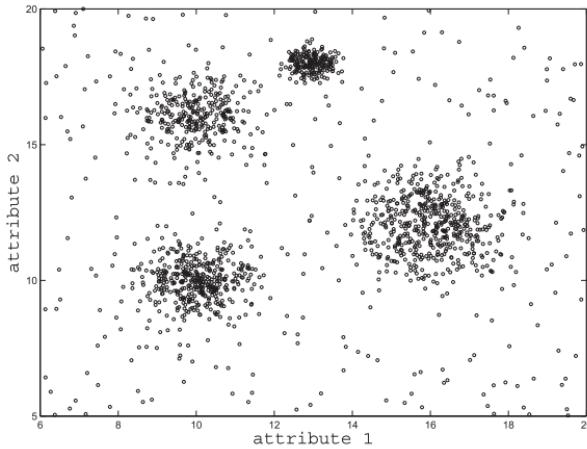


Figura 4 – Figura ilustrativa de um *dataset* sintético com quatro *clusters* e ruído de fundo. A imagem demonstra o tipo de desafio que o algoritmo HDBSCAN é capaz de superar, como a identificação de agrupamentos de densidades e formas variadas, além de tratar *outliers* de forma eficiente. Este comportamento é ideal para a análise de publicações científicas, onde a distribuição dos tópicos tende a ser heterogênea e não segue padrões geométricos rígidos.

Fonte: (Campello, Moulavi e Sander (2013, p. 16))

os tópicos dominantes, mas também lide eficientemente com a diversidade e o ruído natural do corpus da plataforma SIMCC.

Por fim, o BERTopic aplica o *class-based Term Frequency-Inverse Document Frequency* (c-TF-IDF), que trata cada cluster como um único documento. Essa abordagem destaca termos distintivos de cada grupo, permitindo identificar palavras-chave representativas mesmo quando não são as mais frequentes (Gana *et al.* (2024), Grootendorst (2022)). O c-TF-IDF, portanto, fornece uma base interpretável para a descrição de cada tópico.

A combinação dessas etapas — *embeddings* com S-BERT, redução de dimensionalidade com UMAP, clusterização com HDBSCAN e representação com c-TF-IDF — estabelece um fluxo robusto para a modelagem de tópicos. No contexto deste trabalho, esse *pipeline* constitui o núcleo do processo de análise das publicações científicas indexadas na plataforma SIMCC, servindo de ponto de partida para a integração com modelos de linguagem de grande escala, como o GPT-4, que será empregado para enriquecer semanticamente os rótulos dos tópicos e aprimorar sua interpretabilidade.

2.5 Modelos de Linguagem de Grande Escala (LLMs)

Os Modelos de Linguagem de Grande Escala (LLMs) constituem um marco no avanço do Processamento de Linguagem Natural (PLN), permitindo análises textuais sofisticadas e interpretações semânticas em volumes de dados sem precedentes. Fundamentados em arquiteturas baseadas em *transformers*, como o BERT, GPT e suas variantes, esses modelos utilizam aprendizado profundo para construir representações contextuais

dinâmicas de palavras e sentenças. Ao transformar o texto em *embeddings* semânticos, capturam relações latentes complexas entre elementos linguísticos, servindo de alicerce para tarefas como sumarização automática, classificação de documentos e modelagem de tópicos (Meng *et al.* (2024), Gana *et al.* (2024)).

Enquanto LLMs podem ser definidos de forma geral como sistemas de PLN capazes de aprender distribuições linguísticas a partir de grandes corpora não anotados, o modo como tais modelos realizam o pré-treinamento e o ajuste fino (*fine-tuning*) difere significativamente entre arquiteturas. As primeiras tentativas de modelos sequenciais, como Redes Neurais Recorrentes (RNNs) e Redes de Memória de Longo-Curto Prazo (LSTMs), apresentavam limitações na captura de dependências de longo alcance. Esse problema foi mitigado com a introdução do *Transformer* por Vaswani *et al.* (2017), cuja operação se baseia no Mecanismo de Atenção (*self-attention*, permitindo atribuir diferentes pesos às palavras do contexto e, consequentemente, capturar relações semânticas globais de maneira mais eficiente).

O treinamento de LLMs ocorre tipicamente em duas etapas complementares. Na fase de *pré-treinamento*, emprega-se aprendizado não supervisionado para expor o modelo a trilhões de palavras em dados textuais como documentos e dados da internet, consolidando padrões gerais da linguagem. Dois paradigmas se destacam nesse processo: (i) a Modelagem de Linguagem Autorregressiva, como no GPT, onde o modelo aprende a prever o próximo token a partir de uma sequência de tokens anteriores; e (ii) a Modelagem de Linguagem Mascarada, como no BERT, em que lacunas são ocultadas e o modelo deve inferi-las a partir do contexto bidirecional (Devlin *et al.* (2019), Jung *et al.* (2024)). Em seguida, ocorre o *fine-tuning*, etapa supervisionada em que o modelo é ajustado a tarefas específicas, como classificação de textos, análise de sentimentos ou sumarização, garantindo robustez e especialização ((Gana *et al.*, 2024)).

- **Modelagem de Linguagem Autorregressiva:** Modelos como o GPT (*Generative Pre-trained Transformer*) seguem um fluxo sequencial unidirecional, prevendo cada token com base nos anteriores. Essa abordagem favorece a coerência narrativa e a fluidez na geração textual, aspectos essenciais em tarefas de criação de conteúdo (Radford e Narasimhan (2018), Jung *et al.* (2024)).
- **Modelagem de Linguagem Mascarada:** Modelos como o BERT (*Bidirectional Encoder Representations from Transformers*) aplicam mascaramento aleatório em tokens, forçando o modelo a interpretar o contexto bidirecionalmente. Tal característica possibilita uma maior sensibilidade semântica, útil em tarefas como inferência textual e modelagem de tópicos (Devlin *et al.* (2019), Datchanamoorthy, S e B (2023)).

Entre os LLMs mais avançados, destaca-se o **GPT-4**, evolução do **GPT-1** desen-

volvido pela OpenAI por Radford e Narasimhan (2018), que introduziu sua arquitetura a partir de um trabalho seminal. Sua estrutura permanece fundamentada no paradigma *transformer* Vaswani *et al.* (2017), mas incorpora modificações substanciais em relação às versões anteriores. Embora a documentação oficial seja limitada por razões proprietárias, o *Technical Report* da OpenAI (OpenAI *et al.*, 2024) e análises independentes ??Achiam *et al.* (2023) sugerem que o GPT-4 conta com bilhões de parâmetros adicionais em comparação ao GPT-3, além de maior profundidade de camadas de atenção e mecanismos otimizados de paralelização no treinamento distribuído. Essas melhorias resultam em avanços na capacidade de raciocínio semântico, na robustez diante de contextos ambíguos e na generalização para tarefas pouco definidas.

Outro aspecto relevante é o aprimoramento nos métodos de alinhamento e segurança (*alignment*), alcançados por meio de técnicas como o *Reinforcement Learning with Human Feedback* (RLHF), que possibilitam ao modelo produzir respostas mais consistentes com critérios humanos de qualidade e relevância (OpenAI *et al.* (2024), Ouyang *et al.* (2022)). Além disso, o GPT-4 demonstra melhor desempenho em cenários multilíngues e em tarefas de alto nível cognitivo, como resolução de problemas em exames padronizados e síntese de conhecimento interdisciplinar (Achiam *et al.* (2023)). Essas características tornam o modelo especialmente adequado para aplicações acadêmicas e científicas, onde a precisão semântica e a interpretabilidade das respostas são fundamentais.

Ao compararmos o **BERTopic** e o GPT-4, evidenciam-se diferenças fundamentais na natureza e aplicação de cada modelo. O BERTopic, embora baseado em *embeddings* derivados de modelos como BERT, concentra-se em identificar e organizar tópicos latentes a partir de representações vetoriais de documentos, utilizando algoritmos já citados nas seções anteriores. Seu ponto forte está na capacidade de estruturar grandes volumes de dados em *clusters* semanticamente coerentes Grootendorst (2022). Já o GPT-4, além de gerar representações contextuais sofisticadas, pode ser utilizado para atribuir rótulos semânticos refinados a tais *clusters*, ampliando a interpretabilidade dos tópicos e permitindo a construção de narrativas explicativas sobre tendências detectadas nos dados Meng *et al.* (2024), Galli *et al.* (2024).

No contexto deste projeto, a integração de ambos os modelos se mostra justificada. Enquanto o BERTopic viabiliza a organização automática de grandes corpora textuais provenientes da plataforma SIMCC, o GPT-4 agrega valor na etapa de rotulagem, interpretação semântica e análise contextual aprofundada. Tal combinação potencializa tanto a acurácia quanto a inteligibilidade dos resultados, conciliando rigor metodológico com clareza interpretativa. Além disso, a aplicação conjunta favorece a detecção de padrões emergentes em múltiplos idiomas, aspecto essencial dada a heterogeneidade linguística base de dados.

Portanto, ao invés de restringir-se a abordagens puramente estatísticas ou unica-

mente gerativas, este trabalho adota uma perspectiva híbrida, combinando técnicas de modelagem de tópicos e de raciocínio semântico avançado, buscando suprir lacunas de interpretabilidade.

3 TRABALHOS CORRELATOS

A análise de grandes volumes de publicações científicas por meio de técnicas computacionais é um campo de pesquisa em franca expansão. A revisão da literatura, portanto, é fundamental para contextualizar e justificar a abordagem proposta neste trabalho — um *pipeline* que integra *BERTopic* e *GPT-4* para análise do acervo da plataforma SIMCC. Os estudos selecionados a seguir validam as escolhas metodológicas deste projeto, demonstrando a superioridade de modelos de tópicos baseados em *embeddings*, a flexibilidade de seus componentes e a sinergia promissora com *Large Language Models* (LLMs) para enriquecimento semântico.

A decisão de utilizar o *BERTopic* como ferramenta central de modelagem de tópicos é respaldada por estudos comparativos que atestam sua superioridade frente a abordagens tradicionais. A pesquisa de Jung *et al.* (2024) é emblemática nesse sentido, ao realizar uma análise comparativa entre métodos como *Latent Dirichlet Allocation* (LDA), *Nonnegative Matrix Factorization* (NMF) e o *BERTopic*, aplicando-os a dados acadêmicos e de mídia. Os autores concluíram que o *BERTopic*, que combina *embeddings* de texto com técnicas de redução de dimensionalidade e clusterização, "demonstrou superioridade em diversidade e coesão de tópicos" (Jung *et al.* (2024, p. 27)). Essa capacidade de capturar contextos semânticos complexos, superando a abordagem de "saco de palavras" (*bag-of-words*) do LDA, é crucial para o objetivo dessa pesquisa, que é analisar a produção científica interdisciplinar e multilíngue da plataforma SIMCC. A eficácia do *BERTopic* deriva de sua arquitetura, que se apoia em *embeddings* de sentenças, como os gerados pelo *Sentence-BERT* (*SBERT*) (Reimers e Gurevych (2019)), para agrupar documentos com base na similaridade semântica, e não apenas na frequência lexical.

A arquitetura do *BERTopic* não é apenas robusta, mas também modular, permitindo a exploração de diferentes configurações para otimizar os resultados, um ponto investigado por Wijanto, Widiasuti e Yong (2024). Em seu trabalho, os autores exploraram o ajuste de hiperparâmetros em modelos baseados em *BERT*, testando combinações variadas de modelos de *embedding* (como *RoBERTa* e *S-BERT*), técnicas de redução de dimensionalidade (*UMAP* e *PCA*) e algoritmos de clusterização (*K-Means* e *HDBSCAN*). Embora a combinação que eles identificaram como ótima (*RoBERTa* + *PCA* + *K-Means*) divirja da utilizada neste projeto (*S-BERT* + *UMAP* + *HDBSCAN*), o estudo reforça um ponto central: a importância da seleção criteriosa de cada componente do *pipeline* para garantir a geração de tópicos coerentes e interpretáveis. A metodologia dessa pesquisa, portanto, adota a configuração padrão e amplamente validada do *BERTopic* (Grootendorst (2022)), que utiliza *UMAP* para preservação da estrutura topológica dos dados e *HDBSCAN* pela sua capacidade de identificar *clusters* de densidades variadas e lidar com

ruído, sendo ideal para a heterogeneidade esperada nos dados da SIMCC.

Uma vez que os tópicos são *clusterizados*, a etapa de rotulagem torna-se crítica para a interpretabilidade dos resultados. Abordagens tradicionais, baseadas apenas nas palavras-chave mais frequentes, frequentemente geram rótulos genéricos. A literatura recente aponta para o uso de LLMs como uma solução eficaz para este desafio. Kozlowski, Pradier e Benz (2024) propõem uma abordagem que utiliza o *BERTopic* para gerar tópicos e, em seguida, alimenta LLMs como o *GPT-4* com as dez palavras-chave mais representativas de cada tópico para gerar rótulos automáticos. A conclusão dos autores alinha-se diretamente com a segunda etapa do nosso *pipeline*: os modelos *GPT-4* são capazes de rotular os tópicos de forma precisa e acurada, e "rótulos de 3 palavras são preferíveis para capturar a complexidade dos tópicos de pesquisa" (Kozlowski, Pradier e Benz (2024), p. 1). Este achado valida a hipótese central deste projeto de que o *GPT-4* pode ser empregado para enriquecer semanticamente os *clusters* identificados pelo *BERTopic*, gerando rótulos descriptivos que superam as limitações de métodos puramente estatísticos.

A integração de LLMs e *BERTopic* em um *pipeline* coeso para análise bibliométrica já foi validada em contextos similares. Meng *et al.* (2024) propõem uma metodologia que também utiliza *BERTopic* e *GPT* para mapear a evolução da pesquisa científica em um grande volume de publicações. Notavelmente, os autores utilizam a API do *GPT* para gerar os *embeddings* dos documentos antes de aplicar o *BERTopic*, em vez do *SBERT*, demonstrando a flexibilidade e as diferentes possibilidades de integração entre essas tecnologias. Além disso, o trabalho de (Meng *et al.* (2024)) culmina no desenvolvimento de uma plataforma *web* de análise bibliométrica para visualização de redes e tópicos, um objetivo análogo ao proposto nessa pesquisa com o uso da ferramenta *WizMap* para exploração interativa dos resultados.

Seguindo uma linha similar, Gana *et al.* (2024) apresentam um *framework* semi-automático para revisões de literatura que integra LLMs (como o *SOLAR-10.7B*), *BERTopic* e *KeyBERT*. O *framework* utiliza um modelo de *embedding* específico (*Bge-large-en-v1.5*) e segue o *pipeline* padrão do *BERTopic* (*UMAP* e *HDBSCAN*) para identificar tópicos coesos. O LLM é então empregado para refinar a representação dos tópicos, gerando rótulos descriptivos e selecionando documentos representativos. Assim como os estudos de (Meng *et al.* (2024)) e Kozlowski, Pradier e Benz (2024), este trabalho reforça que a combinação de um modelo de tópicos moderno com um LLM constitui uma abordagem de ponta para a análise de literatura científica, validando a arquitetura geral do *pipeline* proposto para a plataforma SIMCC.

Em suma, a análise dos trabalhos correlatos demonstra que a proposta desse projeto está firmemente ancorada em práticas e pesquisas recentes e relevantes. A literatura confirma a superioridade do *BERTopic* sobre métodos tradicionais (Jung *et al.* (2024)), destaca a importância da configuração de seu *pipeline* modular (Wijanto, Widiastuti e

Yong (2024)) e valida o uso de LLMs como o *GPT-4* para a tarefa de enriquecimento semântico dos tópicos gerados (Kozlowski, Pradier e Benz (2024)). Além disso, *frameworks* integrados, que combinam essas duas tecnologias, já foram aplicados com sucesso para mapear e analisar a produção científica (Meng *et al.* (2024); Gana *et al.* (2024)). Este trabalho, portanto, avança ao aplicar essa metodologia de ponta a um contexto específico e de grande relevância regional — a base de dados da plataforma SIMCC —, com o objetivo de não apenas extrair conhecimento, mas também de disponibilizá-lo de forma interativa e visual.

3.1 Síntese Comparativa dos Trabalhos Correlatos

A fim de consolidar a análise da literatura e posicionar de forma clara a contribuição desta pesquisa, o quadro a seguir (Quadro 1) apresenta uma síntese comparativa dos trabalhos correlatos discutidos. A comparação é estruturada com base em critérios essenciais, como o objetivo principal de cada estudo, o pipeline metodológico empregado, as tecnologias de embedding e o uso específico de LLMs. Essa estrutura permite visualizar as sinergias e as particularidades de cada abordagem, destacando como este TCC se fundamenta e avança em relação ao estado da arte.

Tabela 1 – Quadro Resumo: Comparativo de Trabalhos Correlatos

| Referência | Objetivo Principal | Pipeline/Método Utilizado | Modelo de Embedding | Uso do LLM | Relação com Esta Pesquisa |
|-----------------------------------|---|---|--|---|--|
| Jung et al. (2024) | Comparar o desempenho de modelos de tópicos (LDA, NMF, BERTopic) em textos acadêmicos e de notícias sobre LLMs. | Análise comparativa de métricas de coerência e diversidade dos tópicos gerados. | SBERT (implícito no BERTopic). | O estudo é <i>sobre</i> LLMs, mas não os utiliza no pipeline. | Justifica a escolha do BERTopic, demonstrando sua superioridade sobre métodos tradicionais para analisar textos acadêmicos e capturar nuances semânticas. |
| Wijanto, Widiastuti e Yong (2024) | Explorar o ajuste de hiperparâmetros e encontrar a configuração ótima para modelos de tópicos baseados em BERT em artigos científicos. | BERTopic com diferentes combinações de embeddings (RoBERTa, SBERT), redução de dimensionalidade (UMAP, PCA) e clusterização (K-Means, HDBSCAN). | Testou múltiplos, incluindo SBERT e RoBERTa. | Nenhum. O foco é na otimização do pipeline de modelagem de tópicos. | Valida a abordagem metodológica, mostrando que a seleção de cada componente do pipeline é uma etapa crucial. Reforça que a configuração usada neste TCC é uma das mais consolidadas. |
| Kozlowski, Pradier e Benz (2024) | Avaliar a confiabilidade de diferentes LLMs para rotular automaticamente os tópicos gerados pelo BERTopic. | Pipeline em duas etapas: 1. Geração de tópicos com BERTopic; 2. Envio das palavras-chave para um LLM gerar o rótulo. | SBERT (implícito no BERTopic). | Rotulagem de Tópicos (usando flan, GPT-4 mini e GPT-4). | Justifica diretamente a segunda etapa do pipeline, confirmando que o GPT-4 é eficaz para criar rótulos descritivos e precisos. |
| Meng et al. (2024) | Mapear a evolução de um campo de pesquisa científica utilizando uma abordagem integrada de LLM e modelagem de tópicos. | Pipeline integrado: 1. Geração de embeddings com a API do GPT; 2. Clusterização e modelagem com BERTopic. | GPT-3.5 (text-embedding-ada-002). | Geração de Embeddings e análise semântica. | Valida o conceito do pipeline completo (LLM + BERTopic). Apresenta uma arquitetura alternativa e reforça o objetivo de criar uma interface de visualização. |
| Gana et al. (2024) | Propor um framework semi-automático para revisões de literatura, combinando LLMs e BERTopic. | Pipeline integrado: 1. Embedding com BGE; 2. Clusterização com BERTopic (UMAP + HDBSCAN); 3. Refinamento da representação dos tópicos com LLM. | BAAI/bge-large-en-v1.5. | Refinamento da Representação de Tópicos (usando SOLAR-10.7B). | Posiciona este TCC dentro de uma tendência de pesquisa de criar frameworks de análise. Mostra a flexibilidade do pipeline ao usar diferentes modelos. |
| Esta Pesquisa | Analizar as publicações científicas da plataforma SIMCC para identificar padrões temáticos e facilitar a exploração visual do conhecimento. | Pipeline integrado: 1. SBERT; 2. BERTopic (UMAP + HDBSCAN); 3. GPT-4 para rotulagem; 4. Visualização com WizMap. | Sentence-BERT (paraphrase-multilingual-MiniLM-L12-v2). | Rotulagem e Enriquecimento Semântico de Tópicos (GPT-4). | Aplica um pipeline de ponta, validado pela literatura, a um conjunto de dados único (SIMCC) para resolver um problema prático de gestão do conhecimento. |

4 METODOLOGIA

Este estudo adota a Design Science Research (DSR) como sua principal estrutura metodológica para o desenvolvimento e a validação de um artefato tecnológico. A DSR é particularmente adequada para esta pesquisa, pois seu foco reside na criação de soluções com caráter de inovação para problemas práticos, alinhando rigor científico com relevância aplicada (Dresch, Lacerda e Antunes (2015)). O objetivo é construir um pipeline computacional que combina BERTopic e GPT-4 para otimizar a análise de publicações científicas na plataforma SIMCC.

O processo de DSR orienta o projeto de forma iterativa, desde a concepção do problema até a comunicação dos resultados, conforme ilustrado no fluxograma da Figura 5. A seguir, cada etapa do DSR é detalhada e contextualizada no escopo deste trabalho.

4.1 Identificação do Problema e Definição de Objetivos

A primeira fase da DSR, representada na Figura 5 pelos campos “Contexto” e “Problema”, consiste na identificação de uma lacuna relevante. Atualmente, a busca na plataforma SIMCC é limitada a palavras-chave, o que dificulta a identificação de conexões semânticas, temas interdisciplinares e publicações em diferentes idiomas. Essa limitação evidencia a necessidade de uma solução que otimize a gestão do conhecimento acadêmico. O problema de pesquisa, portanto, é: como a combinação de técnicas avançadas de modelagem de tópicos (BERTopic) e modelos de linguagem de grande escala (GPT-4) pode superar as limitações das buscas tradicionais na plataforma SIMCC, melhorando a análise e a classificação de temas emergentes?

A partir disso, formulam-se as “Conjecturas Teóricas”: a integração do BERTopic com o GPT-4 tem o potencial de melhorar significativamente a identificação, classificação e visualização de temas emergentes, oferecendo uma compreensão mais profunda dos dados por meio de uma rotulagem semântica enriquecida e visualizações interativas.

4.2 Desenvolvimento do Artefato

Com base nos objetivos, a etapa seguinte é o desenvolvimento do “Artefato”, que neste trabalho é o pipeline integrado ao SIMCC. Este pipeline é projetado para:

- Realizar a modelagem de tópicos com o BERTopic, utilizando embeddings contextuais para identificar padrões temáticos de forma robusta.

- Utilizar o GPT-4 para a rotulagem semântica, gerando rótulos descritivos e precisos que facilitam a interpretação dos resultados.
- Apresentar os tópicos em uma interface gráfica interativa, como o WizMap, para facilitar a exploração das conexões temáticas.

A construção deste artefato se apoia no “Estado da Técnica”, que envolve uma revisão da literatura sobre Modelagem de Tópicos, Embeddings Contextuais, Redução de Dimensionalidade (UMAP), Algoritmos de Clusterização (HDBSCAN) e LLMs.

4.3 Avaliação e Validação

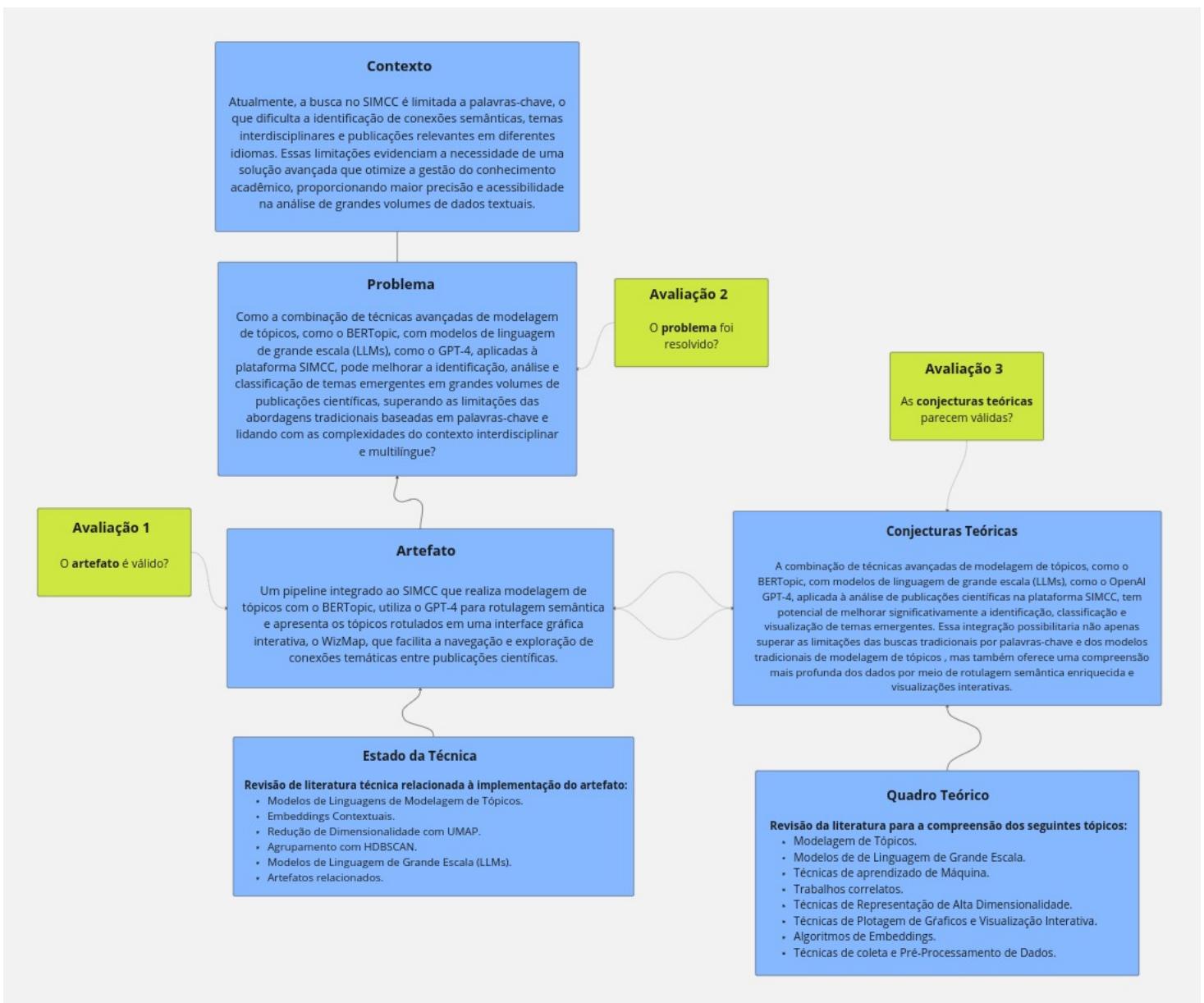
O ciclo de DSR exige uma avaliação rigorosa, representada no diagrama pelas “Avaliações” 1, 2 e 3. A validação do artefato (Avaliação 1: O artefato é válido?) será conduzida por meio de métricas quantitativas e qualitativas. Utilizaremos o *Topic Coherence Score* para medir a coerência semântica dos tópicos gerados pelo pipeline. Adicionalmente, será realizada uma análise comparativa do tempo de análise entre o sistema atual e a solução proposta, visando demonstrar ganhos de eficiência.

A Avaliação 2 (O problema foi resolvido?) e a Avaliação 3 (As conjecturas teóricas parecem válidas?) serão respondidas ao final do experimento. A apresentação dos resultados, mostrando a precisão dos tópicos e a clareza dos rótulos gerados pelo GPT-4, servirá para validar se o artefato efetivamente soluciona o problema identificado e se as conjecturas teóricas se confirmam na prática.

4.4 Apresentação dos Resultados e Comunicação

A etapa final do ciclo DSR, embora não explicitada como uma caixa separada no diagrama, é a comunicação dos achados. Os resultados obtidos serão documentados detalhadamente neste trabalho, destacando o funcionamento do pipeline, a qualidade dos tópicos identificados e a eficácia da interface de visualização. O objetivo é não apenas apresentar uma solução tecnológica, mas também contribuir com conhecimento para as áreas de Ciência da Informação e Processamento de Linguagem Natural.

Figura 5 – Adaptação da Design Science Research para este projeto.



Fonte: O Autor

5 PROJETO DE DESENVOLVIMENTO

O desenvolvimento do artefato proposto neste trabalho segue uma abordagem estruturada e incremental, alinhada à metodologia Design Science Research (DSR) detalhada anteriormente. O objetivo é construir e validar um pipeline computacional que venha a aprimorar e complementar a análise de publicações científicas na plataforma do Sistema de Mapeamento de Competências Científicas da Bahia (SIMCC).

A plataforma SIMCC, conforme descrito por Santos *et al.* (2024), já possui uma arquitetura robusta para o mapeamento de competências, integrando fontes de dados heterogêneas como a Plataforma Lattes, a Plataforma Sucupira e o Journal Citation Reports (JCR). Seu sistema atual utiliza um processo de Extração, Transformação e Carga (ETL) para consolidar as informações em um banco de dados PostgreSQL e já emprega técnicas de Processamento de Linguagem Natural (PLN) para a recuperação de informações baseadas em termos e palavras-chave.

A busca lexical existente é uma ferramenta poderosa para a recuperação de informações diretas, onde o usuário já sabe quais termos procurar. No entanto, como aponta Jorge *et al.* (2025), a grande variação terminológica em domínios científicos complexos representa um desafio para a descoberta de conexões temáticas que não são óbvias. Nesse sentido, há uma oportunidade de incrementar a plataforma com uma nova camada de análise semântica, que permita ao usuário explorar o conhecimento de forma mais intuitiva e visual.

Este projeto de desenvolvimento detalha, portanto, a construção de um pipeline que representa uma evolução para a arquitetura do SIMCC. A solução proposta não visa substituir, mas sim enriquecer a funcionalidade de busca atual, introduzindo a modelagem de tópicos moderna com o BERTopic e o poder de contextualização do GPT-4. O resultado final é a geração de um mapa de clusters interativo, o WizMap, que modifica a forma como os temas são descobertos na plataforma. O foco é transcender a lista de resultados tradicional, permitindo que os usuários naveguem visualmente pelas principais áreas de pesquisa, identifiquem temas emergentes e compreendam as relações entre os diferentes campos do conhecimento de maneira orgânica.

As subseções a seguir descrevem as etapas-chave do desenvolvimento deste artefato, abordando a arquitetura da solução, as tecnologias empregadas e o fluxo de processamento, desde a ingestão e pré-processamento dos dados até a geração dos tópicos, a rotulagem semântica e a visualização interativa dos resultados.

5.1 Tecnologias Utilizadas

O desenvolvimento do artefato proposto neste TCC assenta-se sobre a combinação da infraestrutura tecnológica já consolidada da plataforma SIMCC com um novo pipeline de análise semântica e visualização de dados, construído com ferramentas de ponta em Processamento de Linguagem Natural (PLN) e aprendizado de máquina. Esta seção detalha as tecnologias que compõem tanto a base da plataforma quanto o pipeline de inovação proposto.

5.1.1 Base Tecnológica da Plataforma SIMCC

A arquitetura do SIMCC, que serve como ponto de partida para este trabalho, foi projetada para ser robusta e escalável, utilizando um conjunto de tecnologias de mercado para a gestão de dados acadêmicos.

- **Banco de Dados (PostgreSQL):** A escolha do PostgreSQL como Sistema de Gerenciamento de Banco de Dados (SGBD) para o SIMCC é estratégica. Conforme detalhado por Jorge *et al.* (2025, p. 255), o sistema aproveita os recursos nativos de PLN textual do PostgreSQL para indexação e busca. A sua capacidade de lidar com grandes volumes de dados e sua extensibilidade o tornam ideal para armazenar não apenas os dados estruturados extraídos dos currículos Lattes, mas também para suportar as representações vetoriais (embeddings) que são centrais neste projeto.
- **Orquestração de Dados (Apache Hop):** Para a extração, transformação e carga (ETL) dos dados de fontes diversas como a Plataforma Lattes, Sucupira e JCR, o SIMCC utiliza o Apache Hop. Essa ferramenta é responsável por automatizar e coordenar o fluxo de ingestão de dados, garantindo a consistência e a qualidade das informações que alimentarão o pipeline de análise (Santos *et al.* (2024)).
- **Linguagem de Back-end (Python):** O back-end da plataforma foi desenvolvido em Python, uma escolha que se alinha perfeitamente aos objetivos deste TCC. A linguagem oferece um ecossistema maduro para ciência de dados e PLN, com bibliotecas como a NLTK, já em uso no SIMCC, e as bibliotecas *Transformers* e *BERTopic*, que são o cerne deste trabalho.

5.1.2 Pipeline de Modelagem e Análise de Tópicos

Sobre a base existente do SIMCC, este projeto implementa um novo pipeline focado na descoberta e análise de conhecimento, utilizando as seguintes tecnologias:

- **Modelagem de Tópicos (BERTopic):** Para a identificação dos temas latentes nas publicações científicas, foi escolhido o BERTopic. Diferentemente de abordagens

clássicas como o LDA, o BERTopic é um modelo moderno que utiliza embeddings contextuais para agrupar documentos com base na similaridade semântica. Sua arquitetura modular, que combina a geração de embeddings com Sentence-BERT (SBERT), a redução de dimensionalidade com UMAP e a clusterização com HDBSCAN, permite a extração de tópicos mais coerentes e representativos, sendo ideal para a complexidade e a variação terminológica dos dados acadêmicos.

- **Enriquecimento Semântico (GPT-4):** Uma das principais inovações deste pipeline é o uso do GPT-4 para aprimorar a interpretabilidade dos tópicos. Após o BERTopic identificar os clusters temáticos e extrair palavras-chave, o GPT-4 é utilizado para gerar rótulos descritivos, concisos e semanticamente ricos para cada tópico. Esta etapa transforma uma lista de palavras-chave, muitas vezes ambígua, em um título claro e comprehensível, agregando um valor analítico significativo e facilitando a compreensão dos resultados pelo usuário final.
- **Visualização Interativa (WizMap):** Para apresentar os resultados da modelagem de tópicos de forma intuitiva, a ferramenta WizMap foi selecionada. Conforme descrito por Wang, Hohman e Chau (2023), o WizMap é uma solução de visualização escalável projetada para explorar grandes volumes de embeddings em uma interface inspirada em mapas. Em vez de apresentar os tópicos em listas ou gráficos estáticos, o WizMap permite que o usuário navegue por um mapa de clusters, explore as relações entre os temas, aplique zoom para investigar sub-tópicos e identifique visualmente as áreas de maior densidade de pesquisa, incrementando a forma como o conhecimento é descoberto na plataforma.

5.2 Arquitetura da solução

A arquitetura da solução proposta foi projetada para se integrar de maneira fluida à infraestrutura existente da plataforma SIMCC, adicionando uma nova camada de análise semântica e exploração de conhecimento. O pipeline é composto por uma sequência de etapas modulares que transformam os dados textuais brutos das publicações científicas em um mapa de tópicos interativo e semanticamente rico. O fluxo completo, desde a ingestão dos dados até a visualização, é detalhado nas subseções a seguir.

5.2.1 Coleta e Pré-processamento dos Dados

O ponto de partida do pipeline são os dados já consolidados no banco de dados PostgreSQL da plataforma SIMCC. Conforme documentado por Santos *et al.* (2024), esses dados, que incluem títulos, resumos e metadados de publicações científicas, já passaram por um rigoroso processo de ETL. Para adequá-los à modelagem de tópicos, uma etapa de pré-processamento textual é executada, consistindo em:

- **Limpeza e Normalização:** Remoção de caracteres especiais, conversão de todo o texto para minúsculas e padronização de acentuação para garantir consistência.
- **Tokenização:** Divisão dos textos (títulos e resumos) em unidades menores (tokens), como palavras ou sentenças.
- **Remoção de Stopwords:** Exclusão de palavras funcionalmente importantes mas semanticamente vazias (e.g., "o", "de", "para", "com"), que poderiam gerar ruído na análise de tópicos.
- **Lematização:** Redução das palavras à sua forma canônica (lema) para agrupar diferentes flexões de um mesmo termo (e.g., "pesquisas", "pesquisou" e "pesquisando" são reduzidos a "pesquisar"), consolidando assim o vocabulário.

5.2.2 Geração de Embeddings Contextuais

Após o pré-processamento, os textos são transformados em representações vetoriais numéricas, conhecidas como embeddings. Esta é a etapa fundamental que permite a análise semântica. Para este fim, utiliza-se o modelo *Sentence-BERT* (SBERT), especificamente a variante ***paraphrase-multilingual-MiniLM-L12-v2***. A escolha deste modelo se justifica por sua alta eficiência e por sua capacidade de gerar embeddings semanticamente consistentes em múltiplos idiomas, uma característica essencial para a base de dados multilíngue do SIMCC. Cada documento (publicação) é, então, representado por um vetor denso em um espaço de alta dimensionalidade, onde a proximidade entre vetores indica similaridade semântica.

- **Redução de Dimensionalidade com UMAP:** Os embeddings de alta dimensionalidade são projetados em um espaço de menor dimensão utilizando o algoritmo UMAP (Uniform Manifold Approximation and Projection). Esta técnica é crucial por preservar tanto a estrutura local quanto a global dos dados, garantindo que as relações semânticas entre os documentos sejam mantidas de forma fidedigna.
- **Clusterização com HDBSCAN:** No espaço de dimensionalidade reduzida, os vetores dos documentos são agrupados pelo algoritmo HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Diferentemente de métodos como o K-Means, o HDBSCAN é capaz de identificar clusters de diferentes formas e densidades, além de classificar documentos que não pertencem a nenhum grupo coeso como ruído (outliers). Cada cluster denso de documentos formado nesta etapa representa um tópico latente.
- **Extração de Palavras-Chave com c-TF-IDF:** Para tornar os clusters interpretáveis, o BERTopic emprega uma variação do TF-IDF, chamada c-TF-IDF (class-based

Term Frequency-Inverse Document Frequency). Este método trata todos os documentos de um cluster como um único documento e calcula a importância das palavras, destacando os termos que são mais representativos de cada tópico em comparação com os demais.

5.2.3 Redução de Dimensionalidade

A redução de dimensionalidade é uma etapa fundamental em pipelines que lidam com dados de alta dimensionalidade, como embeddings gerados para textos. Ela transforma os dados em representações mais compactas, preservando informações essenciais e eliminando redundâncias ou ruídos. No projeto, utilizamos o *Uniform Manifold Approximation and Projection (UMAP)*¹ McInnes, Healy e Melville (2018) devido à sua capacidade de preservar tanto as relações locais quanto globais entre os dados, o que é crucial para a análise semântica e a modelagem de tópicos.

O UMAP é um algoritmo de redução de dimensionalidade baseado em princípios matemáticos da teoria de grafos e geometria algébrica. Ele funciona em várias etapas:

- **Construção do Grafo de Alta Dimensionalidade:** O UMAP começa representando os dados originais como um grafo ponderado em alta dimensionalidade. Cada ponto de dado é conectado aos seus vizinhos mais próximos com pesos que indicam a proximidade entre os pontos. Essa etapa é realizada usando algoritmos como *k-nearest neighbors (k-NN)*, que identificam os pontos mais próximos no espaço original.
- **Definição de uma Distribuição de Probabilidade Local:** Uma distribuição de probabilidade local é ajustada ao redor de cada ponto, calculando a probabilidade de uma conexão entre pontos baseando-se na distância.
- **Projeção para Baixa Dimensionalidade:** O algoritmo então otimiza um grafo em um espaço de menor dimensionalidade. Ele tenta preservar a estrutura do grafo original, minimizando a diferença entre os dois grafos usando uma função de perda. Esse processo garante que as relações locais (proximidade entre pontos próximos) e globais (estrutura dos clusters) sejam mantidas.
- **Representação Final:** Os dados são projetados no espaço de menor dimensionalidade (normalmente 2D ou 3D), criando uma representação compacta que pode ser utilizada para agrupamentos (como no HDBSCAN) ou visualizações interativas.

O UMAP utiliza várias técnicas essenciais para a redução de dimensionalidade. A primeira delas é o ***k-Nearest Neighbors (k-NN)***, que identifica os pontos mais próximos

¹ Encontrado em: https://umap-learn.readthedocs.io/en/latest/basic_usage.html

no espaço de alta dimensionalidade para criar um grafo local. A escolha do número de vizinhos (k) é um parâmetro crítico, pois influencia o equilíbrio entre a preservação das relações locais e globais. Em seguida, o algoritmo utiliza um otimizador de gradiente para ajustar o grafo em baixa dimensionalidade, minimizando as diferenças estruturais entre os grafos original e projetado.

Além disso, o *UMAP* aplica geometria algébrica, assumindo que os dados em alta dimensionalidade estão distribuídos em um espaço chamado **manifold**. Ele projeta esses dados em uma variedade de baixa dimensionalidade, preservando tanto as relações locais quanto as estruturas globais dos dados. Essa combinação de técnicas permite ao *UMAP* gerar representações compactas e de alta qualidade, mantendo a integridade semântica dos dados.

Portanto, ele é ideal para o pipeline proposto no *SIMCC* devido à sua capacidade de lidar eficientemente com embeddings densos e multilingues gerados pelos modelos de linguagem. Sua habilidade em preservar estruturas semânticas tanto locais quanto globais permite que clusters de tópicos sejam mais coerentes e representativos. Além disso, a combinação de escalabilidade, flexibilidade e qualidade de projeção o torna uma escolha natural para visualizações interativas, como no caso do *WizMap*, onde os tópicos serão exibidos como clusters organizados.

Dessa forma, ele integra perfeitamente ao restante do pipeline, fornecendo representações compactas e confiáveis que servem como base para o agrupamento de tópicos com o *HDBSCAN* e a análise semântica enriquecida pelo *GPT-4*. Essa combinação garante que a redução de dimensionalidade não seja apenas uma etapa intermediária, mas um componente estratégico para melhorar a eficácia e eficiência da modelagem de tópicos.

5.2.4 Agrupamento de Tópicos (*Clustering*)

O agrupamento de tópicos também é uma etapa essencial no pipeline, onde textos com características semelhantes são agrupados para formar clusters temáticos. Esses clusters são agrupamentos naturais de dados que compartilham características semelhantes. No contexto de modelagem de tópicos, cada cluster representa um grupo de textos que abordam temas relacionados, com base em suas representações numéricas (embeddings). No contexto do projeto, utilizamos o ***Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)*** devido à sua capacidade de lidar com dados complexos, densidades variáveis e ruídos, oferecendo agrupamentos precisos e adaptáveis.

O HDBSCAN é uma extensão hierárquica do algoritmo ***Density-Based Spatial Clustering of Applications with Noise (DBSCAN)***, projetado para detectar clusters em dados de densidades variáveis sem exigir que o número de clusters seja definido previamente. O funcionamento do HDBSCAN envolve várias etapas:

- **Construção do Grafo de Distâncias:** O HDBSCAN inicia criando um grafo onde os pontos de dados são conectados com base na distância entre eles, calculada a partir dos embeddings. Para isso, utiliza a distância métrica escolhida (como euclidiana ou coseno).
- **Árvore de Alcance Mínimo (*Minimum Spanning Tree*):** A partir do grafo, o algoritmo constrói uma árvore de alcance mínimo, que conecta todos os pontos com a menor soma de pesos (distâncias). Isso cria uma representação hierárquica das relações entre os pontos.
- **Densidade e Ruído:** A densidade de cada cluster é avaliada com base no número de pontos próximos e suas distâncias. Pontos que não atingem a densidade mínima são considerados ruídos e não são atribuídos a nenhum cluster.
- **Cluster Condensado:** A árvore é podada iterativamente para remover arestas que representam ruídos, revelando os clusters restantes. Esses agrupamentos finais são obtidos com base em densidades locais, permitindo identificar clusters de diferentes formas e tamanhos.
- **Soft Clustering (Agrupamento Suave):** O HDBSCAN fornece uma saída de probabilidade que indica a força de associação de cada ponto ao cluster, em vez de uma associação rígida. Isso é útil para lidar com pontos que podem estar na fronteira entre clusters.

O HDBSCAN se destaca por sua capacidade de lidar com densidades variadas e formatos complexos de clusters, ao contrário de técnicas como o *k-means*, que assumem densidade uniforme e formas esféricas. Essa adaptabilidade é crucial para dados heterogêneos, como textos acadêmicos. Outra vantagem significativa é sua robustez ao ruído, já que o HDBSCAN identifica pontos que não se encaixam em nenhum cluster e os trata como ruídos, em vez de forçá-los a se agrupar, o que melhora a qualidade geral dos tópicos formados. Além disso, o algoritmo não requer que o número de clusters seja pré-definido, determinando-o automaticamente com base nos dados, eliminando a necessidade de suposições iniciais que podem ser imprecisas. O *soft clustering* é outra característica marcante, pois fornece uma probabilidade de associação para cada ponto, permitindo interpretações mais flexíveis e precisas, especialmente em cenários onde os limites entre *clusters* não são claros. Por fim, o HDBSCAN é altamente escalável, capaz de lidar com grandes volumes de dados textuais, como os do SIMCC, de forma eficiente e robusta.

Comparado a outras técnicas, o HDBSCAN supera o *k-means* MacQueen (1967), que exige o número de clusters como entrada e é limitado a formas esféricas, e o DBSCAN, que não é eficaz em densidades variadas ou em dados de alta dimensionalidade. Além disso,

é mais flexível e escalável do que o *Agglomerative Clustering*, que é adequado para pequenos conjuntos de dados, mas menos eficiente em grandes volumes. Essas características tornam o HDBSCAN uma escolha superior para a modelagem de tópicos em corpora complexos e extensos, como os encontrados na plataforma SIMCC.

Estudos recentes, como os de Gana *et al.* (2024) e Jung *et al.* (2024), destacam o sucesso do HDBSCAN em tarefas de modelagem de tópicos, especialmente quando combinado com UMAP para reduzir a dimensionalidade antes do agrupamento. Essa combinação oferece clusters mais coesos e representativos.

5.2.5 Rotulagem de Tópicos

No processo de modelagem temática, a etapa de rotulagem desempenha um papel essencial, garantindo que os clusters sejam identificados com rótulos representativos e facilmente interpretáveis. Para este projeto, o GPT-4 foi escolhido como ferramenta principal para enriquecer semanticamente os rótulos, aproveitando sua habilidade avançada de gerar textos contextualmente precisos e alinhados com o significado subjacente dos tópicos.

Uma vez que os clusters de tópicos são formados pelo BERTopic, suas palavras-chave representativas (identificadas por métodos como o c-TF-IDF) são enviadas para a API da OpenAI. O GPT-4 interpreta essas palavras-chave e o contexto geral do cluster, gerando rótulos mais descritivos, claros e semanticamente ricos. Esse processo transcende a rotulagem padrão automatizada do BERTopic, que utiliza apenas as palavras mais frequentes ou relevantes de cada cluster para gerar nomes, frequentemente resultando em rótulos genéricos ou pouco informativos.

O GPT-4 oferece vantagens significativas para a rotulagem de tópicos, destacando-se pela sua capacidade de contextualização e coerência. Graças ao treinamento em um extenso corpus textual, o GPT-4 consegue interpretar não apenas as palavras-chave de um cluster, mas também as relações contextuais entre elas. Isso permite a criação de rótulos mais detalhados e representativos, capturando o significado subjacente dos tópicos de maneira precisa. Essa capacidade é especialmente útil para lidar com temas complexos, onde as conexões contextuais são fundamentais para a interpretação.

Outro ponto forte do GPT-4 é sua habilidade de enriquecer semanticamente os rótulos, indo além das limitações de abordagens estatísticas tradicionais, como o c-TF-IDF utilizado no BERTopic. Em clusters interdisciplinares, o GPT-4 é capaz de identificar conexões que podem não ser evidentes apenas pelas palavras-chave, oferecendo rótulos mais ricos e informativos. Além disso, sua flexibilidade multilingue é crucial para projetos como o SIMCC, permitindo a geração de rótulos consistentes e semânticos em idiomas como português, inglês e espanhol. Essa capacidade de traduzir conceitos entre idiomas de

forma precisa reforça sua aplicação em contextos multilíngues. Por fim, o GPT-4 possibilita rotulagem personalizada, permitindo ajustes nas instruções para gerar rótulos detalhados ou simplificados, dependendo da necessidade do projeto ou do público-alvo, tornando o processo ainda mais versátil e adaptado aos objetivos específicos da análise.

A técnica padrão de rotulagem do BERTopic utiliza algoritmos como c-TF-IDF para identificar as palavras mais relevantes de cada cluster. Embora eficiente, essa abordagem é limitada à seleção automática de palavras-chave, o que muitas vezes resulta em rótulos genéricos ou pouco descritivos, especialmente em tópicos complexos ou com sobreposição temática. O BERTopic não considera o contexto semântico amplo ou as relações entre palavras no cluster, tornando os rótulos menos precisos em temas interdisciplinares.

Por outro lado, o GPT-4 supera essas limitações ao interpretar os clusters de maneira contextual e ao enriquecer os rótulos com descrições detalhadas. Por exemplo, se um cluster contém palavras como "transformers", "BERT" e "PLN", o BERTopic poderia gerar um rótulo genérico como "Modelos de Transformadores". Já o GPT-4 poderia contextualizar essas palavras para criar um rótulo mais informativo, como "Aplicações de Modelos de Transformadores em Processamento de Linguagem Natural".

A integração do GPT-4 no pipeline para rotulagem de tópicos representa um avanço significativo em termos de qualidade e usabilidade dos rótulos gerados. Comparado à técnica padrão do BERTopic, ele oferece uma compreensão mais profunda e contextual dos clusters, resultando em rótulos mais representativos e úteis para análise. Essa abordagem é especialmente vantajosa para a plataforma SIMCC, onde a clareza e a precisão semântica dos tópicos são cruciais para facilitar a exploração e a compreensão das produções científicas indexadas.

5.2.6 Visualização Interativa (*WizMap*)

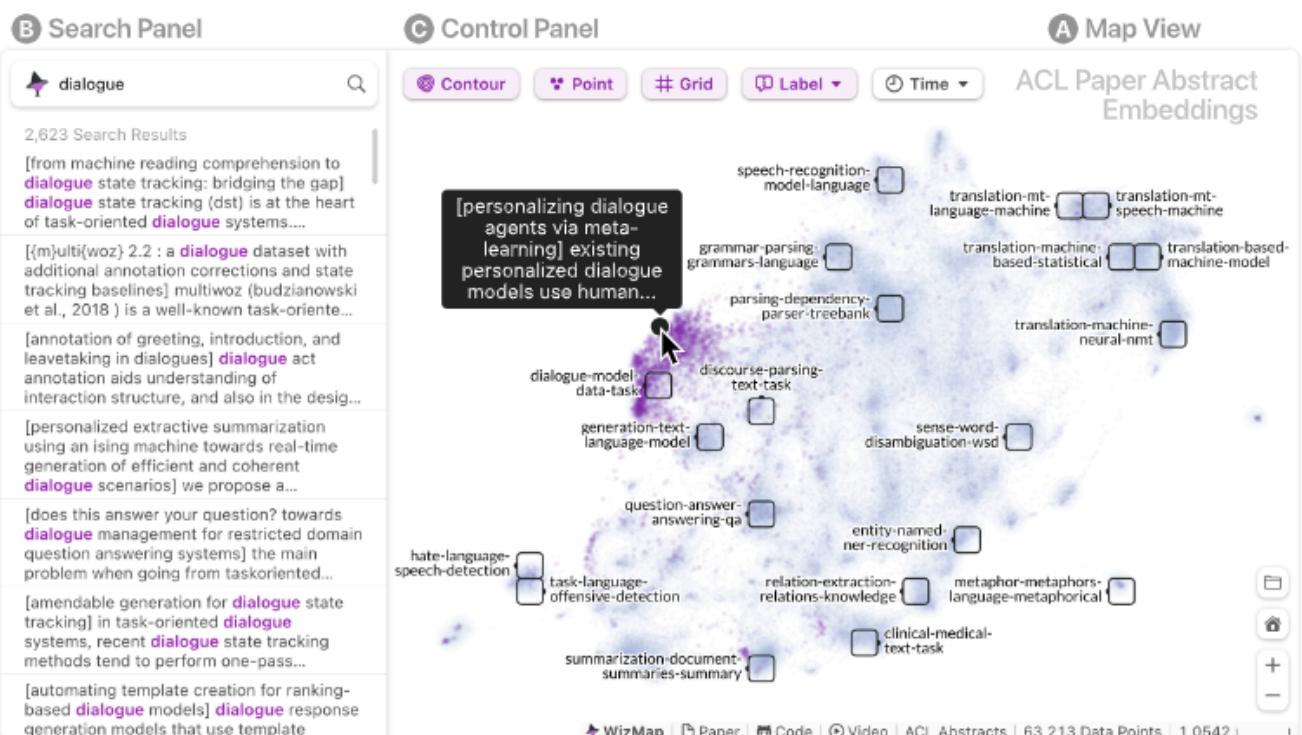
O *WizMap* é uma ferramenta avançada de visualização interativa projetada para explorar e interpretar grandes volumes de *embeddings* em espaços de alta dimensionalidade, como descrito por Wang, Hohman e Chau (2023). Essa solução é particularmente eficaz para lidar com os desafios apresentados pelo volume e pela complexidade de *embeddings* gerados por modelos de aprendizado de máquina, permitindo que pesquisadores compreendam a estrutura global e local desses dados. O *WizMap* adota uma abordagem de visualização inspirada em mapas, integrando técnicas como o uso de contornos de densidade e resumos adaptativos de *embeddings* em múltiplas resoluções.

No contexto do SIMCC, a aplicação do *WizMap* oferece uma oportunidade única para tornar a exploração de tópicos emergentes mais intuitiva e acessível. Após a modelagem de tópicos e a geração de *embeddings* com o BERTopic, combinada com técnicas de redução

de dimensionalidade, como o UMAP, o *WizMap*² organiza os tópicos em uma interface visual. Essa interface utiliza *quadtree* para segmentar os dados e summarizar os tópicos em diferentes níveis de granularidade, permitindo que os usuários explorem *clusters* de tópicos em detalhes ou obtenham uma visão geral, ajustando a profundidade da análise conforme necessário.

Comparado a métodos tradicionais, como *scatter plots* estáticos ou projeções simples, o *WizMap* se destaca pela capacidade de manipular milhões de pontos em navegadores sem a necessidade de infraestrutura pesada, graças ao uso de tecnologias como *WebGL*³ e *Web Workers*⁴. Isso torna a ferramenta altamente escalável e adequada para o *SIMCC*, que lida com grandes volumes de dados acadêmicos. Além disso, a funcionalidade de buscar palavras-chave diretamente no painel de pesquisa e a possibilidade de observar mudanças temporais nos *embeddings* ampliam significativamente as capacidades analíticas, conforme ilustrado por Wang, Hohman e Chau (2023). Essa abordagem aprimora a experiência do usuário, promovendo insights mais profundos e contribuindo para a eficiência na análise de publicações científicas.

Figura 6 – Diagrama esquemático detalhado da ferramenta Wizmap para visualização de dados.



Fonte: Wang, Hohman e Chau (2023, Traduzido, p. 1)

² Encontrado em: <https://poloclub.github.io/wizmap/>

³ Encontrado em: <https://get.webgl.org/>

⁴ Encontrado em: https://developer.mozilla.org/en-US/docs/Web/API/Web_Workers_API/Using_web_workers

6 RESULTADOS ESPERADOS

O estudo proposto espera alcançar uma série de resultados significativos que visam melhorar a análise e a gestão de publicações científicas na plataforma SIMCC. Em primeiro lugar, espera-se que a integração do BERTopic com o GPT-4 permita buscas mais precisas e contextuais, superando as limitações das abordagens tradicionais baseadas em palavras-chave exatas. Com essa combinação, os usuários da plataforma devem obter resultados mais relevantes e completos, especialmente em áreas interdisciplinares, onde sinônimos e variações terminológicas são frequentes.

Outro resultado esperado é a capacidade do pipeline proposto de identificar tópicos emergentes de forma mais eficiente. Ao capturar nuances semânticas que escapam das abordagens tradicionais, espera-se que os tópicos gerados sejam mais coesos e representativos, facilitando a descoberta de tendências e áreas de pesquisa em ascensão. Essa melhoria na identificação de tópicos emergentes deve contribuir para uma análise mais aprofundada e estratégica das publicações científicas.

A implementação do WizMap como ferramenta de visualização interativa também é um dos principais resultados esperados. Essa abordagem deve proporcionar uma experiência de navegação mais intuitiva e visual dos tópicos, permitindo que os usuários explorem os clusters de tópicos em diferentes níveis de granularidade. A visualização interativa deve facilitar a compreensão das relações entre os diferentes temas, tornando a análise mais acessível e enriquecedora.

Além disso, espera-se que o pipeline proposto reduza significativamente o tempo necessário para a análise de grandes volumes de publicações científicas. A automação e a otimização das etapas de modelagem de tópicos e rotulagem semântica devem tornar o processo mais eficiente e acessível aos pesquisadores, permitindo que eles se concentrem em insights e descobertas em vez de tarefas manuais e demoradas.

A rotulagem semântica realizada pelo GPT-4 também deve apresentar melhorias significativas. Espera-se que os rótulos gerados sejam mais descritivos e contextualizados, superando as limitações da rotulagem automática tradicional. Esses rótulos mais ricos e informativos devem facilitar a interpretação e a exploração dos tópicos, contribuindo para uma análise mais clara e precisa.

A validação do pipeline será realizada por meio de métricas de avaliação, como o Topic Coherence Score, que mede a coerência dos tópicos identificados. Espera-se que os tópicos gerados apresentem alta coerência semântica, demonstrando a eficácia do pipeline na análise de publicações científicas. Essa validação deve confirmar que a solução proposta é robusta e capaz de lidar com os desafios da análise de grandes volumes de dados textuais.

Por fim, espera-se que a solução proposta contribua significativamente para a gestão do conhecimento acadêmico. Ao oferecer uma ferramenta robusta e escalável para a análise e categorização de publicações científicas, a plataforma SIMCC deve ser aprimorada, proporcionando uma experiência mais eficiente e enriquecedora para os usuários. Além disso, os resultados obtidos serão documentados e publicados em eventos científicos e periódicos relevantes, contribuindo para o avanço do conhecimento na área de Ciência da Informação e Processamento de Linguagem Natural. A disseminação dos resultados deve inspirar novas pesquisas e aplicações práticas em diferentes contextos acadêmicos e industriais.

Em resumo, os resultados esperados indicam que a integração do BERTopic com o GPT-4 e a visualização interativa com o WizMap proporcionarão uma solução inovadora e eficaz para a análise de publicações científicas. Essa abordagem deve superar as limitações das abordagens tradicionais, oferecendo uma ferramenta valiosa para a gestão do conhecimento acadêmico e contribuindo para o avanço da ciência e tecnologia.

REFERÊNCIAS

- ACHIAM, J. *et al.* Gpts are gpts: An early look at the labor market impact potential of large language models. **arXiv preprint**, 2023. Disponível em: <https://arxiv.org/abs/2304.02142>. Citado na página 24.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, MIT Press, v. 3, n. Jan, p. 993–1022, 2003. ISSN 1532-4435. Disponível em: <http://jmlr.org/papers/v3/blei03a.html>. Citado na página 18.
- CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J. *et al.* (Ed.). **Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 160–172. Citado nas páginas 21 e 22.
- 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 nas páginas 12, 18, 19, 20 e 23.
- DEERWESTER, S. *et al.* Indexing by latent semantic analysis. **Journal of the American Society for Information Science**, v. 41, n. 6, p. 391–407, 1990. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9>. Citado na página 18.
- DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: <https://arxiv.org/abs/1810.04805>. Citado nas páginas 12, 16, 19 e 23.
- 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.
- DRESCH, A.; LACERDA, D.; ANTUNES, J. **Design Science Research: Método de Pesquisa para Avanço da Ciência e Tecnologia**. [S.l.]: Editora FGV, 2015. ISBN 978-85-8260-298-0. Citado na página 30.
- 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 nas páginas 12, 14, 15 e 24.
- GANA, B. *et al.* Leveraging llms for efficient topic reviews. **Applied Sciences**, v. 14, n. 17, 2024. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/14/17/7675>. Citado nas páginas 22, 23, 27, 28 e 40.
- GEORGE, L.; SUMATHY, P. An integrated clustering and bert framework for improved topic modeling. **International Journal of Information Technology**, v. 15, n. 4, p. 2187–2195, 2023. Disponível em: <https://doi.org/10.1007/s41870-023-01268-w>. Citado nas páginas 18 e 20.

- 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, 13, 17, 19, 22, 24 e 26.
- HOFMANN, T. Probabilistic latent semantic indexing. In: **Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. Berkeley, CA, USA: ACM Press, 1999. p. 50–57. ISBN 1-58113-096-1. Citado na página 18.
- HOFMANN, T. Hofmann, t.: Unsupervised learning by probabilistic latent semantic analysis. *machine learning* 42(1-2), 177-196. **Machine Learning**, v. 42, p. 177–196, 01 2001. Citado na página 18.
- JORGE, E. M. F. *et al.* Recuperando especialistas em energias renováveis por meio de taxonomia facetada e técnicas de processamento de linguagem natural: um experimento de mineração de dados acadêmicos aplicados por pesquisadores das universidades estaduais da bahia. **Informação & Informação**, v. 30, n. 2, p. 242–268, 2025. Citado nas páginas 33 e 34.
- JUNG, H. S. *et al.* Expansive data, extensive model: Investigating discussion topics around llm through unsupervised machine learning in academic papers and news. **PLOS ONE**, Public Library of Science, v. 19, n. 5, p. 1–18, 05 2024. Disponível em: <https://doi.org/10.1371/journal.pone.0304680>. Citado nas páginas 17, 20, 23, 26, 27 e 40.
- KIM, K.; KOGLER, D. F.; MALIPHOL, S. Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTTopic. **Palgrave Communications**, v. 11, n. 1, p. 1–15, December 2024. Disponível em: https://ideas.repec.org/a/pal/palcom/v11y2024i1d10.1057_s41599-024-03044-y.html. Citado na página 14.
- KOZLOWSKI, D.; PRADIER, C.; BENZ, P. **Generative AI for automatic topic labelling.** 2024. Disponível em: <https://arxiv.org/abs/2408.07003>. Citado nas páginas 27 e 28.
- MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008. Disponível em: <http://jmlr.org/papers/v9/vandermaaten08a.html>. Citado na página 20.
- MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability**. Berkeley: University of California Press, 1967. (*Statistics*, v. 1), p. 281–297. Disponível em: <http://projecteuclid.org/euclid.bsmsp/1200512992>. Citado na página 39.
- MCINNES, L.; HEALY, J.; MELVILLE, J. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.** 2018. Disponível em: <https://arxiv.org/abs/1802.03426>. Citado nas páginas 20, 21 e 37.
- MENG, F. *et al.* Demand-side energy management reimagined: A comprehensive literature analysis leveraging large language models. **Energy**, v. 291, p. 130303, 2024. ISSN 0360-5442. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360544224000744>. Citado nas páginas 23, 24, 27 e 28.

- MIFRAH, S.; BENLAHMAR, E. H. Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. **International Journal of Advanced Trends in Computer Science and Engineering**, 08 2020. Citado na página 18.
- MIKOLOV, T. *et al.* **Efficient Estimation of Word Representations in Vector Space**. 2013. Disponível em: <https://arxiv.org/abs/1301.3781>. Citado na página 15.
- 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.
- OPENAI *et al.* **GPT-4 Technical Report**. 2024. Disponível em: <https://arxiv.org/abs/2303.08774>. Citado na página 24.
- OUYANG, L. *et al.* Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, v. 35, p. 27730–27744, 2022. Disponível em: <https://arxiv.org/abs/2203.02155>. Citado na página 24.
- PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://aclanthology.org/D14-1162>. Citado na página 15.
- POLYZOS, E.; WANG, F. Twitter and market efficiency in energy markets: Evidence using lda clustered topic extraction. **Energy Economics**, v. 114, p. 106264, 2022. ISSN 0140-9883. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0140988322004017>. Citado na página 18.
- RADFORD, A.; NARASIMHAN, K. Improving language understanding by generative pre-training. In: . [s.n.], 2018. Disponível em: <https://api.semanticscholar.org/CorpusID:49313245>. Citado nas páginas 16, 19, 23 e 24.
- REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019. Disponível em: <https://arxiv.org/abs/1908.10084>. Citado nas páginas 17, 19 e 26.
- SANTOS, M. S. d. *et al.* Solução para mapeamento e consulta das competências dos pesquisadores: uma arquitetura para extração, integração e consultas de informações acadêmicas. **Cadernos de Prospecção**, v. 17, n. 2, p. 671–688, abr. 2024. Disponível em: <https://periodicos.ufba.br/index.php/nit/article/view/56670>. Citado nas páginas 33, 34 e 35.
- VASWANI, A. *et al.* **Attention Is All You Need**. 2017. Disponível em: <https://arxiv.org/abs/1706.03762>. Citado nas páginas 12, 15, 16, 19, 23 e 24.
- 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 nas páginas 13, 35, 41 e 42.
- 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 nas páginas 13 e 17.

WIJANTO, M. C.; WIDIASTUTI, I.; YONG, H.-S. Topic modeling for scientific articles: Exploring optimal hyperparameter tuning in bert. **International Journal on Advanced Science, Engineering and Information Technology**, v. 14, n. 3, p. 912–919, Jun. 2024. Disponível em: <https://ijaseit.insightsociety.org/index.php/ijaseit/article/view/19347>. Citado nas páginas 26 e 28.

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 nas páginas 12, 14 e 18.