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 PARA ANÁLISE DE PUBLICAÇÕES CIENTÍFICAS: UM ESTUDO DE CASO NO OBSERVATÓRIO DE DADOS PÚBLICOS DE CIÊNCIA E TECNOLOGIA DA BAHIA

SALVADOR
2025

JEOSTON ARAUJO DA CRUZ JÚNIOR

UM PIPELINE COM BERTOPIC PARA ANÁLISE DE PUBLICAÇÕES CIENTÍFICAS: UM ESTUDO DE CASO NO OBSERVATÓRIO DE DADOS PÚBLICOS DE CIÊNCIA E TECNOLOGIA DA BAHIA

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

Orientador: Prof. Dr. Eduardo Manuel de Freitas Jorge

SALVADOR
2025

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 PARA ANÁLISE DE PUBLICAÇÕES CIENTÍFICAS: UM ESTUDO DE CASO NO OBSERVATÓRIO DE DADOS PÚBLICOS DE CIÊNCIA E TECNOLOGIA DA BAHIA

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

O crescente e contínuo volume de publicações científicas representa um desafio significativo para a organização, exploração e descoberta de padrões temáticos, especialmente quando as abordagens de busca se limitam a palavras-chave. Para mitigar essa lacuna e aprimorar a capacidade de análise do conhecimento, este estudo propõe o desenvolvimento de uma **solução de mapeamento interativo de conhecimento**. Esta abordagem visa transformar a maneira como pesquisadores exploram acervos científicos, revelando conexões semânticas e tópicos emergentes de forma intuitiva. A proposta será aplicada ao acervo do **Observatório de dados públicos de ciência e tecnologia da Bahia**, que centraliza dados de produção acadêmica de fontes como Currículos Lattes, Sucupira e OpenAlex. O artefato tecnológico consiste em um *pipeline* computacional que integra a modelagem de tópicos via **BERTopic** (utilizando *embeddings* contextuais, UMAP e HDBSCAN) com o refinamento de rótulos por meio da **MMR** (*Maximal Marginal Relevance*) para aumentar a interpretabilidade. Os resultados do mapeamento são então integrados à ferramenta de visualização interativa **WizMap**, permitindo a exploração visual e dinâmica da estrutura do conhecimento. Esta metodologia busca oferecer uma compreensão mais estruturada e acessível do acervo científico do Observatório, facilitando a identificação de tendências e a navegação por áreas de pesquisa.

Palavras-chave: Processamento de Linguagem Natural; Inteligência Artificial; Modelagem de Tópicos; BERTopic; Visualização de Dados; WizMap; Observatório de C&T da Bahia.

ABSTRACT

The growing and continuous volume of scientific publications represents a significant challenge for the organization, exploration, and discovery of thematic patterns, especially when search approaches are limited to keywords. To mitigate this gap and enhance knowledge analysis capabilities, this study proposes the development of an **interactive knowledge mapping solution**. This approach aims to transform the way researchers explore scientific collections, revealing semantic connections and emerging topics in an intuitive manner. The proposal will be applied to the collection of the **Observatory of Public Data on Science and Technology of Bahia**, which centralizes academic production data from sources such as Lattes Curricula, Sucupira, and OpenAlex. The technological artifact consists of a computational *pipeline* that integrates topic modeling through **BERTopic** (using contextual *embeddings*, UMAP, and HDBSCAN) with label refinement via **MMR** (*Maximal Marginal Relevance*) to enhance interpretability. The mapping results are then integrated into the interactive visualization tool **WizMap**, enabling dynamic and visual exploration of the knowledge structure. This methodology seeks to provide a more structured and accessible understanding of the Observatory's scientific collection, facilitating the identification of trends and navigation across research areas.

Keywords: Natural Language Processing; Artificial Intelligence; Topic Modeling; BERTopic; Data Visualization; WizMap; Science and Technology Observatory of Bahia.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O cenário da pesquisa científica registra um crescimento na produção acadêmica nas últimas décadas, gerando um volume de dados que impõe dificuldades aos métodos convencionais de organização e análise. Para navegar nesse conjunto de informações, pesquisadores confiam em plataformas de busca, como *Web of Science*[1], *Scopus*[2] e *IEEE Xplore*[3], baseando-se majoritariamente em palavras-chave. Contudo, essa abordagem de recuperação de informações é limitada pela ambiguidade e pela diversidade do léxico científico, o que pode dificultar o retorno de resultados completos e a identificação de tendências na literatura (GALLI *et al.*, 2024).

Segundo Datchanamoorthy, S e B (2023), a complexidade desses acervos e a necessidade de uma análise mais profunda têm impulsionado a aplicação de técnicas avançadas de Processamento de Linguagem Natural (PLN). Essa integração entre ciência da informação, Inteligência Artificial (IA) e linguística computacional, auxilia na construção de soluções para a gestão do conhecimento acadêmico. Estudos como os de Mohammadi e Karami (2020) e Xie *et al.* (2020), que analisaram tendências de pesquisa em *Big Data* por meio de mineração de texto, destacam a relevância da integração de técnicas de modelagem de tópicos com modelos baseados em transformadores.

A arquitetura de Transformadores, apresentada por Vaswani *et al.* (2017), introduziu o mecanismo de autoatenção (*self-attention*) no campo da PLN. Modelos subsequentes, como o *Bidirectional Encoder Representations from Transformers* (BERT)[4], proposto por Devlin *et al.* (2019), permitiram capturar relações contextuais em textos. A partir dessa base, surgiram os *embeddings*, representações numéricas que codificam o significado semântico de palavras e frases, abordando as limitações de modelos tradicionais de *bag-of-words* e de modelagem de tópicos como o *Latent Dirichlet Allocation* (LDA) (GALLI *et al.*, 2024).

Nesse contexto, a técnica de *Bidirectional Encoder Representations Transformers for Topic Modeling* (BERTopic)[5], proposto por Grootendorst (2022), surge como uma abordagem atualizada. Seu diferencial reside na utilização dos *embeddings* contextuais de modelos como o BERT para a modelagem de tópicos, permitindo a identificação de agrupamentos e o tratamento de nuances semânticas em textos interdisciplinares.

Esta pesquisa concentra-se no desenvolvimento de um *pipeline* computacional para o mapeamento interativo de publicações científicas. A proposta central é construir um

[1] Disponível em: https://access.clarivate.com/login?app=wos.
[2] Disponível em: https://www.scopus.com/home.uri.
[3] Disponível em: https://ieeexplore.ieee.org/.
[4] Disponível em: https://huggingface.co/docs/transformers/model_doc/bert.
[5] Disponível em: https://github.com/MaartenGr/BERTopic.

artefato que combina a modelagem de tópicos do BERTopic (GROOTENDORST, 2022) com a ferramenta de visualização *Knowledge Map Visualization Tool* (WizMap)[6] (WANG; HOHMAN; CHAU, 2023). O artefato computacional é aplicado ao acervo do Observatório de dados públicos de ciência e tecnologia da Bahia, esse Observatório coleta informações de fontes como Currículos Lattes, Plataforma Sucupira e *OpenAlex*, e tem um papel fundamental na gestão do conhecimento científico regional.

A metodologia *Design Science Research* (DSR) é adotada como a estrutura principal deste estudo, orientando a criação do *pipeline* como artefato principal resultado do trabalho. O objetivo é transformar a base de dados textual do Observatório em um mapa de conhecimento navegável. Nessa solução, o BERTopic é empregado para extrair os padrões temáticos e o WizMap é utilizado para a exploração visual e interativa desses tópicos. Essa integração visa permitir a identificação de temas e a compreensão da estrutura dos dados científicos da plataforma.

O trabalho está organizado da seguinte forma: O Capítulo 2 estabelece o Referencial Teórico, abordando os conceitos de Ciência da Informação, a arquitetura dos Transformadores e as técnicas de modelagem de tópicos. A seguir, o Capítulo 3 analisa os Trabalhos Correlatos, contextualizando esta pesquisa frente ao estado da arte. O Capítulo 4 detalha a Metodologia (DSR). O Capítulo 5,descreve o Projeto de Desenvolvimento e a arquitetura do *pipeline*. Por fim, o Capítulo 6 discute os Resultados e a validação da solução proposta.

[6] Disponível em: https://github.com/poloclub/wizmap.

2 REFERENCIAL TEÓRICO

Este capítulo estabelece a fundamentação teórica que sustenta o desenvolvimento do *pipeline* proposto. A revisão da literatura aborda os pilares conceituais necessários para a análise de publicações científicas e para a construção do artefato de mapeamento interativo.

Iniciamos pela Ciência da Informação, contextualizando o desafio central do crescimento exponencial da produção científica e as limitações das abordagens tradicionais de recuperação. Em seguida, aprofundamos nos fundamentos técnicos do Processamento de Linguagem Natural (PLN), explorando a arquitetura dos Transformadores e o conceito de *embeddings*, que são a base dos modelos modernos.

Posteriormente, detalhamos as Abordagens de Modelagem de Tópicos, comparando métodos tradicionais, como o LDA, com a arquitetura moderna do BERTopic, justificando sua escolha. Por fim, discutimos a importância da Visualização da Informação como ferramenta analítica, fundamentando a integração da ferramenta WizMap como etapa final do artefato.

2.1 Ciência da Informação e Análise de Publicações Científicas

O crescimento da produção científica global nas últimas décadas, impulsionado pela acessibilidade tecnológica e pela colaboração interdisciplinar, apresenta desafios para a Ciência da Informação. Conforme Kim, Kogler e Maliphol (2024), o volume de publicações dificulta a atualização dos pesquisadores e a identificação de áreas emergentes do conhecimento. Os autores reforçam essa problemática no resumo de seu trabalho:

> A produção científica global está se expandindo exponencialmente, o que, por sua vez, exige uma melhor compreensão da ciência da ciência e, especialmente, de como as fronteiras dos campos científicos se expandem através de processos de emergência. Kim, Kogler e Maliphol (2024, Traduzido, p. 1)

Nesse contexto, estratégias tradicionais de busca baseadas em palavras-chave apresentam limitações, pois frequentemente não capturam a variação semântica do léxico científico. Esse fator pode resultar na omissão de trabalhos relevantes, e dificultar o mapeamento do progresso em determinados campos de pesquisa.

Um aspecto que amplia essa complexidade é a diversidade linguística no ambiente científico. Segundo Xie *et al.* (2020), embora o inglês desempenhe papel predominante, uma parcela significativa da produção científica ocorre em outros idiomas. Metodologias

convencionais de análise mostram-se insuficientes para o tratamento multilíngue, o que pode restringir a circulação do conhecimento e a visibilidade de estudos regionais.

> A maioria dos estudos até hoje sobre análise de tópicos tem sido baseada em publicações em língua inglesa e tem dependido fortemente da análise de evolução de tópicos baseada em citações. [...] metodologias baseadas em citações não são adequadas para analisar relações de tópicos de pesquisa multilíngues. Xie *et al.* (2020, Traduzido, p. 1)

Diante desse cenário, técnicas de *Topic Modeling*, em especial aquelas fundamentadas em *embeddings*, têm sido investigadas como alternativas para a análise documental. De acordo com Galli *et al.* (2024), a utilização de representações densas derivadas de modelos como o BERT viabiliza a análise de grandes volumes textuais, permitindo capturar aspectos semânticos que vão além da correspondência lexical exata. Sobre os *embeddings*, os autores definem:

> Um componente essencial para alcançar a compreensão semântica são os *embeddings* — representações numéricas que codificam o significado de palavras ou mesmo de frases — que são fundamentais no PLN para capturar relacionamentos complexos entre palavras e frases usando arquiteturas especiais conhecidas como transformadores. Galli *et al.* (2024, Traduzido, p. 2)

Essa característica favorece a identificação de padrões temáticos em documentos que não compartilham necessariamente o mesmo vocabulário. Métodos como o BERTopic, oferecem uma estrutura metodológica para a extração de tópicos a partir dessas representações vetoriais densas. A literatura indica que a aplicação dessas ferramentas é adequada para textos científicos heterogêneos e multilíngues, devido à capacidade de processar nuances semânticas independentemente do idioma (XIE *et al.*, 2020).

2.2 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) é um campo multidisciplinar, situado na interseção da Inteligência Artificial, da Linguística Computacional e da Ciência da Informação. O objetivo central da área é desenvolver métodos computacionais capazes de processar, analisar, compreender e gerar a linguagem humana — seja em formato de texto ou voz — de maneira funcional (JURAFSKY; MARTIN, 2009)[1].

Historicamente, o PLN fundamentava-se em abordagens estatísticas e regras linguísticas manuais para modelar a linguagem, conforme descrito por (MANNING; SCHUTZE,

[1] Refere-se à obra *Speech and Language Processing*, de Daniel Jurafsky e James H. Martin. É amplamente considerado o livro-texto acadêmico padrão e a referência canônica para o ensino e estudo do Processamento de Linguagem Natural em todo o mundo.

1999)[2]. As técnicas de PLN são projetadas para extrair significado e estrutura de dados textuais, que são inerentemente não estruturados. Este processo envolve uma série de tarefas que variam desde a análise sintática (a estrutura gramatical) até a análise semântica (o significado por trás das palavras). Entre as aplicações comuns estão a classificação de textos, a tradução automática, a sumarização de documentos e a Modelagem de Tópicos (*Topic Modeling*), que é o foco desta pesquisa.

A evolução recente do campo foi impulsionada pelo *Deep Learning* (Aprendizado Profundo), que permitiu a criação de representações vetoriais mais precisas. Como destacam Galli *et al.* (2024), o PLN moderno passou a depender da capacidade de capturar o significado contextual, superando a análise baseada apenas na contagem de palavras. Essa transição para uma abordagem focada na compreensão semântica viabilizou os avanços recentes em modelagem de tópicos.

2.3 A Evolução das Representações Vetoriais em PLN

O progresso na área de PLN tem sido caracterizado pela investigação de representações vetoriais capazes de capturar não apenas a estrutura sintática, mas também os aspectos semânticos e contextuais dos textos. A evolução dessas representações partiu de abordagens estáticas para modelos dinâmicos baseados em contexto.

2.3.1 Embeddings Estáticos: Limitações do Bag-of-Words

As primeiras abordagens de sucesso, como o *Word2Vec* proposto por Mikolov *et al.* (2013)[3] e o *GloVe* proposto por Pennington, Socher e Manning (2014)[4], consolidaram o conceito de *embeddings*. Nesse caso, a operação algébrica subtrai o vetor de "Homem" do vetor de "Rei", isolando o conceito de realeza, e adiciona o vetor de "Mulher", resultando em uma representação vetorial espacialmente próxima à de "Rainha". Isso demonstra que o modelo é capaz de codificar conceitos abstratos, como gênero, através da direção e distância entre os vetores. Estes consistem em vetores em espaços de alta dimensionalidade capazes de representar o significado aproximado de uma palavra.

A contribuição desses modelos foi permitir a quantificação do significado semântico. Em vez de tratar palavras como identificadores discretos (como em uma abordagem *bag-of-words*), os *embeddings* posicionam termos com significados similares próximos uns dos

[2] Refere-se à obra *Foundations of Statistical Natural Language Processing* (Manning e Schütze, 1999), considerada o trabalho seminal que consolidou as abordagens estatísticas como o padrão do PLN antes da ascensão das redes neurais profundas.

[3] O *Word2Vec* (2013) foi seminal por introduzir duas arquiteturas eficientes, *Skip-gram* e *CBOW*, que aprendem vetores de palavras prevendo o contexto em que elas aparecem, baseando-se na hipótese distribucional.

[4] O *GloVe* (2014), ou "Global Vectors", diferencia-se por combinar as estatísticas globais de coocorrência de palavras (como o LSA) com a modelagem baseada em janelas de contexto (como o *Word2Vec*), capturando relações lineares entre palavras.

outros no espaço vetorial. Isso permite que relações semânticas sejam capturadas matematicamente, como no exemplo clássico "Rei - Homem + Mulher ≈ Rainha" (MIKOLOV *et al.*, 2013). Xie *et al.* (2020) na literatura de PLN refere-se a este espaço vetorial como um "espaço semântico".

O aprendizado desses vetores ocorre através do treinamento de redes neurais em tarefas de previsão de contexto, conforme ilustrado na Figura 1. O artigo seminal de Mikolov *et al.* (2013) introduziram duas arquiteturas principais:

1. ***Continuous Bag-of-Words* (CBOW):** A arquitetura prevê a palavra atual (saída) com base em uma janela de palavras do contexto (entrada).
2. ***Skip-gram*:** A arquitetura inverte a lógica e usa a palavra atual (entrada) para prever as palavras do contexto (saída).

É importante destacar que os *embeddings* não são o produto final, mas sim um subproduto do treinamento: os vetores aprendidos na camada oculta da rede (*PROJECTION* na figura) tornam-se a representação semântica da palavra, como indica Mikolov *et al.* (2013, p. 4).

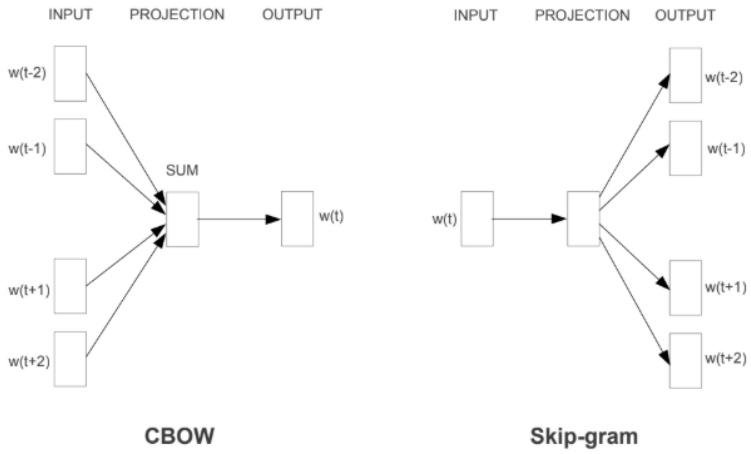

Figura 1 – Arquiteturas CBOW e *Skip-gram*.

Fonte: Mikolov *et al.* (2013, p. 5)

Apesar da utilidade em capturar similaridades lexicais, esses modelos apresentavam a limitação de atribuir um único vetor fixo a cada termo, independentemente do contexto de ocorrência. Por exemplo, a palavra "banco" teria a mesma representação vetorial em "banco financeiro" e "banco da praça". Tal restrição, usualmente referida como o problema da ambiguidade do significado da palavra (*ambiguity of word meaning*), compromete a precisão em tarefas que exigem desambiguação semântica.

2.3.2 A Revolução dos Transformadores e o Mecanismo de Atenção

Uma mudança significativa no paradigma ocorreu com a introdução do modelo de Transformadores (*Transformers*), proposto por Vaswani *et al.* (2017) no artigo seminal *Attention Is All You Need*[5]. Essa arquitetura diferencia-se das *Recurrent Neural Network* (RNN) e convolucionais, por fundamentar-se inteiramente no mecanismo de autoatenção (*self-attention*).

Por meio da autoatenção, o modelo atribui pesos diferenciados a *tokens* em uma sequência, permitindo processar simultaneamente e de forma bidirecional a totalidade do contexto textual. A arquitetura do Transformador, conforme apresentado na Figura 2, segue uma estrutura de codificador-decodificador (*encoder-decoder*). O lado esquerdo do diagrama representa o Codificador, enquanto o lado direito representa o Decodificador.

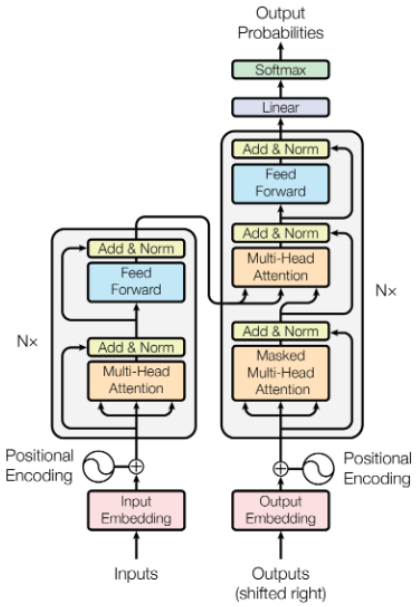

Figura 2 – Arquitetura do modelo Transformador.

Fonte: Vaswani *et al.* (2017, p. 3)

O **Codificador** (*Encoder*) é composto por uma pilha de N camadas idênticas (no artigo original, N=6). Cada camada, por sua vez, é composta por duas subcamadas principais: um mecanismo de autoatenção *multi-head* (*multi-head self-attention*) e uma rede neural *feed-forward* (rede neural de alimentação direta) simples e totalmente conectada. Conexões residuais seguidas de normalização de camada (*Add & Norm*) são aplicadas ao redor de cada subcamada.

[5] Este artigo é considerado um dos trabalhos mais influentes da PLN moderna. Sua principal contribuição foi propor uma arquitetura de rede neural que dispensa totalmente as camadas recorrentes (RNN) e convolucionais, baseando-se unicamente em mecanismos de atenção para modelar dependências globais entre a entrada e a saída (VASWANI *et al.*, 2017, p. 1).

O **Decodificador** (*Decoder*), de forma similar, é uma pilha de N camadas. Além das duas subcamadas presentes no codificador, o decodificador insere uma terceira subcamada, que realiza a atenção *multi-head* sobre a saída da pilha do codificador. Crucialmente, a subcamada de autoatenção do decodificador é "mascarada" (*Masked Multi-Head Attention*). Esse mascaramento é o que garante que a previsão para uma posição i só possa depender das saídas conhecidas em posições anteriores a i, preservando a propriedade autorregressiva do modelo.

Embora a arquitetura completa do Transformador tenha sido projetada para tarefas de transdução de sequência (como a tradução automática), foi a sua pilha de **Codificadores** (*Encoder*) que se mostrou revolucionária para tarefas de *compreensão* de linguagem. A capacidade do Codificador de processar texto de forma bidirecional e gerar representações numéricas ricas em contexto estabeleceu a base para uma nova classe de modelos focados exclusivamente na representação semântica, como será detalhado a seguir.

2.3.3 Embeddings Contextuais: BERT e SBERT

Sobre a base arquitetônica dos Transformadores, foram desenvolvidos os modelos pré-treinados, entre os quais se destaca o BERT, introduzido por Devlin *et al.* (2019). O BERT utiliza a arquitetura do Codificador (*Encoder*) do Transformador para gerar representações de linguagem.

A inovação fundamental do BERT foi o pré-treinamento bidirecional, que diferentemente de abordagens anteriores, como o *Generative Pre-trained Transformer* (GPT) de Radford e Narasimhan (2018), que utilizava um treinamento unidirecional (da esquerda para a direita), o BERT foi projetado para "pré-treinar representações profundamente bidirecionais, condicionando conjuntamente o contexto esquerdo e direito em todas as camadas" como apontam Devlin *et al.* (2019, p. 1, Traduzido).

Para alcançar essa bidirecionalidade sem que o modelo "visse a resposta", Devlin *et al.* (2019) introduziu o objetivo do *Masked Language Model* (MLM)[6]. A Figura 3 ilustra a diferença fundamental entre as arquiteturas de pré-treinamento, mostrando como o BERT é capaz de processar informações de toda a sequência em todas as suas camadas.

[6] O MLM é inspirado na tarefa *Cloze* (TAYLOR, 1953), onde o modelo deve prever palavras que foram omitidas (mascaradas) de uma sentença, usando o contexto de ambas as direções (esquerda e direita) para fazer a previsão (DEVLIN *et al.*, 2019, p. 1).

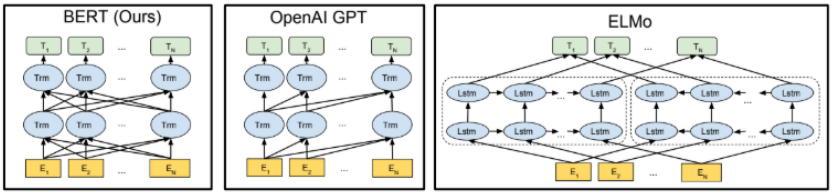

Figura 3 – Diferenças nas arquiteturas de pré-treinamento. BERT é bidirecional, GPT é unidirecional (da esquerda para a direita) e *Embeddings from Language Models* (ELMo).

Fonte: Devlin *et al.* (2019, p. 13)

Apesar do desempenho em tarefas de classificação, a arquitetura do original do BERT apresentou limitações para tarefas de busca de similaridade semântica ou *clustering*. Conforme observador por Reimers e Gurevych (2019), o uso do BERT "requer que ambas as sentenças sejam alimentadas na rede, o que causa um overhead computacional massivo". Uma busca de similaridade em 10.000 sentenças, por exemplo, exigiria cerca de 50 milhões de inferências (aproximadamente 65 horas), tornando-o inviável para grandes bases de dados. Além disso, estudos empíricos demonstraram que usar os *embeddings* "crus" do BERT (seja pela média das saídas ou pelo vetor do *token* '[CLS]') produz resultados insatisfatórios, muitas vezes piores do que os *embeddings* estáticos como o *GloVe*.

Para solucionar essa questão, Reimers e Gurevych (2019) propuseram o *Sentence-BERT* (SBERT). Ele modifica o BERT pré-treinado, adicionando uma operação de *pooling* (sendo a média, *MEAN-strategy*, a mais comum) à saída do BERT para criar um *embedding* de sentença de tamanho fixo.

Crucialmente, o SBERT utiliza redes siamesas[7] para fazer o *fine-tuning* desses *embeddings* de sentença. A Figura 4 ilustra a arquitetura de inferência do SBERT, onde duas sentenças (A e B) são processadas por redes BERT idênticas (com pesos compartilhados), gerando vetores de sentença $\mathbf{u}$ e $\mathbf{v}$. Esses vetores podem, então, ser comparados eficientemente usando uma medida de similaridade, como a similaridade de cosseno (*cosine-similarity*).

[7] Redes siamesas são uma arquitetura onde duas ou mais redes neurais idênticas (com pesos compartilhados) processam entradas diferentes de forma independente. Elas são otimizadas para aprender uma função de similaridade, aproximando os vetores de saída para entradas similares e afastando-os para entradas diferentes.

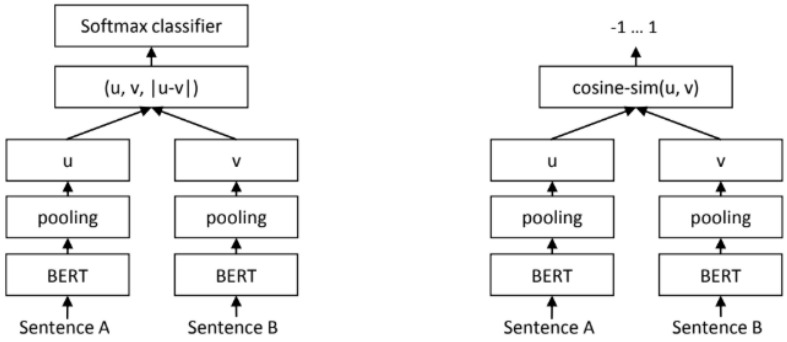

Figura 4 – Arquitetura de inferência do SBERT para computar similaridade.

Fonte: Adaptado de Reimers e Gurevych (2019, p. 3)

Reimers e Gurevych (2019) demonstraram que essa abordagem reduz o custo computacional de encontrar o par mais similar em 10.000 sentenças de 65 horas (com BERT) para cerca de 5 segundos. Essa otimização para similaridade de sentenças permite o uso eficiente desses vetores em cenários multilíngues. A utilização de modelos pré-treinados em múltiplos idiomas (como o *paraphrase-multilingual-MiniLM-L12-v2*[8]) torna-se particularmente relevante, visto que tais modelos produzem *embeddings* semanticamente consistentes mesmo em diferentes idiomas.

2.4 Abordagens Tradicionais de Modelagem de Tópicos

Com o aumento do volume de dados textuais e a necessidade de organizar a informação, a modelagem de tópicos estabeleceu-se como uma técnica relevante na área de PLN. Em termos gerais, consiste de um conjunto de métodos estatísticos cujo objetivo é identificar estruturas semânticas latentes[9], denominadas *tópicos*, em coleções de documentos. Essas técnicas permitem inferir distribuições temáticas que não são explicitamente observáveis, mas que emergem a partir de regularidades no uso do vocabulário.

Entre as abordagens iniciais destacam-se três marcos históricos: a *Latent Semantic Analysis* (LSA), a *Probabilistic Latent Semantic Analysis* (PLSA) e a 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), baseia-se da decomposição de matrizes termo-documento por meio da técnica de *Singular Value Decomposition* (SVD)[10]. Nesse

[8] Disponível em: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.

[9] O termo "latente" significa que os tópicos não são diretamente observáveis, mas sim inferidos estatisticamente a partir dos padrões de coocorrência de palavras no *corpus*.

[10] A SVD é uma técnica de álgebra linear para a decomposição de matrizes que permite encontrar a

método, documentos e termos são projetados em um espaço vetorial de dimensionalidade reduzida, o que permite atenuar ruídos lexicais e capturar relações de similaridade latentes. No entanto, a linearidade da LSA e sua insensibilidade a variações contextuais limitam seu desempenho em cenários onde relações semânticas complexas são determinantes (GEORGE; SUMATHY, 2023; XIE *et al.*, 2020).

Para abordar essas limitações, Hofmann (1999), Hofmann (2001) introduziu 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; 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*[11]. 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; BENLAHMAR, 2020) e eficiência energética (POLYZOS; 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*[12] (saco de palavras), que ignora a ordem e o contexto local das palavras. Segundo George e Sumathy (2023), Xie *et al.* (2020), isso frequentemente conduz a representações semânticas superficiais em textos técnicos ou multilíngues. Datchanamoorthy, S e B (2023) também reitera que a

melhor aproximação de uma matriz por outra de posto inferior, sendo fundamental para a redução de dimensionalidade em espaços vetoriais de termos.

[11] A LDA é um modelo generativo Bayesiano. O uso das distribuições de *Dirichlet* (uma distribuição de probabilidade sobre outras distribuições) permite ao modelo tratar as misturas de tópicos nos documentos e as misturas de palavras nos tópicos como variáveis aleatórias, conferindo maior robustez e melhor generalização.

[12] O *Bag-of-Words* (Saco de Palavras) é um modelo de representação de texto que ignora a ordem e a estrutura gramatical das palavras, tratando um documento apenas como um conjunto (ou multiconjunto) de suas palavras e suas frequências.

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.

> 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; B, 2023, Traduzido).

Essas restrições indicam que, tais métodos podem ser insuficientes para tarefas que exigem compreensão semântica profunda, especialmente em bases textuais heterogêneas onde a ambiguidade lexical é alta, evidenciando a necessidade de abordagens que superem o modelo *bag-of-words*.

2.5 BERTopic: Uma Abordagem Moderna

As limitações das abordagens tradicionais de modelagem de tópicos, especialmente sua dependência do paradigma *bag-of-words* e a falha em capturar o contexto semântico, motivaram o desenvolvimento de novos métodos. Pesquisas recentes indicam a viabilidade de tratar a modelagem de tópicos como uma tarefa de *clustering* (agrupamento) de *embeddings*, notavelmente nos trabalhos que introduziram o *Top2Vec* (ANGELOV, 2020) e em estudos comparativos como o de Sia, Dalmia e Mielke (2020).

Nesse contexto, Grootendorst (2022) propôs o BERTopic, um modelo que estende a abordagem de *clustering* ao introduzir uma variação do *Term Frequency-Inverse Document Frequency* (TF-IDF) baseada em classes para extrair representações de tópicos. O BERTopic funciona como um *pipeline* modular que consiste em três etapas principais: 1) geração de *embeddings* de documentos, 2) *clustering* desses *embeddings* e 3) representação dos tópicos com *Class-based Term Frequency-Inverse Document Frequency* (c-TF-IDF) (GROOTENDORST, 2022, p. 1-2).

A Figura 5 ilustra o fluxo geral dessa arquitetura.

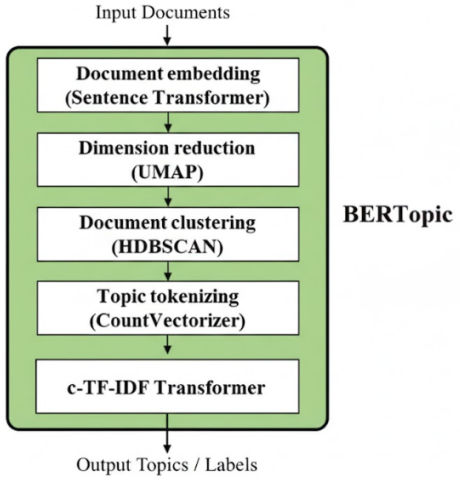

Figura 5 – Diagrama esquemático do *pipeline* BERTopic.

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

Na primeira etapa, *Document embeddings*, os documentos são convertidos em representações vetoriais (embeddings). O BERTopic utiliza nativamente a biblioteca SBERT (*Sentence-BERT*) proposta por Reimers e Gurevych (2019), garantindo que documentos semanticamente similares sejam posicionados próximos no espaço vetorial (GROOTENDORST, 2022, p. 2).

A segunda etapa, *Dimension reduction*, é o *reduction* desses *embeddings* de alta dimensionalidade. Para que os algoritmos de *reduction* funcionem de forma eficiente, é necessário primeiro combater a "maldição da dimensionalidade" (*curse of dimensionality*), um fenômeno onde as distâncias entre os pontos se tornam pouco significativas em espaços com muitas dimensões (GROOTENDORST, 2022, p. 2). Para isso, o BERTopic emprega o *Uniform Manifold Approximation and Projection* (UMAP) (MCINNES; HEALY; MELVILLE, 2018).

Antes de passar para as próximas etapas, precisamos contextualizar o UMAP, que é uma técnica de redução de dimensionalidade que se destaca por preservar tanto a estrutura local quanto a estrutura global dos dados em um espaço de dimensão reduzida[13] (GROOTENDORST, 2022, p. 2-3). A Figura 6 demonstra o impacto de seus dois principais hiperparâmetros.

[13] O UMAP é fundamentado em geometria Riemanniana e topologia algébrica. Ele constrói uma representação topológica dos dados em alta dimensão e busca uma representação em baixa dimensão que tenha uma estrutura topológica o mais equivalente possível (MCINNES; HEALY; MELVILLE, 2018, p. 3-4).

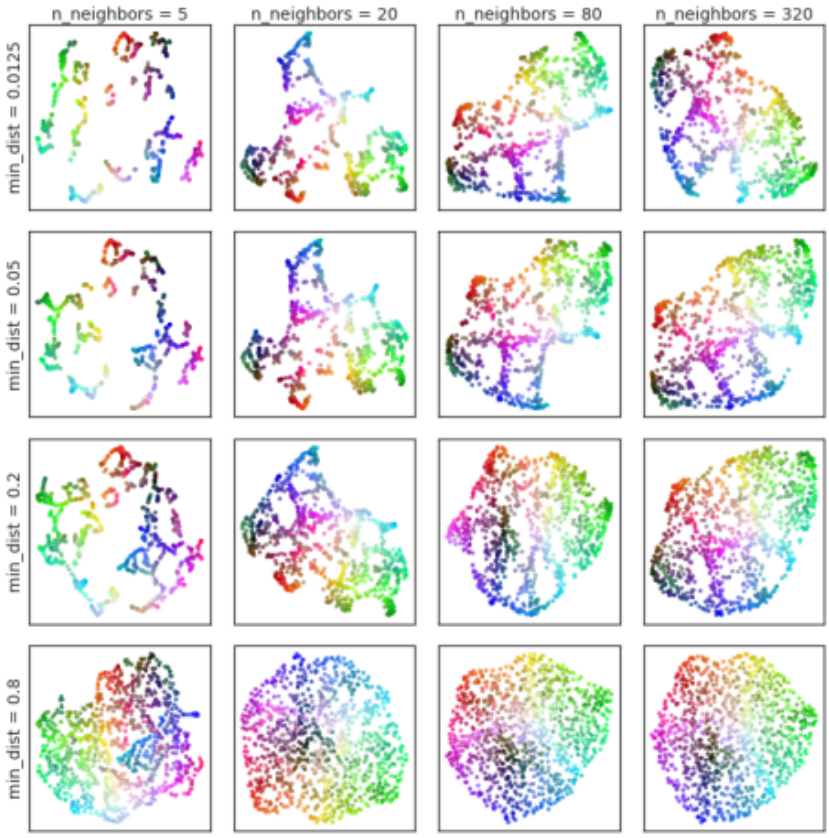

Figura 6 – Diagrama ilustrativo do UMAP, demonstrando a relação entre os hiperparâmetros `n_neighbors` e `min_dist` e a representação visual dos dados.

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

Conforme ilustrado na Figura 6, o parâmetro `n_neighbors` (número de vizinhos) controla o equilíbrio entre a preservação da estrutura global (valores altos) e local (valores baixos). O parâmetro `min_dist` (distância mínima) ajusta a densidade dos agrupamentos, determinando a proximidade entre os pontos no espaço de baixa dimensionalidade.

Com os vetores em dimensão reduzida pelo UMAP, a etapa seguinte da Figura 5, *Document clustering*, é o *clustering* através do *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) (MCINNES; HEALY; MELVILLE, 2018). A escolha deste método justifica-se pelas limitações dos algoritmos tradicionais de particionamento, como o *K-Means*[14].

O *K-Means* assume que todos os agrupamentos (clusters) possuem formato esférico e densidades similares, além de forçar a inclusão de todos os pontos em algum grupo . No entanto, dados reais de publicações científicas raramente seguem esse padrão: tópicos podem ter formatos irregulares e muitos documentos podem não pertencer a nenhum tema

[14] O *K-Means* é um dos algoritmos de *clustering* mais populares. Ele particiona n observações em k agrupamentos, onde cada observação pertence ao *cluster* cujo centro (média) é o mais próximo. Sua simplicidade é uma vantagem, mas ele assume *clusters* de forma esférica e sensibilidade à inicialização dos centroides (MACQUEEN, 1967).

específico (ruído).

A Figura 7 apresenta um cenário sintético que ilustra exatamente esse desafio de "densidade variável e ruído", típico de dados não estruturados. Ao analisar a Figura 7, observa-se a coexistência de três situações distintas no mesmo conjunto de dados:

1. **Clusters de Alta Densidade:** Agrupamentos compactos (topo e esquerda), representando temas muito específicos e coesos.
2. **Clusters de Baixa Densidade:** Agrupamentos mais dispersos (direita), representando temas mais amplos ou menos consolidados.
3. **Ruído (*Noise*)** Pontos isolados espalhados pelo fundo, que não se conectam claramente a nenhum grupo.

O HDBSCAN supera esse desafio por ser um algoritmo baseado em densidade. Diferentemente de métodos que buscam apenas a distância até um centro, o HDBSCAN identifica "ilhas" de alta densidade em um "mar" de pontos dispersos. Essa característica permite que o algoritmo:

- Identifique clusters de formatos arbitrários e densidades variadas simultaneamente;
- Classifique pontos isolados como outliers (ruído), atribuindo-lhes o rótulo -1, em vez de força-los a integrar um tópico incoerente.

A Figura 7 ilustra a capacidadade de identificar agrupamentos de densidades e formas variadas, demonstrando 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. 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.

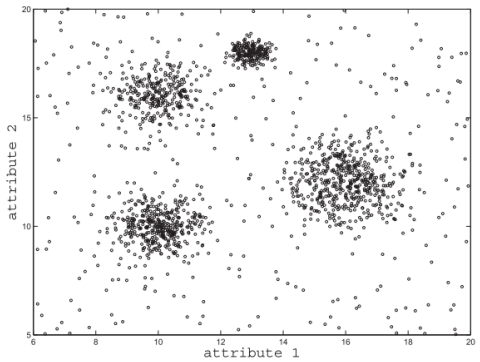

Figura 7 – Figura ilustrativa de um *dataset* sintético com quatro *clusters* e ruído de fundo.

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

As duas etapas finais do *pipeline*, é onde ocorre a geração da representação dos tópicos. Abordagens anteriores, como o *Top2Vec* (ANGELOV, 2020), baseiam-se em encontrar o centroide (o ponto médio) do *cluster* e identificar as palavras mais próximas a ele. Grootendorst (2022) argumenta que essa abordagem é falha, pois "um *cluster* nem sempre se situa dentro de uma esfera ao redor de um centroide" (GROOTENDORST, 2022, p. 1, Traduzido).

Para resolver o cénario dos clusters ao redor do centroide, o BERTopic introduz o c-TF-IDF (*Class-based Term Frequency-Inverse Document Frequency*). A abordagem primeiro trata todos os documentos dentro de um *cluster* (tópico) como um único documento concatenado. Em seguida, modifica a fórmula padrão do TF-IDF para operar a nível de classe, e não de documento.

O TF-IDF clássico é definido por Joachims (1997) como:

$$W_{t,d} = tf_{t,d} \cdot \log\left(\frac{N}{df_t}\right) \tag{2.1}$$

onde $W_{t,d}$ é a pontuação da palavra t no documento d, $tf_{t,d}$ é a frequência da palavra t no documento d, N é o número total de documentos e df_t é o número de documentos que contêm a palavra t.

O c-TF-IDF adapta essa lógica, onde a frequência do termo (tf) é calculada para a palavra t dentro da classe c inteira (o *cluster* de documentos concatenados). A frequência inversa do documento (idf) é substituída pela "frequência inversa da classe", que mede a importância da palavra t em relação a todas as outras classes. A fórmula é então ajustada para:

$$W_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{tf_t}\right) \tag{2.2}$$

onde $tf_{t,c}$ é a frequência da palavra t na classe c, A é o número médio de palavras por classe (total de palavras dividido pelo número de classes), e tf_t é a frequência total da palavra t em todas as classes (GROOTENDORST, 2022, p. 3). O resultado é uma lista de palavras que destaca os termos que são mais distintivos e representativos de um tópico específico.

Embora eficaz, o c-TF-IDF pode gerar palavras-chave redundantes (ex: "modelo", "modelagem"). O Grootendorst (2022, p. 8) sugere que isso pode ser resolvido "aplicando *Maximal Marginal Relevance* (MMR) às n palavras principais de um tópico". O MMR, introduzido por Carbonell e Goldstein (1998), é uma técnica projetada especificamente para otimizar o equilíbrio entre relevância e diversidade na recuperação de informações. O algoritmo funciona de forma iterativa: ele primeiro seleciona o termo de maior relevância (maior pontuação c-TF-IDF); em seguida, para cada termo candidato subsequente, ele aplica uma penalidade com base na similaridade desse candidato com os termos já selecionados. O resultado é um conjunto de palavras-chave que não apenas representa o tema central, mas

também cobre diferentes facetas semânticas desse tema, aumentando significativamente a interpretabilidade humana.

Contudo, mesmo com representações de tópicos robustas e rótulos semanticamente diversos, analisar a estrutura latente e as inter-relações de centenas de tópicos em um *corpus* massivo permanece um desafio. A geração de um modelo de tópicos é apenas a primeira etapa; a descoberta de conhecimento emerge da capacidade de explorar esses resultados de forma intuitiva. Isso destaca a necessidade de técnicas que superem listas estáticas e permitam uma análise exploratória, um desafio que é central no campo da Visualização Científica e de Dados.

2.6 Visualização de Dados para Análise Científica

A geração de modelos de tópicos e *embeddings*, conforme discutido nas seções anteriores, produz representações vetoriais de alta dimensionalidade que capturam a semântica do domínio. No entanto, a interpretação e o uso prático desses *embeddings* representam um desafio significativo, dada a sua "opacidade, alta dimensionalidade e o grande tamanho dos conjuntos de dados modernos" (WANG; HOHMAN; CHAU, 2023, p. 1, Traduzido).

Para tornar esses vetores complexos inteligíveis, pesquisadores frequentemente aplicam técnicas de redução de dimensionalidade, como o UMAP (MCINNES; HEALY; MELVILLE, 2018) ou o *t-distributed Stochastic Neighbor Embedding* (t-SNE) (MAATEN; HINTON, 2008), para projetar os *embeddings* em um espaço bidimensional (**2d**) ou tridimensional (**3d**). Embora essa projeção permita a visualização dos dados em um gráfico de dispersão (*scatter plot*), a análise em larga escala permanece um desafio: em conjuntos de dados com milhões de pontos, "é exaustivo ou mesmo implausível inspecionar os dados ponto a ponto para entender a estrutura global" (WANG; HOHMAN; CHAU, 2023, p. 2, Traduzido).

Abordagens alternativas, como gráficos de contorno (*contour plots*), podem resumir a distribuição global, mas "restringem a exploração das estruturas locais de um *embedding*" (WANG; HOHMAN; CHAU, 2023, p. 2, Traduzido). Para preencher a lacuna entre a visão global (contornos) e a exploração local (pontos), ferramentas de visualização interativa tornam-se essenciais.

Neste contexto, surge o WizMap[15], "uma ferramenta de visualização interativa escalável que capacita pesquisadores e especialistas de domínio a explorar e interpretar *embeddings* com milhões de pontos" (WANG; HOHMAN; CHAU, 2023, p. 2, Traduzido). A ferramenta emprega um "design de interação familiar semelhante a um mapa" (*map-like*

[15] O repositório de código aberto do WizMap está disponível em: https://github.com/poloclub/wizmap.

interaction design), permitindo que o usuário navegue pelo espaço semântico com ações de *pan* e *zoom*.

A interface do WizMap, ilustrada na Figura 8, é dividida em três componentes principais: (A) A Visão de Mapa (*Map View*), que integra as camadas de visualização; (B) O Painel de Busca (*Search Panel*), que permite a filtragem por texto; e (C) O Painel de Controle (*Control Panel*), para customização da visualização (WANG; HOHMAN; CHAU, 2023, p. 1).

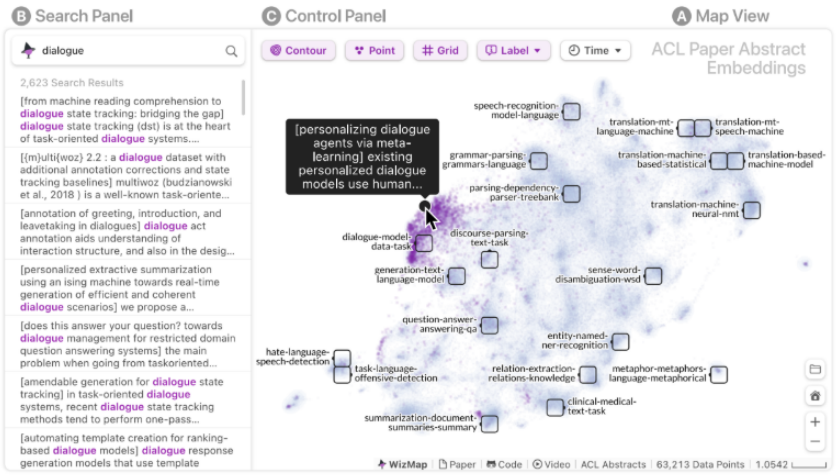

Figura 8 – Interface da ferramenta WizMap e seus componentes principais.

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

A principal inovação do WizMap é a sua capacidade de escalar para milhões de pontos diretamente no navegador do usuário, sem a necessidade de servidores dedicados. Isso é alcançado através do uso de tecnologias web modernas, como *Web Graphics Library* (WebGL)[16] para renderização gráfica, *Web Workers*[17] para paralelização, e a *Streams API*[18] para o carregamento de dados (WANG; HOHMAN; CHAU, 2023, p. 2, 4).

A "Visão de Mapa" (*Map View*), sua interface primária, integra três camadas de visualização (WANG; HOHMAN; CHAU, 2023, p. 4):

1. **Contorno de Distribuição:** Utiliza *Kernel Density Estimation* (KDE) para fornecer uma visão geral da estrutura global e das áreas de alta densidade.

2. **Gráfico de Dispersão (*Scatter Plot*):** Permite a investigação de *embeddings* individuais em nível local.

3. **Rótulos Multi-Resolução:** Permite uma interpretação contextual em diferentes níveis de granularidade.

[16] Disponível em: https://developer.mozilla.org/pt-BR/docs/Web/API/WebGL_API
[17] Disponível em: https://developer.mozilla.org/pt-BR/docs/Web/API/Web_Workers_API.
[18] Disponível em: https://developer.mozilla.org/pt-BR/docs/Web/API/Streams_API.

Para implementar os Rótulos Multi-Resolução, o WizMap utiliza uma estrutura de dados *quadtree*, conforme detalhado na Figura 9. O *quadtree* particiona recursivamente o espaço **2d** (A) em quadrantes, que são representados como nós em uma árvore (B). A ferramenta então agrega as informações de baixo para cima, permitindo que os rótulos se "ajustem em resolução à medida que os usuários aumentam o *zoom*" (WANG; HOHMAN; CHAU, 2023, p. 2-3).

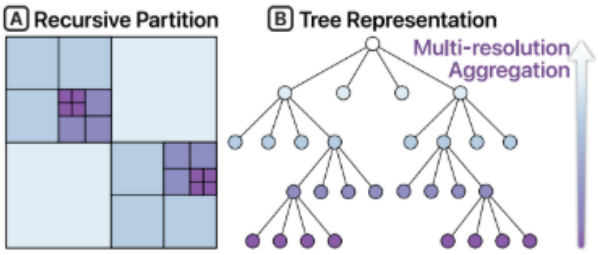

Figura 9 – Estrutura de dados *Quadtree* usada pelo WizMap para agregação multi-resolução. (A) Particionamento recursivo do espaço **2d**. (B) Representação em árvore.

Fonte: Wang, Hohman e Chau (2023, p. 3)

Ferramentas como o WizMap, que combinam redução de dimensionalidade com interfaces interativas de multi-resolução e busca semântica, são fundamentais para traduzir a saída matemática de modelos como o BERTopic em mapas de conhecimento navegáveis, facilitando a descoberta de padrões latentes em grandes base de dados textuais.

3 TRABALHOS CORRELATOS

Este capítulo apresenta uma revisão de literatura conduzida para fundamentar o estudo de metodologias de modelagem de tópicos baseadas em *embeddings*, focando em sua aplicação e avaliação no domínio da análise de tendências científicas.

Estudos comparativos recentes têm se dedicado a avaliar a eficácia de modelos de tópicos modernos frente a abordagens tradicionais. A pesquisa de Jung *et al.* (2024), por exemplo, apresenta uma análise comparativa entre métodos como LDA, *Non-negative 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 predominância em diversidade e coesão de tópicos" (JUNG *et al.*, 2024, p. 27). Essa capacidade de capturar contextos semânticos complexos, superando a abordagem de *bag-of-words* do LDA, é uma capacidade relevante para a análise de produção científica interdisciplinar.

De forma similar, Kim, Kogler e Maliphol (2024) propõem uma combinação de análise de redes e BERTopic para identificar a emergência de campos científicos interdisciplinares. O estudo valida o BERTopic como uma técnica de *embedded topic modeling* (modelagem de tópicos embarcada) que, ao contrário das abordagens baseadas em frequência, "permite considerar o conhecimento contextual de grandes conjuntos de dados de texto" (KIM; KOGLER; MALIPHOL, 2024, p. 3, Traduzido). A arquitetura empregada utiliza os componentes de *embeddings* BERT, UMAP, HDBSCAN e c-TF-IDF.

A aplicação do BERTopic para a análise de publicações científicas, especificamente para a triagem de revisões sistemáticas, também foi explorada por Galli *et al.* (2024). O estudo aplicou o *pipeline* (SBERT/UMAP/HDBSCAN/c-TF-IDF) em *datasets* da literatura médica e concluiu que a ferramenta foi eficaz na segmentação e filtragem de artigos irrelevantes, reduzindo a carga de trabalho manual (GALLI *et al.*, 2024, p. 1, 18). Notavelmente, o trabalho também identificou a representação de tópicos padrão do c-TF-IDF como "muitas vezes obscura" (GALLI *et al.*, 2024, p. 6, Traduzido).

Enquanto os estudos anteriores comparam o BERTopic com modelos clássicos, Gerasimenko *et al.* (2023) o utilizam como *baseline* para uma nova técnica de detecção de tendências científicas em tempo real. A pesquisa oferece uma análise detalhada do desempenho do BERTopic, concluindo que o modelo apresenta alta performance na distinção de documentos por tópicos, extraindo 90 dos 91 tópicos de tendência. No entanto, o estudo também identifica que o BERTopic "tem muita dificuldade na extração de palavras-chave" (GERASIMENKO *et al.*, 2023, p. 10, Traduzido), indicando um desempenho diferenciado entre as tarefas de *clustering* de documentos e de representação

de tópicos.

A arquitetura do BERTopic é também modular, permitindo a exploração de diferentes configurações para otimizar os resultados, um ponto investigado por Wijanto, Widiastuti 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 SBERT), técnicas de redução de dimensionalidade (UMAP e *Principal Component Analysis* (PCA)) e algoritmos de clusterização (*K-Means* e HDBSCAN). O estudo reforça a importância da seleção criteriosa de cada componente do *pipeline* para garantir a geração de tópicos coerentes e interpretáveis, sendo uma configuração validada o uso de SBERT, UMAP e HDBSCAN (GROOTENDORST, 2022) para documentos heterogêneos.

A literatura também aponta para a validação de *pipelines* coesos de modelagem de tópicos para análise bibliométrica e visualização. Meng *et al.* (2024), por exemplo, propõem uma metodologia que utiliza *BERTopic* para mapear a evolução da pesquisa científica em um grande volume de publicações. O trabalho de Meng *et al.* (2024) culmina no desenvolvimento de uma plataforma *web* de análise bibliométrica para visualização de redes e tópicos, validando a aplicação de *pipelines* de modelagem semântica como base para ferramentas de exploração interativa.

Em suma, a análise dos trabalhos correlatos indica que o BERTopic é uma ferramenta validada pela literatura recente para a análise de publicações científicas. A literatura confirma sua predominância sobre métodos tradicionais em métricas de coerência (JUNG *et al.*, 2024) e sua capacidade de usar contexto semântico como apontam Kim, Kogler e Maliphol (2024), Galli *et al.* (2024). Também aponta para a importância de sua modularidade (WIJANTO; WIDIASTUTI; YONG, 2024) e para um desempenho diferenciado entre a clusterização de documentos (onde é forte) e a extração de palavras-chave (onde é mais fraco) (GERASIMENKO *et al.*, 2023; GALLI *et al.*, 2024). Por fim, a literatura valida o uso de *pipelines* de modelagem como base para o desenvolvimento de plataformas de visualização interativa (MENG *et al.*, 2024).

3.1 Síntese Comparativa dos Trabalhos Correlatos

A fim de consolidar a análise da literatura e posicionar de forma clara o estado da arte, 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 e as tecnologias de *embedding*. Essa estrutura permite visualizar as sinergias e as particularidades de cada abordagem.

Tabela 1 – Quadro Resumo: Comparativo de Trabalhos Correlatos

Referência	**Objetivo Principal**	**Pipeline/Método Utilizado**	**Modelo de Embedding**	**Relação com o Estado da Arte**
Jung *et al.* (2024)	Comparar o desempenho de modelos de tópicos (LDA, NMF, BERTopic) em textos acadêmicos e de notícias sobre LLMs (JUNG *et al.*, 2024).	Análise comparativa de métricas de coerência e diversidade dos tópicos gerados (JUNG *et al.*, 2024).	SBERT (implícito no BERTopic) (JUNG *et al.*, 2024).	Estabelece o BERTopic como uma ferramenta superior aos métodos tradicionais (LDA, NMF) para a análise de textos acadêmicos (JUNG *et al.*, 2024).
Kim, Kogler e Maliphol (2024)	Identificar a emergência de ciência interdisciplinar em metadados de publicações científicas (KIM; KOGLER; MALIPHOL, 2024).	Combinação de análise de redes de coocorrência (Etapa 1) e modelagem de tópicos com BERTopic (Etapa 2) (KIM; KOGLER; MALIPHOL, 2024).	BERT (usado para *embedding vectorization*) (KIM; KOGLER; MALIPHOL, 2024).	Valida o BERTopic como ferramenta superior às abordagens baseadas em frequência, por usar "conhecimento contextual", para analisar publicações científicas (KIM; KOGLER; MALIPHOL, 2024).
Galli *et al.* (2024)	Explorar como o BERTopic pode ser aplicado para acelerar a triagem de literatura em revisões sistemáticas de publicações científicas (GALLI *et al.*, 2024).	Pipeline BERTopic padrão (SBERT → UMAP → HDBSCAN → c-TF-IDF) para identificar e filtrar *clusters* irrelevantes (GALLI *et al.*, 2024, p. 4).	'all-mpnet-base-v2' (SBERT) (GALLI *et al.*, 2024, p. 4).	Valida o *pipeline* SBERT/UMAP/HDBSCAN para analisar publicações científicas e corrobora que os rótulos de c-TF-IDF são "muitas vezes obscuros" (GALLI *et al.*, 2024, p. 6).
Gerasimenko *et al.* (2023)	Extrair tópicos de tendências científicas ("trend topics") em tempo real a partir de publicações (GERASIMENKO *et al.*, 2023).	Propõe um modelo ARTM incremental e o compara com *baselines*, incluindo PLSA, LDA e BERTopic (GERASIMENKO *et al.*, 2023).	Sentence-Transformers (para o *baseline* BERTopic) (GERASIMENKO *et al.*, 2023).	Fornece uma análise comparativa do BERTopic, destacando sua alta performance em clusterização de documentos e sua fraqueza na extração de palavras-chave (GERASIMENKO *et al.*, 2023).
Meng *et al.* (2024)	Mapear a evolução de um campo de pesquisa científica utilizando uma abordagem integrada de modelagem de tópicos e uma plataforma web (MENG *et al.*, 2024).	Pipeline integrado: 1. Geração de embeddings (via API de LLM); 2. Clusterização e modelagem com BERTopic; 3. Plataforma de visualização (MENG *et al.*, 2024, p. 3-4).	GPT-3.5 (text-embedding-ada-002) (MENG *et al.*, 2024, p. 4).	Valida a aplicação de um *pipeline* de modelagem de tópicos como base para uma plataforma *web* de visualização, um objetivo relevante para a análise de grandes *corpora* (MENG *et al.*, 2024, p. 18).

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 utilizada 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; ANTUNES, 2015).

O objetivo é construir um *pipeline* computacional para o mapeamento interativo de publicações científicas do Observatório de dados públicos de ciência e tecnologia da Bahia, combinando a modelagem de tópicos do BERTopic com a visualização interativa do WizMap.

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 10. Este capítulo está estruturado para detalhar cada etapa desse processo: inicia-se pela **Identificação do Problema e Definição de Objetivos** (Seção 4.1), descreve o **Desenvolvimento do Artefato** (Seção 4.2), detalha os procedimentos de **Avaliação e Validação** (Seção 4.3) e conclui com a **Apresentação dos Resultados e Comunicação** (Seção 4.4). A seguir, cada etapa do DSR é detalhada e contextualizada no escopo deste trabalho.

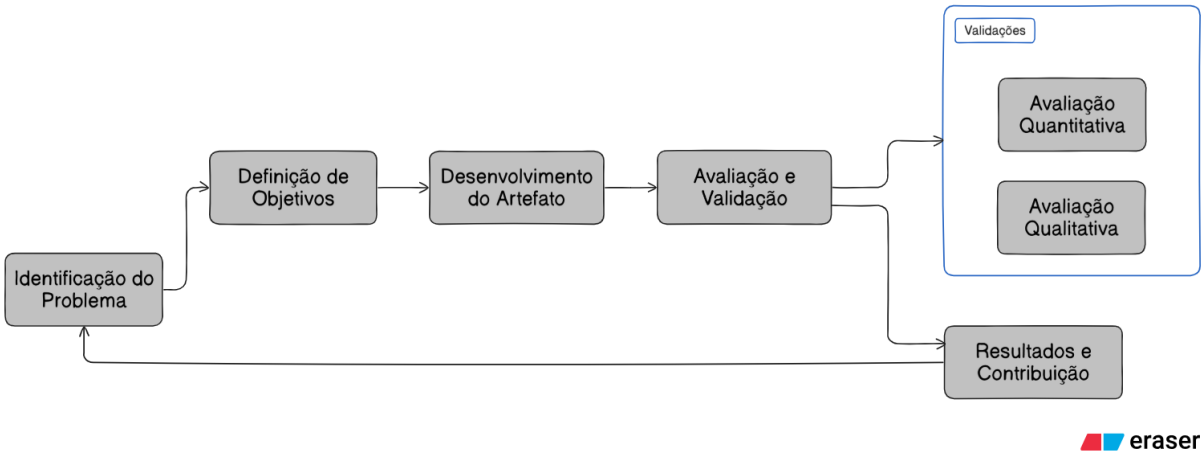

Figura 10 – Adaptação da Design Science Research para este projeto.

Fonte: O Autor

4.1 Identificação do Problema e Definição de Objetivos

A primeira fase da DSR (Figura 10) consiste na identificação de uma lacuna relevante. Atualmente, a exploração do acervo no Observatório de dados públicos de

ciência e tecnologia da Bahia é limitada a abordagens de recuperação de informação baseadas em palavras-chave. Embora funcional, essa abordagem dificulta a identificação de conexões semânticas, a descoberta de temas interdisciplinares e a compreensão da estrutura global do conhecimento científico.

Essa limitação evidencia a necessidade de uma solução que vá além da busca lexical e ofereça uma **exploração semântica e visual** da distribuição do arcervo de pesquisa na reunidas na plataforma. O problema de pesquisa, portanto, é: como um *pipeline* de modelagem de tópicos (BERTopic) pode ser integrado a uma ferramenta de visualização interativa (WizMap) para mapear o acervo de publicações do Observatório, melhorando a análise e a descoberta de temas emergentes?

A partir disso, formulam-se as "Conjecturas Teóricas": a integração do BERTopic com uma interface de mapa interativo (como o WizMap) tem o potencial de melhorar significativamente a identificação, classificação e exploração de temas, oferecendo uma compreensão mais profunda dos dados por meio de visualizações interativas.

4.2 Desenvolvimento do Artefato

Com base nos objetivos definidos na seção anterior, a etapa seguinte da DSR é o desenvolvimento do "Artefato". Neste trabalho, o artefato é um *pipeline* computacional para o mapeamento interativo do conhecimento científico contido no acervo do Observatório.

Este *pipeline* é projetado para executar as seguintes funções-chave:

1. Realizar a modelagem de tópicos com o BERTopic, utilizando *embeddings* contextuais (Seção 2.3.3) para identificar padrões temáticos de forma semântica e com alta coerência, capazes de agrupar documentos mesmo com variações lexicais.

2. Aplicar o MMR (*Maximal Marginal Relevance*), detalhado na Seção 2.5, como uma etapa de pós-processamento para refinar os rótulos de tópicos derivados do c-TF-IDF.

3. Apresentar os tópicos e documentos em uma interface gráfica interativa (WizMap) para facilitar a exploração visual das conexões temáticas (Seção 2.6).

A construção deste artefato se apoia no "Estado da Técnica", detalhado no Capítulo 2. O *pipeline* integra os conceitos de Modelagem de Tópicos (BERTopic), Embeddings Contextuais (SBERT), Redução de Dimensionalidade (UMAP), Algoritmos de Clusterização (HDBSCAN) e Visualização de Dados Interativa (WizMap).

4.3 Avaliação e Validação

O ciclo de DSR exige uma avaliação com o caráter mais rigoroso, representada no diagrama pelas "Avaliações" 1, 2 e 3. Para garantir o rigor científico e a validade do artefato desenvolvido, será empregada uma abordagem de validação mista, combinando métricas quantitativas e análises qualitativas. Esta etapa é fundamental para aferir se os resultados gerados são consistentes, coerentes e se resolvem o problema de exploração de conhecimento identificado.

4.3.1 Validação Quantitativa

A validação quantitativa foca em avaliar objetivamente a qualidade dos tópicos gerados pelo *pipeline* do BERTopic. Serão utilizadas métricas consolidadas na literatura de PLN:

- **Coerência de Tópicos (*Topic Coherence*):** A coerência avalia a consistência semântica e a interpretabilidade de um tópico. Um tópico é considerado coerente se as palavras que o representam aparecem frequentemente juntas no *corpus*, indicando que formam um conceito semântico lógico (JUNG *et al.*, 2024). Para esta avaliação, será calculada a pontuação ***Normalized Pointwise Mutual Information* (NPMI)** (*Normalized Pointwise Mutual Information*). O cálculo da NPMI consiste em extrair os N termos mais representativos de cada tópico (ex: $N = 10$), formar pares desses termos, e calcular a probabilidade de sua coocorrência (aparição conjunta) nos documentos do *corpus* original, normalizada pela probabilidade de suas ocorrências individuais. A pontuação NPMI varia de +1 (coocorrência perfeita) a -1 (nunca aparecem juntas). Um escore positivo indica que os termos coexistem mais do que o esperado pelo acaso, e escores mais altos indicarão um modelo mais coerente.

- **Diversidade de Tópicos (*Topic Diversity*):** A diversidade é utilizada para garantir que o modelo não está gerando tópicos redundantes, ou seja, *clusters* diferentes descritos pelas mesmas palavras-chave. Um modelo ideal deve apresentar tópicos que sejam, ao mesmo tempo, coerentes (alta NPMI) e distintos entre si (alta diversidade). A métrica é calculada como a "porcentagem de palavras únicas em todos os tópicos" (GROOTENDORST, 2022, p. 5, Traduzido) e é definida pela fórmula $M/(N \times k)$, onde M é o número de palavras únicas extraídas nos k termos principais de todos os N tópicos (JUNG *et al.*, 2024, p. 5). O resultado varia de 0 (todos os tópicos são idênticos) a 1 (todos os termos em todos os tópicos são únicos).

4.3.2 Validação Qualitativa

A validação qualitativa é essencial para confirmar se os resultados quantitativos se traduzem em valor prático e interpretabilidade. Esta etapa responde diretamente à "Avaliação 2" (O problema foi resolvido?), verificando se o artefato de fato soluciona o problema da exploração de conhecimento identificado. Esta validação será focada em dois componentes e conduzida por um especialista da área:

- **Análise de Rótulos (Interpretabilidade):** Para validar a eficácia da etapa de refinamento (Seção 4.2, item 2), será conduzida uma comparação direta. Para uma amostra de tópicos, será apresentado ao especialista o rótulo de linha de base (ex: as 3 palavras-chave do c-TF-IDF puro) e o rótulo refinado (as 3 palavras-chave selecionadas pelo MMR). O especialista avaliará qual dos dois rótulos é mais informativo, menos redundante e mais preciso, permitindo aferir a melhoria na interpretabilidade.
- **Inspeção Visual dos *Clusters* (Validade Semântica):** Para validar se os *clusters* possuem coerência semântica interna (i.e., se estão "bem formados") e se o mapa reflete uma estrutura de conhecimento válida, o especialista inspecionará o WizMap. A validação consistirá em selecionar alguns *clusters* e analisar os documentos que os compõem (através dos *tooltips* de cada *datapoint*). Se os documentos dentro de um mesmo *cluster* forem coesos e pertencerem ao tema descrito pelo rótulo, o artefato será considerado válido para resolver o problema da exploração semântica.

4.4 Apresentação dos Resultados e Comunicação

A etapa final do ciclo DSR, conforme o *framework* de Dresch, Lacerda e Antunes (2015), é a "Comunicação". Esta fase é dedicada à documentação e disseminação dos achados obtidos durante o desenvolvimento e a avaliação do artefato.

Os resultados da aplicação do *pipeline* e de sua validação (conforme definido na Seção 4.3) serão detalhados nos capítulos subsequentes. O **Capítulo 5** apresentará a implementação técnica do artefato e a arquitetura da solução. O **Capítulo 6** analisará os dados gerados, a qualidade dos tópicos (via NPMI e Diversidade) e a validade da interface de visualização.

O objetivo desta comunicação é apresentar e validar uma solução tecnológica para o problema de exploração de conhecimento no Observatório, documentando o processo de desenvolvimento e os resultados obtidos. Este trabalho visa demonstrar a aplicação prática de técnicas de modelagem de tópicos e visualização interativa, consolidando o aprendizado e a experiência adquiridos no desenvolvimento de um sistema funcional.

5 PROJETO DE DESENVOLVIMENTO

O desenvolvimento do artefato proposto neste trabalho segue uma abordagem estruturada e incremental, alinhada à metodologia DSR detalhada no Capítulo 4. O objetivo é construir e validar um *pipeline* computacional que venha a aprimorar e complementar a análise de publicações científicas na plataforma do **Observatório de dados públicos de ciência e tecnologia da Bahia**[1].

Conforme descrito na literatura Santos *et al.* (2024), Jorge *et al.* (2025), o Observatório já possui uma arquitetura robusta para o mapeamento de competências, integrando fontes de dados heterogêneas como a Plataforma Lattes[2], a Plataforma Sucupira[3], o OpenAlex[4] e o *Journal Citation Reports* (JCR)[5]. Seu sistema atual utiliza um processo de Extract, Transform and Load (ETL) com Apache Hop[6] para consolidar as informações em um banco de dados PostgreSQL[7] e já emprega técnicas de 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. No entanto, como discutido na metodologia, a grande variação terminológica em domínios científicos 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 capítulo detalha, portanto, a construção de um *pipeline* que representa uma evolução para a arquitetura do Observatório. A solução proposta não visa substituir, mas sim enriquecer a funcionalidade de busca atual, introduzindo a modelagem de tópicos moderna com o BERTopic, o refinamento de rótulos com MMR e a visualização interativa com o WizMap. 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.

[1] Disponível em: https://simcc.uesc.br/observatorio
[2] Disponível em: http://lattes.cnpq.br/
[3] Disponível em: https://sucupira.capes.gov.br/
[4] Disponível em: https://openalex.org/
[5] Disponível em: https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports/
[6] Disponível em: https://hop.apache.org/
[7] Disponível em: https://www.postgresql.org/

5.1 Tecnologias Utilizadas

O desenvolvimento do artefato proposto neste projeto assenta-se sobre a combinação de duas arquiteturas tecnológicas distintas: (1) a infraestrutura consolidada do **Observatório de dados públicos de ciência e tecnologia da Bahia**, que serve como fonte de dados e contexto de aplicação; e (2) o *pipeline* de modelagem e visualização desenvolvido, que constitui o artefato central deste estudo.

5.1.1 Base Tecnológica do Observatório

A arquitetura do Observatório, que serve como ponto de partida para este trabalho, foi projetada para ser robusta e escalável, utilizando um conjunto de tecnologias consolidadas para a gestão de dados acadêmicos, conforme detalhado por Santos *et al.* (2024) e Jorge *et al.* (2025).

Suas principais tecnologias incluem:

- **Banco de Dados (PostgreSQL):** O Observatório utiliza o Database Management System (DBMS) PostgreSQL para armazenar e consolidar as informações. O sistema aproveita os recursos nativos de busca textual (*Full-Text Search*) do PostgreSQL para a recuperação de informações baseada em termos.
- **Orquestração de Dados (Apache Hop):** Para o ETL dos dados de fontes diversas (Lattes, Sucupira, JCR, etc.), o Observatório utiliza o Apache Hop. Essa ferramenta é responsável por automatizar e coordenar o fluxo de ingestão de dados, garantindo a consistência das informações.
- **Infraestrutura de Aplicação (Python e React):** O *back-end* da plataforma é desenvolvido em Python, utilizando o *framework* Flask para a *Application Programming Interface* (API). A interface de usuário (*front-end*) é construída com a biblioteca React JS.

A Figura 11 ilustra essa arquitetura existente, mostrando o fluxo de dados desde as fontes externas (como Lattes e Sucupira) até o banco de dados, que será utilizado pelo framework para exibição das informações no front-end.

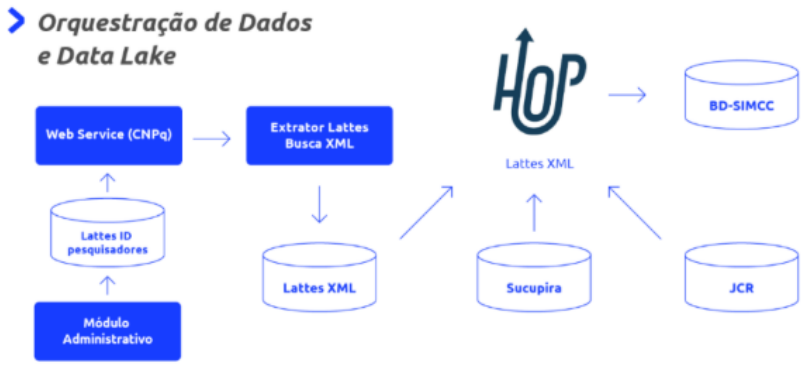

Figura 11 – Arquitetura Geral do Observatório (SIMCC).

Fonte: Jorge *et al.* (2025)

5.1.2 Pipeline de Modelagem e Visualização

Sobre a base conceitual do Observatório, este projeto implementa um novo *pipeline* focado na descoberta e análise semântica do conhecimento. Dada a natureza deste trabalho como um estudo de caso (conforme metodologia DSR), o *pipeline* foi desenvolvido e validado em um ambiente de prototipagem computacional.

As tecnologias e bibliotecas utilizadas para a construção do artefato foram:

- **Ambiente de Desenvolvimento (Google Colaboratory):** Todo o processamento, desde a leitura dos dados até a geração dos resultados, foi executado na plataforma Google Colab[8], um ambiente interativo baseado em notebooks Jupyter que fornece acesso a recursos computacionais (CPUs e GPUs).
- **Manipulação de Dados e Pré-processamento:** Para a ingestão do *dump* do banco de dados e para as etapas de limpeza e pré-processamento textual (normalização, remoção de *stopwords*, etc.), foram utilizadas bibliotecas centrais do ecossistema Python para ciência de dados, como `Pandas` para a manipulação dos dados tabulares e `NLTK` e `spaCy` para as tarefas de PLN.
- **Modelagem de Tópicos (BERTopic):** Para a identificação dos temas latentes, a biblioteca `bertopic` foi a tecnologia central. Ela orquestra todo o fluxo de modelagem descrito no Capítulo 2, integrando os seguintes componentes-chave:
  - **Embeddings (Sentence-BERT):** A geração dos vetores semânticos dos documentos foi realizada pela biblioteca `sentence-transformers`.
  - **Redução de Dimensionalidade (UMAP):** A projeção dos *embeddings* para um espaço de baixa dimensão foi executada com a biblioteca `umap-learn`.

[8] Disponível em: https://colab.research.google.com/drive/1IdeIJ14TeLEZuJAB176UqALWFRwFXA2p

  - **Clusterização (HDBSCAN):** O agrupamento dos vetores para a formação dos tópicos foi feito com a biblioteca `hdbscan`.

- **Refinamento de Rótulos (MMR):** Para o refinamento e a diversificação dos rótulos dos tópicos, foi utilizada a classe `MaximalMarginalRelevance` do próprio BERTopic, que implementa o algoritmo MMR para selecionar palavras-chave mais informativas e menos redundantes.

- **Visualização Interativa (WizMap):** Para a apresentação final dos resultados, foi utilizada a biblioteca `wizmap`. A função desta biblioteca foi empregada para gerar o arquivo HTML final, que renderiza o mapa interativo a partir dos dados processados (coordenadas 2D, IDs de tópicos e metadados) exportados pelo *pipeline*.

- **Hospedagem de Dados (Gist):** Para viabilizar a renderização do WizMap, que opera no lado do cliente (navegador), os arquivos de dados em formato JSON gerados pelo *pipeline* foram hospedados em um Gist (GitHub Gist)[9], permitindo que a ferramenta de visualização os consumisse de forma pública e estática.

5.2 Arquitetura da Solução

Esta seção detalha a implementação prática do artefato computacional, descrevendo o fluxo de dados e as etapas de processamento que compõem o *pipeline* de análise. Conforme a metodologia DSR (Capítulo 4), esta é a fase de "Desenvolvimento do Artefato", que aplica as tecnologias descritas na Seção 5.1.

O objetivo central deste *pipeline* é transformar o acervo textual bruto das publicações científicas do Observatório, em um mapa de conhecimento interativo e semanticamente navegável.

Para alcançar esse objetivo, o *pipeline* foi estruturado em quatro etapas macro, que serão detalhadas nas subseções seguintes:

1. **Coleta e Pré-processamento dos Dados (Seção 5.2.1):** Extração dos dados textuais (títulos e resumos) da base de dados e sua subsequente limpeza e normalização.

2. **Modelagem de Tópicos (Seção 5.2.2):** Geração dos *embeddings* de sentenças (SBERT), redução de dimensionalidade (UMAP) e agrupamento (HDBSCAN) para a identificação dos *clusters* temáticos.

3. **Refinamento e Representação dos Tópicos (Seção 5.2.3):** Extração das palavras-chave via c-TF-IDF e aplicação do MMR para gerar rótulos semanticamente diversos e interpretáveis.

[9] Disponível em: https://gist.github.com/jeoaraujx/c9a610202e70139054c7e37eab937b93

4. **Geração e Exportação para Visualização (Seção 5.2.4):** Formatação dos dados processados (coordenadas 2D, metadados e rótulos) e exportação para o formato JSON consumível pela ferramenta WizMap.

O fluxograma completo deste artefato, desde a ingestão da base de dados até a geração da visualização final no WizMap, está representado esquematicamente na Figura 12.

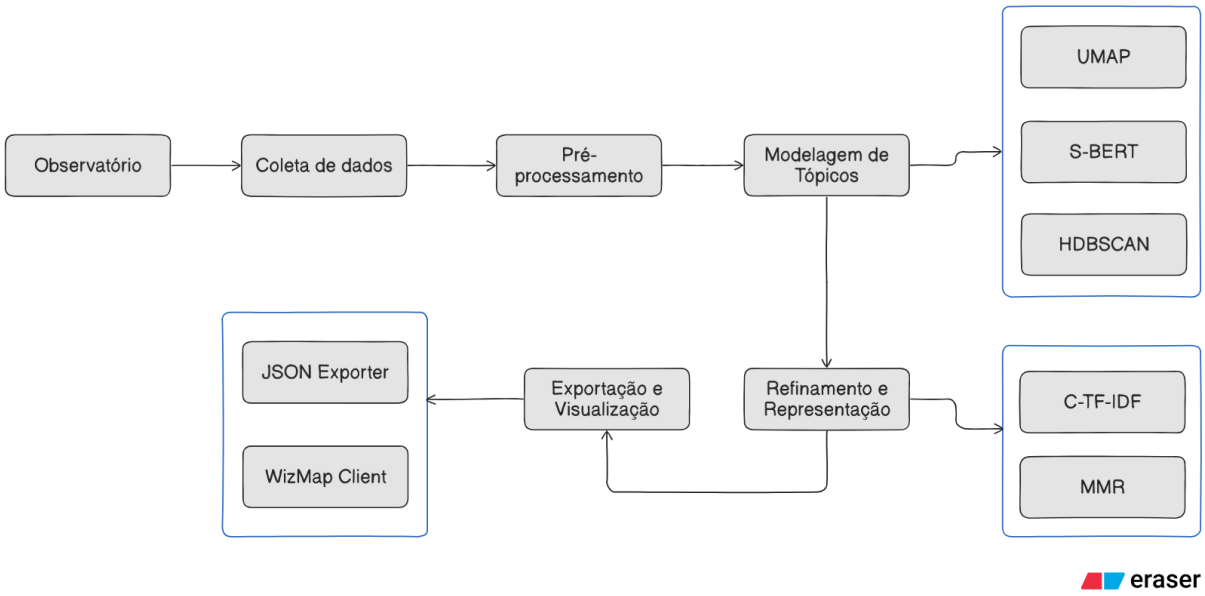

Figura 12 – Arquitetura do Pipeline de Modelagem e Visualização.

Fonte: O Autor

5.2.1 Coleta e Pré-processamento dos Dados

O ponto de partida do *pipeline* foi a coleta e a preparação do conjunto de dados textuais. Para este estudo de caso, foi utilizado um *dump* estático da base de dados (representado pelo arquivo **npai_database.csv**), contendo os metadados das publicações científicas. A coluna principal utilizada para a modelagem de tópicos foi a **title**, contendo os títulos das publicações.

O processo de preparação dos dados, executado no ambiente Google Colab, foi dividido em duas fases: a limpeza em nível de *dataset* e o pré-processamento textual em nível de documento.

Na primeira fase, utilizando a biblioteca **Pandas**, foi realizada uma filtragem inicial do *dataset*. Removeram-se todas as entradas onde o título era nulo ou consistia apenas em espaços em branco. Em seguida, aplicou-se a remoção de duplicatas exatas (via

drop_duplicates), uma etapa metodológica crucial para evitar que publicações idênticas distorcessem a frequência dos termos e biassem o modelo de tópicos.

Na segunda fase, cada título passou por um rigoroso *pipeline* de pré-processamento multilíngue, projetado para normalizar e limpar o texto, preservando apenas seu núcleo semântico. Este processo seguiu, para cada documento, a seguinte sequência de operações:

1. **Detecção de Idioma:** O texto foi analisado pela biblioteca **langdetect** para identificar o idioma (primariamente português ou inglês), determinando quais ferramentas de PLN seriam aplicadas.

2. **Normalização e Limpeza de Caracteres:** O texto foi padronizado. Primeiramente, foi convertido para minúsculas. Em seguida, aplicou-se a normalização Unicode (**NFKD**) para decompor caracteres acentuados, e a codificação para American Standard Code for Information Interchange (ASCII) (com **ignore**) foi usada para remover todas as acentuações (ex: *'educação'* → *'educacao'*). Por fim, uma expressão regular (**RegEx**) removeu todos os números, pontuações e caracteres não alfabéticos.

3. **Lematização e Tokenização (spaCy):** Com base no idioma detectado, o modelo **spaCy** correspondente (**pt_core_news_sm** para português ou **en_core_web_sm** para inglês) foi aplicado. Este modelo realizou a tokenização (divisão do texto em palavras) e a lematização, reduzindo cada palavra à sua forma canônica (ex: *'pesquisas'* ou *'pesquisando'* → *'pesquisar'*).

4. **Remoção de Stopwords (NLTK):** Utilizando as listas de *stopwords* (termos de parada) da biblioteca **NLTK** para ambos os idiomas, os termos semanticamente vazios (como *'o'*, *'para'*, *'the'*, *'is'*) foram removidos.

5. **Filtragem Final de Tokens:** Foram descartados todos os *tokens* (lemas) resultantes com menos de três caracteres, bem como duplicatas de *tokens* dentro de um mesmo documento, para criar uma representação final limpa e concisa.

Ao final desta etapa, as publicações que resultaram em textos vazios (ex: títulos que continham apenas *stopwords* ou siglas curtas) foram removidas do conjunto final. O *dataset* resultante, salvo como **npai_processed_texts.csv**, consistiu em uma lista de documentos processados (denominados **title_processed**), prontos para a etapa de geração de *embeddings*.

5.2.2 Modelagem de Tópicos

A fase de modelagem de tópicos é o cerne do *pipeline*, onde os textos pré-processados são transformados em representações numéricas e agrupados em tópicos coerentes. Esta

etapa segue a arquitetura do BERTopic, integrando modelos de *embeddings*, técnicas de redução de dimensionalidade e algoritmos de *clusterização*, conforme detalhado na Seção 2.5.

5.2.2.1 Geração de Embeddings

Para a representação semântica dos títulos das publicações, foi utilizado o modelo pré-treinado **paraphrase-multilingual-MiniLM-L12-v2** da biblioteca **sentence-transformers**. Este modelo foi escolhido por sua capacidade de gerar *embeddings* de sentenças de alta qualidade para múltiplos idiomas, incluindo português e inglês, e por sua eficiência computacional.

O modelo **SentenceTransformer** transformou cada título pré-processado em um vetor numérico de 384 dimensões, capturando o significado contextual das palavras e a similaridade semântica entre os documentos. Os *embeddings* resultantes foram armazenados em um arquivo binário (**npai_embeddings.npy**) para reuso, otimizando o tempo de processamento em execuções subsequentes.

5.2.2.2 Redução de Dimensionalidade (UMAP)

Após a geração dos *embeddings*, a alta dimensionalidade dos vetores (384 dimensões) foi reduzida para facilitar a *clusterização* e a visualização. O emprego direto de algoritmos de *clusterização* nesse espaço dimensional elevado é computacionalmente custoso e impactado pela "maldição da dimensionalidade" referida no seção 2.5. Para essa finalidade, empregou-se o algoritmo UMAP, que se destaca por sua capacidade de preservar a estrutura topológica dos dados (conforme Seção 2.5) de forma performática.

A instância do UMAP foi configurada com os seguintes parâmetros, ajustados para otimizar a descoberta de tópicos de nicho no *dataset* de teste:

- **n_neighbors=5**: Define o número de vizinhos a serem considerados. Um valor baixo como **5** foi escolhido para forçar o algoritmo a focar na **estrutura local** dos dados. Dado o volume reduzido do *dataset* de teste (NPAI), esta abordagem é ideal para identificar *clusters* de nicho, pequenos e coesos, em vez de tentar inferir uma macroestrutura global que exigiria mais dados.
- **n_components=2**: Os *embeddings* foram projetados para duas dimensões. Este valor foi um **requisito direto** da ferramenta de visualização WizMap, que necessita de coordenadas bidimensionais (x, y) para renderizar o mapa interativo.
- **min_dist=0.0**: Controla a distância mínima entre os pontos no espaço 2D. Um valor de **0.0** foi selecionado para permitir que os *clusters* se tornassem o mais compactos

possível, priorizando uma **separação visual clara** entre os tópicos no mapa, o que facilita a interpretação humana dos resultados.

- **metric="cosine"**: Define a métrica de distância para comparar os vetores no espaço original de 384 dimensões. A métrica **"cosine"** é o padrão para dados de PLN, pois ela mede o ângulo entre os vetores (similaridade semântica) e ignora suas magnitudes, que não são relevantes para *embeddings* de sentenças.
- **random_state=42**: Fixa a semente do gerador de números aleatórios. Como o UMAP é um algoritmo estocástico (seus resultados podem variar ligeiramente a cada execução), definir um estado fixo é uma prática de rigor científico **essencial para garantir a reprodutibilidade** completa do experimento.

É fundamental ressaltar que este conjunto de parâmetros (**n_neighbors=5**, **min_dist=0.0**) foi ajustado para a natureza exploratória e o volume reduzido do *dataset* de teste. Em uma futura implementação sobre a base de dados completa do Observatório, que é ordens de magnitude maior, esses parâmetros (especialmente o **n_neighbors**) deveriam ser reavaliados e aumentados, a fim de capturar uma estrutura global mais significativa do conhecimento.

5.2.2.3 Clusterização (HDBSCAN)

Com os *embeddings* já projetados em um espaço bidimensional pelo UMAP, a etapa seguinte consistiu em identificar os agrupamentos densos de documentos, que efetivamente formam os tópicos. Para esta tarefa, foi utilizado o algoritmo HDBSCAN.

A escolha pelo HDBSCAN (conforme fundamentado na Seção 2.5) é metodologicamente superior a alternativas como o K-Means, pois não exige a definição prévia do número de tópicos (K). Além disso, sua natureza baseada em densidade permite identificar *clusters* de formas e densidades variadas, e sua capacidade intrínseca de classificar pontos como ruído (*outliers*) é ideal para dados textuais, onde nem toda publicação pertence a um tópico bem definido.

Os parâmetros do modelo HDBSCAN foram ajustados para a natureza exploratória deste estudo de caso:

- **min_cluster_size=4**: Define o número mínimo de documentos que podem constituir um tópico. Dado que o *pipeline* foi executado sobre um *dataset* de teste de volume reduzido, um valor baixo como **4** foi **essencial para permitir a descoberta** de tópicos de nicho ou temas emergentes, que possuem poucas publicações. Um valor maior poderia resultar na não identificação de nenhum tópico.

- **min_samples=2**: Controla o quão "denso" um grupo de pontos precisa ser para ser considerado um *cluster* central (influenciando diretamente o que é classificado como ruído). Um valor baixo como **2** torna o algoritmo mais permissivo, **reduzindo a agressividade na classificação de ruído** e permitindo que *clusters* em áreas de menor densidade (temas menos frequentes) possam ser formados.

- **metric="euclidean"**: Define a métrica de distância para calcular a proximidade entre os pontos. Uma vez que o UMAP já havia processado a similaridade semântica (com a métrica **"cosine"**) e projetado os dados para um espaço 2D, a métrica **"euclidean"** (distância geométrica padrão) é a **mais adequada e computacionalmente eficiente** para medir a proximidade neste novo espaço bidimensional.

- **cluster_selection_method="eom"**: O HDBSCAN é um algoritmo hierárquico; este parâmetro define como "cortar" a árvore de *clusters*. O método **"eom"** (Excess of Mass) é o padrão e o mais robusto, pois seleciona os *clusters* que demonstram a **maior estabilidade e persistência** ao longo da hierarquia de densidade, permitindo que a própria estrutura dos dados defina os tópicos mais significativos.

- **prediction_data=True**: Este é um **requisito técnico** do BERTopic. Ao ser definido como **True**, o HDBSCAN armazena informações adicionais que permitem ao modelo (posteriormente) classificar novos documentos em tópicos já existentes, garantindo que o artefato seja não apenas descritivo, mas também preditivo.

A integração destes componentes na classe **BERTopic**, juntamente com o modelo de *embedding* e os parâmetros do UMAP, permitiu a execução do método **fit_transform**. Este método orquestrou o fluxo completo, culminando na atribuição de um ID de tópico (ou -1 para ruído) para cada publicação do *dataset*.

5.2.2.4 Configuração Final do BERTopic

A instância final do modelo foi configurada para utilizar o **SentenceTransformer** como **embedding_model**, as instâncias customizadas de **UMAP** e **HDBSCAN** e, de forma notável, o **MaximalMarginalRelevance** (MMR) como **representation_model** para o refinamento dos rótulos de tópicos, que será detalhado na Seção 5.2.3. O parâmetro **language="multilingual"** foi especificado, alinhando-se ao pré-processamento bilíngue.

Após a configuração, o modelo foi treinado com os textos processados e seus respectivos *embeddings* pré-gerados (**topic_model.fit_transform(processed_texts, embeddings=embeddings_array)**). Os resultados, incluindo os IDs dos tópicos para cada documento e suas probabilidades, foram então obtidos. O modelo treinado foi salvo persistentemente (**npai_bertopic_model**) para facilitar sua reutilização.

5.2.3 Refinamento e Representação dos Tópicos (MMR)

Após a etapa de *clusterização* (Seção 5.2.2.3), o *pipeline* produziu um conjunto de tópicos, cada um consistindo em um grupo de documentos. No entanto, esses tópicos eram apenas agrupamentos numéricos; para que tivessem valor analítico, precisavam de uma representação textual interpretável.

O BERTopic realiza essa tarefa através do c-TF-IDF (Class-based Term Frequency-Inverse Document Frequency), um algoritmo que, conforme descrito na Seção 2.5, trata todos os documentos de um tópico como um único documento grande e, em seguida, calcula as pontuações TF-IDF para extrair as palavras-chave mais representativas.

Embora eficaz, uma limitação conhecida da abordagem c-TF-IDF pura é a **redundância semântica**. Como o método se baseia unicamente na pontuação de relevância, ele frequentemente retorna termos que são variações uns dos outros (ex: *"pesquisa", "pesquisas", "pesquisador"*) ou sinônimos próximos (ex: *"modelo", "modelagem"*). Isso polui a representação do tópico e compromete a sua interpretabilidade imediata, forçando o analista a inferir o conceito central a partir de palavras redundantes.

Para mitigar este problema e criar rótulos de tópicos mais claros e informativos, este artefato substituiu o modelo de representação padrão pelo MMR (Maximal Marginal Relevance). O MMR é um algoritmo de diversificação (fundamentado na Seção 2.5) que refina a lista de palavras-chave gerada pelo c-TF-IDF.

Conforme definido no código de treinamento, o modelo foi instanciado com o parâmetro `MaximalMarginalRelevance(diversity=0.3)`. O MMR opera balanceando duas métricas:

1. **Relevância:** A pontuação original da palavra (seu *score* c-TF-IDF).
2. **Diversidade:** A dissimilaridade (distância de cosseno) entre uma palavra candidata e as palavras já selecionadas para o rótulo.

O hiperparâmetro `diversity=0.3` instrui o algoritmo a priorizar fortemente a relevância (70% de peso), ao mesmo tempo em que introduz um fator moderado de diversidade (30% de peso). Na prática, isso permite que o modelo selecione a palavra mais relevante (ex: *"pesquisa"*), mas penalize termos semanticamente muito próximos (como *"pesquisas"*), favorecendo a escolha de uma palavra subsequente que, embora ainda relevante, cubra uma faceta diferente do tópico.

A aplicação do MMR como `representation_model` (passado diretamente para a instância do BERTopic) busca garantir que a saída final do modelo seja um conjunto de rótulos de tópicos semanticamente diversos, menos redundantes e mais humanamente inteligíveis, resolvendo uma das fraquezas centrais do c-TF-IDF puro.

5.2.4 Geração e Exportação para Visualização (WizMap)

A etapa final do *pipeline* consistiu na geração do artefato de visualização. Esta fase foi focada em consolidar os diversos resultados da modelagem e exportá-los para o formato de dados específico exigido pela ferramenta WizMap.

Primeiramente, foi construída uma estrutura de dados unificada. As coordenadas bidimensionais (x, y) de cada publicação, que foram calculadas pelo UMAP (Seção 5.2.2.2), foram consolidadas. A elas, foram associados os metadados originais (como o título da publicação) e os resultados da modelagem (o ID do tópico atribuído e o rótulo textual refinado pelo MMR).

Para garantir a clareza na interface final, os rótulos dos tópicos passaram por uma etapa de pós-processamento para remover prefixos numéricos (ex: `1_pesquisa_dados` → `pesquisa dados`), tornando-os mais legíveis. Adicionalmente, foi formatado um texto de *tooltip* (dica de contexto) para cada publicação, permitindo ao usuário final inspecionar o título e o tópico de um ponto específico ao interagir com o mapa.

Por fim, a biblioteca `wizmap` foi utilizada para processar essa estrutura de dados consolidada. Esta ferramenta gerou os dois arquivos JavaScript Object Notation (JSON) essenciais para a renderização do mapa, conforme a arquitetura de visualização descrita na Seção 2.6:

1. **Um arquivo de dados brutos:** Contendo a lista completa de todas as publicações, suas coordenadas 2D e o texto do *tooltip* personalizado.
2. **Um arquivo de grade (grid):** Contendo a estrutura de dados *quadtree* de multi-resolução. Este arquivo é a inovação técnica do WizMap, pois permite ao navegador renderizar de forma eficiente milhões de pontos e agregar rótulos de forma hierárquica em diferentes níveis de *zoom*.

Estes dois arquivos JSON representam o produto final do *pipeline* de engenharia de dados, prontos para serem hospedados (conforme Seção 5.1.2) e consumidos pela interface *html* do WizMap, que renderiza o mapa de conhecimento interativo.

6 RESULTADOS E DISCUSSÃO

Este capítulo apresenta a avaliação do artefato computacional desenvolvido, conforme a metodologia DSR detalhada no Capítulo 4. Após a descrição da arquitetura e implementação do *pipeline* no Capítulo 5, esta seção tem como objetivo central executar a etapa de "Avaliação e Validação" .

A avaliação visa responder às conjecturas teóricas e ao problema de pesquisa: a integração do BERTopic com uma ferramenta de visualização interativa (WizMap) de fato melhora a exploração e a descoberta de temas emergentes no acervo de publicações?

Para conduzir esta análise, o capítulo está estruturado da seguinte forma:

- **Validação Quantitativa (Seção 6.1):** Apresenta as métricas objetivas de qualidade dos tópicos gerados, conforme planejado na Seção 4.3.1, com foco na Coerência (NPMI) e Diversidade dos tópicos.
- **Validação Qualitativa (Seção 6.2):** Realiza a análise semântica e interpretativa dos resultados, conforme a Seção 4.3.2. Esta seção analisa os principais tópicos identificados, avalia a eficácia do refinamento de rótulos com MMR e discute a validade do mapa de conhecimento gerado.
- **Limitações do Estudo (Seção 6.3):** Discute as limitações inerentes ao uso de um *dataset* de teste (NPAI) e a natureza de protótipo do artefato, indicando caminhos para a aplicação futura no acervo completo do Observatório.

6.1 Validação Quantitativa

Conforme definido na metodologia (Seção 4.3.1), a primeira etapa da avaliação do artefato consistiu na medição objetiva da qualidade dos tópicos gerados. Para isso, o *pipeline* foi avaliado em duas métricas centrais: Diversidade de Tópicos e Coerência de Tópicos (NPMI).

6.1.1 Diversidade de Tópicos

A diversidade de tópicos mede o quão distintos os tópicos são entre si, calculando a porcentagem de palavras únicas entre os 10 termos mais representativos de todos os tópicos gerados. Esta métrica é utilizada para avaliar se o modelo está produzindo agrupamentos redundantes.

No experimento, o modelo alcançou um índice de Diversidade de Tópicos de **0.9214** (ou 92,14%).

Este resultado indica a proporção de palavras-chave únicas que compõem os rótulos dos tópicos. Um índice nesta magnitude sugere uma baixa sobreposição lexical entre as representações textuais de cada tópico. Tal fato indica que a combinação de parâmetros do BERTopic (Seção 5.2.2) resultou em *clusters* com representações de palavras-chave distintas entre si.

6.1.2 Coerência de Tópicos (NPMI)

A Coerência de Tópicos, calculada pela métrica NPMI (*Normalized Pointwise Mutual Information*), avalia a interpretabilidade de um tópico medindo a frequência com que suas palavras-chave mais representativas co-ocorrem (aparecem juntas) no *corpus* de texto original. A Figura 13 ilustra a pontuação de cada tópico individualmente.

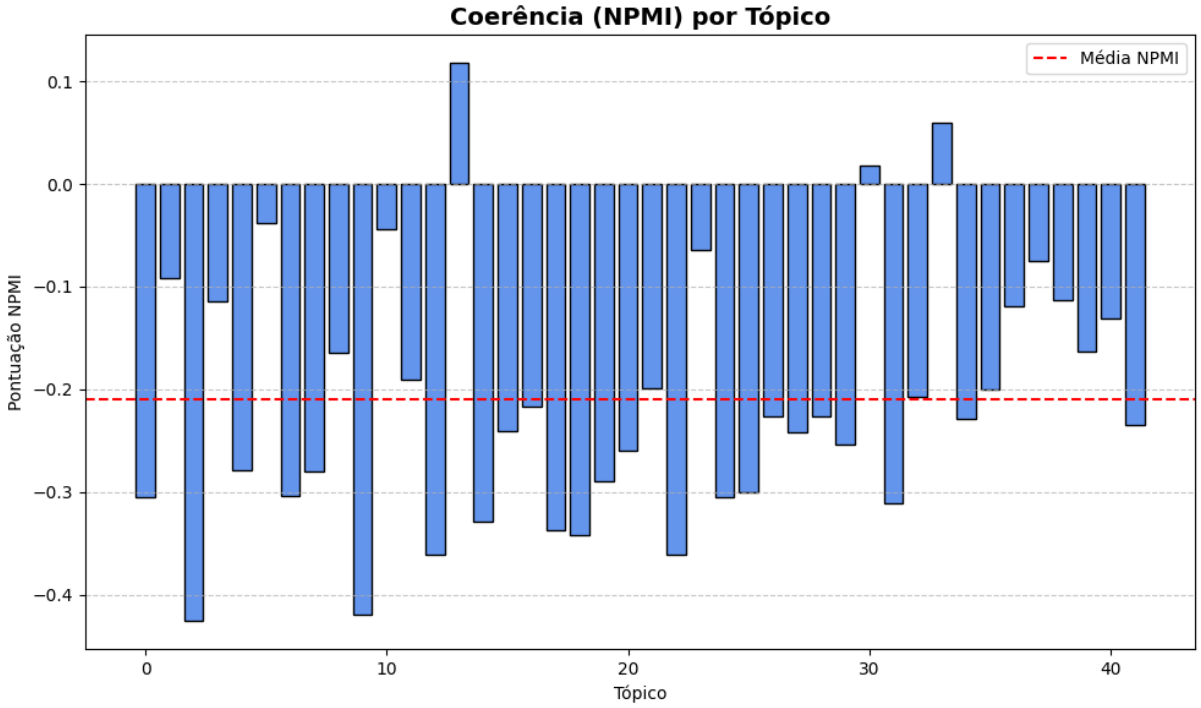

Figura 13 – Gráfico de Coerência NPMI por Tópico.

Fonte: O Autor

A pontuação média de NPMI em todos os tópicos foi de **-0.2095**.

Um escore de NPMI negativo, como o obtido, indica que as palavras-chave que definem os tópicos raramente co-ocorrem nos documentos originais. Este resultado é **metodologicamente esperado** e deve ser interpretado à luz de três fatores contextuais deste projeto:

1. **Natureza dos Dados (Textos Curtos):** O *pipeline* foi treinado exclusivamente em **títulos** de publicações. Títulos são, por natureza, textos muito curtos (baixa co-ocorrência). A métrica NPMI foi projetada para documentos longos (resumos ou textos

completos), onde palavras de um mesmo conceito (ex: “rede” e “neural”) têm espaço para aparecer juntas.

2. **Volume dos Dados (Dataset de Teste):** O modelo foi treinado em um *dataset* de teste (NPAI) de volume reduzido. Métricas de coerência estatística como a NPMI necessitam de um *corpus* massivo para encontrar padrões de co-ocorrência significativos.

3. **Metodologia (BERTopic vs. LDA):** O BERTopic é um modelo que agrupa por **similaridade semântica** (via *embeddings*) e não por co-ocorrência estatística de palavras (como o LDA). É documentado na literatura (Seção 3) que modelos baseados em *embeddings* frequentemente apresentam escores NPMI mais baixos que o LDA, mesmo que seus tópicos sejam qualitativamente mais lógicos e úteis para a interpretação humana.

Considerando o contexto deste estudo de caso, o índice de diversidade de 92,14% indica uma baixa sobreposição de palavras entre os tópicos. Por outro lado, o escore de coerência NPMI negativo (-0.2095) sugere que as palavras-chave de um mesmo tópico raramente co-ocorrem nos textos originais (títulos). A literatura aponta que a métrica NPMI é dependente de co-ocorrência, um padrão menos frequente em textos curtos e em modelos baseados em similaridade semântica (como o BERTopic), em oposição a modelos estatísticos (como o LDA). Diante disso, a avaliação do artefato prossegue com a Validação Qualitativa, que se concentra na análise da interpretabilidade e na utilidade prática dos tópicos gerados.

6.2 Validação Qualitativa

A validação qualitativa do artefato é crucial para complementar as métricas quantitativas (Capítulo 4.3.1), focando na interpretabilidade e na utilidade prática dos tópicos identificados. Esta seção avalia a capacidade do *pipeline* em gerar agrupamentos temáticos coerentes e semanticamente ricos e discute como a visualização interativa (WizMap) contribui para a exploração do conhecimento.

6.2.1 Análise dos Tópicos Identificados

A inspeção manual dos tópicos, utilizando as palavras-chave mais representativas geradas pelo MMR (Seção 5.2.3), revelou a formação de agrupamentos temáticos distintos. O modelo identificou um total de **42 tópicos** no *dataset* de teste do NPAI (excluindo o tópico -1, que representa ruído).

A Tabela 2 apresenta os 10 maiores tópicos identificados, ordenados pela quantidade de documentos, junto com suas 5 palavras-chave mais representativas.

Tabela 2 – Tópicos Mais Populosos e suas Palavras-Chave Representativas.

ID Tópico	Palavras-Chave	Nº Documentos
0	fibromialgia, dor, cronico, postural, afetivo	22
1	arquitetura, academico, educacao, professor, fundamental	13
2	robotica, educacional, escolar, robotico, crianca	12
3	design, inovador, criativo, processo, empreendedor	12
4	dengue, febre, chikungunya, zika, infeccao	12
5	software, programacao, sistema, desenvolvimento, gestao	11
6	amazonia, floresta, chuva, infeccao, ambiental	10
7	wavelet, fractal, condutor, dimensao, sinal	9
8	cancer, imune, social, colaborativo, saude	9
9	ontologia, framework, desenvolvimento, sistemas, arquitetura	9

Fonte: O Autor

A análise da Tabela 2 e da Figura 14 (que ilustra as nuvens de palavras para os tópicos mais relevantes) permite observar a capacidade do BERTopic em extrair temas emergentes e bem definidos, mesmo em um *corpus* desafiador de títulos curtos.

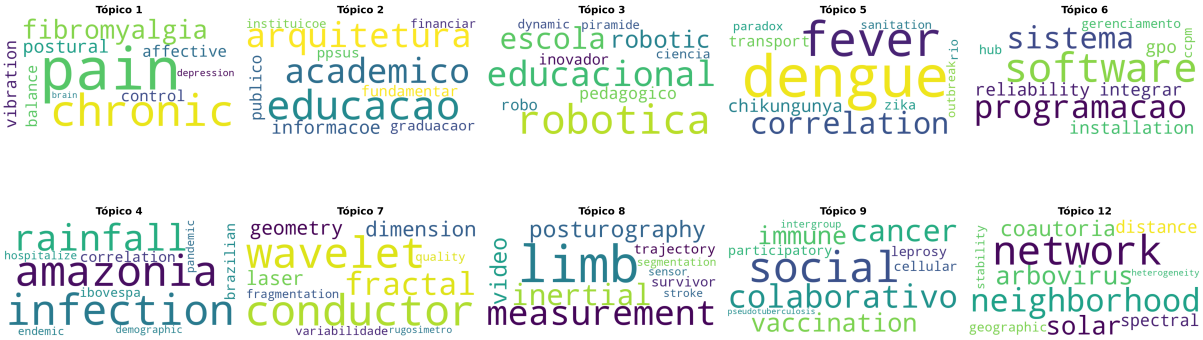

Figura 14 – Nuvens de Palavras para Tópicos Selecionados.

Fonte: O Autor

Observa-se, por exemplo:

- **Tópico 0 (Fibromialgia e Dor Crônica):** Definido por termos médicos, indicando uma área de pesquisa focada em saúde e condições crônicas.
- **Tópico 1 (Arquitetura e Educação):** Agrupa pesquisas sobre o ensino e o ambiente acadêmico, sugerindo um nicho de estudos pedagógicos na área.
- **Tópico 2 (Robótica Educacional):** Evidencia um subcampo de aplicação da robótica voltado para o contexto escolar e infantil.

- **Tópico 4 (Dengue e Arboviroses):** Um tópico de saúde pública, indicando pesquisas sobre doenças transmitidas por vetores. Nota-se a presença de termos relacionados como “dengue”, “febre”, “chikungunya” e “infecção”.

Esses exemplos ilustram que, apesar do escore NPMI médio negativo (discutido na Seção 6.1.2), os rótulos gerados com o auxílio do MMR apresentam coerência qualitativa. A aplicação do MMR busca reduzir a redundância, apresentando palavras-chave que cobrem diferentes aspectos de um mesmo tópico.

6.2.2 Discussão do Mapa de Conhecimento Interativo

O artefato final do *pipeline* é um mapa de conhecimento interativo, gerado pela ferramenta WizMap (Figura 15). Esta visualização posiciona cada publicação (representada por um ponto) no espaço bidimensional calculado pelo UMAP (Seção 5.2.2.2).

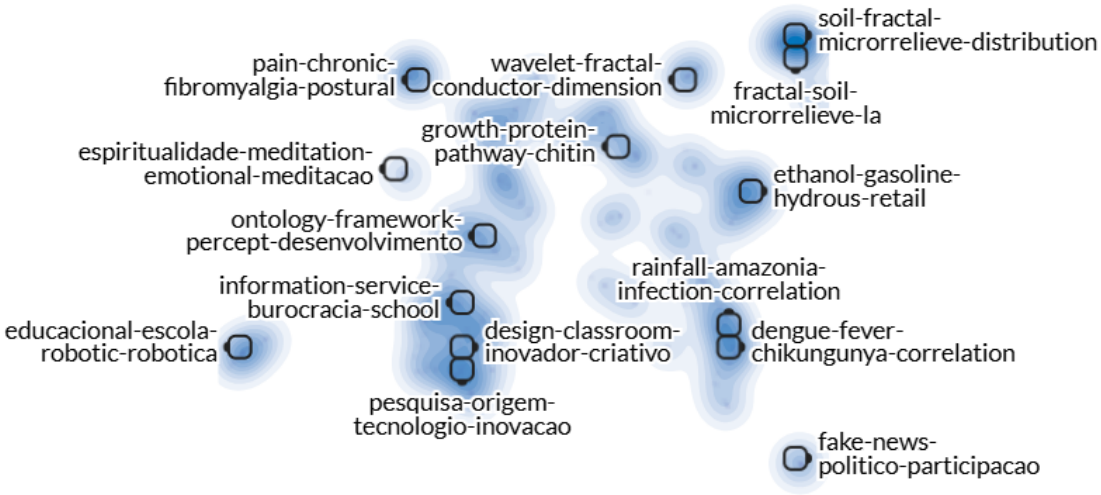

Figura 15 – Visão Geral do Mapa de Conhecimento Interativo (WizMap).

Fonte: O Autor

Na Figura 15 , as áreas de maior densidade representam os *clusters* de publicações (tópicos). Os rótulos ilustram os temas centrais desses agrupamentos. A análise visual da projeção 2D aponta os seguintes padrões espaciais:

- **Proximidade Espacial:** Tópicos com termos relacionados aparecem, em alguns casos, espacialmente próximos. Observa-se, por exemplo, a adjacência entre o tópico “dengue-fever-chikungunya-correlation” e o tópico “rainfall-amazonia-infection-

correlation". Esta justaposição sugere uma relação contextual nos dados entre arboviroses e fatores ambientais/regionais.

- **Agrupamentos Temáticos:** Algumas áreas do mapa exibem uma concentração de múltiplos tópicos. Nota-se um agrupamento que contém temas como "growth-protein-pathway-chitin" e "wavelet-fractal-conductor-dimension", sugerindo uma intersecção, no *dataset*, de pesquisas em biologia molecular com processamento de sinais ou física.
- **Distanciamento Temático:** Tópicos com semântica dissimilar, como "fake-news-politico-participacao" e "educacional-escola-robotico-robotica", aparecem em regiões espacialmente distantes no mapa, indicando a separação de domínios de pesquisa.
- **Funcionalidade de Exploração:** A interface do WizMap permite a exploração dinâmica dos dados. O usuário pode aplicar *zoom* para inspecionar *clusters* densos, ou passar o mouse sobre pontos individuais para visualizar seus *tooltips* (contendo o título e o tópico, conforme definido na Seção 5.2.4). Esta funcionalidade de navegação oferece um método de exploração dos temas e suas relações espaciais, que difere da análise de tabelas ou listas estáticas.

O mapa, portanto, funciona como uma interface navegável para os dados gerados pelo *pipeline*. Ele apresenta a organização dos tópicos e sua distribuição espacial, permitindo ao usuário explorar o *dataset* e as relações temáticas de forma visual.

Para explorar o mapa de conhecimento interativo online, clique aqui: Mapa interativo de conhecimento – WizMap

6.3 Limitações do Estudo

A validade e a generalização dos resultados apresentados neste capítulo devem ser consideradas no contexto de limitações metodológicas específicas, inerentes à natureza deste estudo de caso como um experimento de DSR.

- **Conjunto de Dados (Prototipagem vs. Produção):** A principal limitação deste estudo é a fonte dos dados. O *pipeline* foi desenvolvido e validado sobre um *dataset* de teste estático (um *dump* do NPAI) e não sobre o acervo de dados completo e dinâmico do **Observatório de dados públicos de ciência e tecnologia da Bahia**. Embora este *dataset* tenha sido suficiente para a validação do protótipo (o artefato), os resultados aqui apresentados (como a quantidade e a natureza dos tópicos) não são representativos do acervo real do Observatório.
- **Dependência dos Hiperparâmetros:** A configuração do *pipeline*, especialmente os hiperparâmetros de modelagem (Seções 5.2.2.2 e 5.2.2.3), foi diretamente influenciada

pelo volume reduzido do *dataset* de teste. Parâmetros como `n_neighbors=5` (UMAP) e `min_cluster_size=4` (HDBSCAN) foram ajustados para permitir a descoberta de tópicos de nicho em um *corpus* pequeno. Estes parâmetros não seriam diretamente transferíveis para o acervo completo do Observatório, que, sendo ordens de magnitude maior, exigiria uma nova etapa de *tuning* (ajuste fino).

- **Natureza do Texto de Entrada:** O *pipeline* foi treinado utilizando exclusivamente os **títulos** das publicações. Textos curtos, por definição, oferecem baixa co-ocorrência de palavras. Este fator contextual é a explicação técnica mais provável para o escore negativo da métrica NPMI (Seção 6.1.2), que é uma métrica estatística dependente de co-ocorrência. Os resultados de coerência poderiam ser diferentes se o modelo fosse treinado em textos mais longos, como os resumos (*abstracts*).

- **Limitação da Métrica de Coerência:** Conforme discutido na Seção 6.1.2, a própria métrica NPMI apresenta uma limitação contextual. Ela foi projetada para avaliar modelos estatísticos (como o LDA) e pode não ser a ferramenta de avaliação quantitativa ideal para modelos baseados em similaridade semântica (como o BERTopic). A validação qualitativa (Seção 4.3.2) foi, portanto, necessária para avaliar a interpretabilidade dos tópicos.

- **Natureza do Artefato (DSR):** O artefato desenvolvido é um **protótipo** executado em um ambiente de desenvolvimento (Google Colab). Ele não está integrado ao *pipeline* de produção, ao banco de dados dinâmico ou à interface de usuário existente do Observatório (descritos na Seção 5.1.1). A sua função neste estudo é demonstrar a *viabilidade* da metodologia, e não apresentar uma solução de *software* em produção.

7 CONCLUSÃO

Este Estudo de Caso propôs-se a enfrentar o desafio da exploração de grandes acervos de publicações científicas, cujas abordagens tradicionais de busca lexical (palavras-chave) limitam a descoberta de conhecimento. O objetivo central foi projetar, desenvolver e avaliar um artefato computacional, seguindo os princípios da DSR, capaz de transformar o acervo textual do **Observatório de dados públicos de ciência e tecnologia da Bahia** em um mapa de conhecimento interativo.

Para atingir este objetivo, foi desenvolvido um *pipeline* computacional (Capítulo 5) que integra técnicas modernas de PLN. O *pipeline* executa um fluxo de quatro etapas: (1) pré-processamento de dados textuais (títulos); (2) modelagem de tópicos com BERTopic, utilizando *embeddings* `Sentence-BERT`, redução de dimensionalidade UMAP e *clusterização* HDBSCAN; (3) refinamento de rótulos com MMR para reduzir a redundância semântica; e (4) exportação dos dados processados (coordenadas 2D e metadados) para a ferramenta de visualização WizMap.

A avaliação do artefato (Capítulo 6), realizada sobre um *dataset* de teste (NPAI), demonstrou a viabilidade técnica da solução. A validação quantitativa (Seção 4.3.1) indicou uma alta diversidade de tópicos (0.9214), sugerindo baixa sobreposição lexical entre os temas. O escore de coerência NPMI (-0.2095) foi negativo, um resultado metodologicamente discutido como sendo uma limitação da métrica NPMI quando aplicada a modelos semânticos (como o BERTopic) e a textos curtos (títulos), onde a co-ocorrência de palavras é estatisticamente baixa.

A validação qualitativa (Seção 4.3.2) complementou esta análise, demonstrando que, apesar do escore NPMI, os tópicos gerados apresentaram coerência semântica interpretável (ex: “Dengue e Arboviroses”, “Robótica Educacional”). A visualização final no WizMap (Figura 15) demonstrou a capacidade do artefato em posicionar tópicos semanticamente próximos em regiões adjacentes do mapa, oferecendo uma interface navegável para a exploração da paisagem de pesquisa.

Considera-se, portanto, que os objetivos delineados pela metodologia DSR foram atingidos no que tange à construção e validação de um protótipo funcional. O artefato representa uma solução viável para o problema de descoberta de conhecimento no acervo do Observatório.

7.1 Trabalhos Futuros

As limitações identificadas durante o desenvolvimento (Seção 6.3) e a natureza de protótipo deste estudo abrem caminho para diversas frentes de trabalho futuro:

- **Validação em Larga Escala:** Aplicar o *pipeline* sobre o acervo de dados completo e dinâmico do Observatório de dados públicos de ciência e tecnologia da Bahia, o que constitui o próximo passo natural para validar a solução em um ambiente de produção.
- **Ajuste Fino de Hiperparâmetros:** A execução em larga escala exigirá uma nova etapa de *tuning* (ajuste fino) dos hiperparâmetros (ex: `n_neighbors` no UMAP e `min_cluster_size` no HDBSCAN) para adequá-los a um volume de dados massivo.
- **Enriquecimento do Texto de Entrada:** Incluir os **resumos** (*abstracts*) das publicações, além dos títulos, no processo de modelagem. Textos mais longos podem fornecer mais contexto, potencialmente gerando tópicos mais ricos e impactando positivamente as métricas de coerência estatística como o NPMI.
- **Integração do Artefato:** Evoluir o protótipo (atualmente executado no Google Colab) para um módulo de *software* integrado à arquitetura de produção do Observatório (Seção 5.1.1), permitindo a atualização automática e periódica do mapa de conhecimento.

REFERÊNCIAS

ANGELOV, D. **Top2Vec: Distributed Representations of Topics**. 2020. Disponível em: https://arxiv.org/abs/2008.09470. Citado nas páginas 24 e 28.

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 23.

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 na página 27.

CARBONELL, J. G.; GOLDSTEIN, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: **Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. Melbourne, Australia: ACM, 1998. p. 335–336. Disponível em: https://doi.org/10.1145/290941.291025. Citado na página 28.

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 13, 23 e 24.

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 22.

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 13, 20 e 21.

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 nas páginas 35 e 38.

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 13, 16, 17, 32, 33 e 34.

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 na página 23.

GERASIMENKO, N. *et al.* Incremental topic modeling for scientific trend topics extraction. In: **Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"**. [S.l.: s.n.], 2023. p. 88–103. Citado nas páginas 32, 33 e 34.

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 13, 14, 24, 25, 28, 33 e 37.

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 23.

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 23.

JOACHIMS, T. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: . [S.l.: s.n.], 1997. p. 143–151. Citado na página 28.

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 39, 40 e 41.

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 25, 32, 33, 34 e 37.

JURAFSKY, D.; MARTIN, J. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Pearson Prentice Hall, 2009. (Prentice Hall series in artificial intelligence). ISBN 9780131873216. Disponível em: https://books.google.com.br/books?id=fZmj5UNK8AQC. Citado na página 16.

KIM, K.; KOGLER, D. F.; MALIPHOL, S. Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTopic. **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 nas páginas 15, 32, 33 e 34.

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 29.

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 26.

MANNING, C.; SCHUTZE, H. **Foundations of Statistical Natural Language Processing**. MIT Press, 1999. (Foundations of Statistical Natural Language Processing). ISBN 9780262133609. Disponível em: https://books.google.com.br/books?id=YiFDxbEX3SUC. Citado na página 17.

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 25, 26 e 29.

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 33 e 34.

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 23.

MIKOLOV, T. *et al.* **Efficient Estimation of Word Representations in Vector Space**. 2013. Disponível em: https://arxiv.org/abs/1301.3781. Citado nas páginas 17 e 18.

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 13.

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 17.

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 23.

RADFORD, A.; NARASIMHAN, K. Improving language understanding by generative pre-training. In: . [s.n.], 2018. Disponível em: https://api.semanticscholar.org/CorpusID:49313245. Citado na página 20.

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019. Disponível em: https://arxiv.org/abs/1908.10084. Citado nas páginas 21, 22 e 25.

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 39 e 40.

SIA, S.; DALMIA, A.; MIELKE, S. J. **Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!** 2020. Disponível em: https://arxiv.org/abs/2004.14914. Citado na página 24.

TAYLOR, W. L. Cloze procedure: A new tool for measuring readability. **Journalism Quarterly**, v. 30, p. 415–433, 1953. Citado na página 20.

VASWANI, A. *et al.* **Attention Is All You Need**. 2017. Disponível em: https://arxiv.org/abs/1706.03762. Citado nas páginas 13 e 19.

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 14, 29, 30 e 31.

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 na página 33.

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 13, 15, 16, 18 e 23.