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. Hugo Saba Pereira Cardoso
Examinador interno (DCET-I/UNEB)

Prof. Dr. Vagner de Souza Fonseca
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 buscou transformar a maneira como pesquisadores exploram acervos científicos, revelando conexões semânticas e tópicos emergentes de forma intuitiva. A proposta foi 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* para aumentar a interpretabilidade. Os resultados são integrados à ferramenta *WizMap*, permitindo a exploração interativa da estrutura dos dados. A validação experimental indicou a geração de agrupamentos semanticamente coesos e a viabilidade da ferramenta WizMap na projeção espacial de domínios correlatos. Os resultados indicam que a técnica de *embeddings* mostrou-se adequada para organizar acervos científicos, oferecendo uma alternativa prática aos métodos estatísticos tradicionais.

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 sought to transform the way researchers explore scientific collections, revealing semantic connections and emerging topics in an intuitive manner. The proposal was 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 via *BERTopic* (using contextual *embeddings*, *UMAP*, and *HDBSCAN*) with label refinement via *MMR* to enhance interpretability. The results are integrated into the *WizMap* tool, enabling the interactive exploration of the data structure. The experimental validation indicated the generation of semantically cohesive clusters and the viability of the WizMap tool in the spatial projection of correlated domains. The results indicate that the *embedding* technique proved adequate for organizing scientific collections, offering a practical alternative to traditional statistical methods.

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

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API *Application Programming Interface*. 40

BERT *Bidirectional Encoder Representations from Transformers*. 8, 13, 16, 20–22, 32, 33, 44

BERTopic *Bidirectional Encoder Representations Transformers for Topic Modeling*. 13–16, 24, 25, 28, 31–37, 39, 41, 45, 46, 54

c-TF-IDF *Class-based Term Frequency-Inverse Document Frequency*. 24, 28, 32, 34, 36, 38, 42, 45, 46

CBOW *Continuous Bag-of-Words*. 8, 18

DBMS Database Management System. 40

DSR *Design Science Research*. 14, 35, 36, 38, 39, 41, 42, 48

ELMo *Embeddings from Language Models*. 8, 21

ETL Extract, Transform and Load. 39, 40

GPT *Generative Pre-trained Transformer*. 8, 20, 21

HDBSCAN *Hierarchical Density-Based Spatial Clustering of Applications with Noise*. 26, 32, 33, 36, 44

IA Inteligência Artificial. 13

JCR *Journal Citation Reports*. 39, 40

JSON JavaScript Object Notation. 46, 47

KDE *Kernel Density Estimation*. 30

LDA *Latent Dirichlet Allocation*. 13, 15, 22–24, 32, 54

LSA *Latent Semantic Analysis*. 22, 23

MLM *Masked Language Model*. 20

MMR *Maximal Marginal Relevance*. 28, 36, 38, 39, 41, 42, 45, 46, 48, 52

NMF *Non-negative Matrix Factorization*. 32

NPMI *Normalized Pointwise Mutual Information*. 37, 38, 49, 52, 54, 57

PCA *Principal Component Analysis*. 33

PLN Processamento de Linguagem Natural. 13, 15–19, 22, 37, 41

PLSA *Probabilistic Latent Semantic Analysis*. 22, 23

RNN *Recurrent Neural Network*. 19

SBERT *Sentence-BERT*. 8, 21, 22, 25, 33, 36

SVD *Singular Value Decomposition*. 22

t-SNE *t-distributed Stochastic Neighbor Embedding*. 29

TF-IDF *Term Frequency-Inverse Document Frequency*. 24, 28, 45

UMAP *Uniform Manifold Approximation and Projection*. 8, 25, 26, 29, 32, 33, 36, 44, 46, 52

WebGL *Web Graphics Library*. 30

WizMap *Knowledge Map Visualization Tool*. 8, 14, 15, 29–31, 35, 36, 38, 39, 42, 47, 50, 52, 53

SUMÁRIO

1 INTRODUÇÃO

O cenário da pesquisa científica registra um crescimento exponencial na produção acadêmica nas últimas décadas, gerando um volume de dados que impõe dificuldades aos métodos convencionais de organização e análise. Para navegar nesse vasto conjunto de informações, pesquisadores confiam majoritariamente em plataformas de busca baseadas em palavras-chave, como *Web of Science*[1], *Scopus*[2] e *IEEE Xplore*[3]. 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 **estruturas temáticas latentes** na literatura (GALLI *et al.*, 2024).

Segundo Datchanamoorthy, S e B (2023), a complexidade desses acervos e a necessidade de uma análise semântica 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 mais robustas para a gestão do conhecimento acadêmico. Estudos como os de Mohammadi e Karami (2020) e Xie *et al.* (2020), que analisaram agrupamentos de pesquisa em *Big Data* por meio de mineração de texto, destacam a relevância da integração de técnicas de modelagem de tópicos com modelos de linguagem baseados em transformadores.

A arquitetura de Transformadores, apresentada por Vaswani *et al.* (2017), introduziu o mecanismo de autoatenção (*self-attention*) no campo do PLN. Modelos subsequentes, como o BERT[4], proposto por Devlin *et al.* (2019), permitiram capturar relações contextuais bidirecionais em textos. A partir dessa base, consolidaram-se os *embeddings*, representações vetoriais que codificam o significado semântico de palavras e frases, superando as limitações de modelos tradicionais de *bag-of-words* e de abordagens probabilísticas clássicas como o *Latent Dirichlet Allocation* (LDA) (GALLI *et al.*, 2024).

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

O foco desta pesquisa reside na instanciação de um artefato computacional para o mapeamento visual e interativo de publicações científicas. A proposta central consiste

[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.

na integração da modelagem de tópicos do BERTopic com a ferramenta de visualização WizMap[6] (WANG; HOHMAN; CHAU, 2023), aplicada ao contexto de dados regionais. O estudo de caso utiliza como base de validação (Prova de Conceito) o acervo do Núcleo de Pesquisa Aplicada e Inovação (NPAI), visando demonstrar a viabilidade da solução para futura expansão ao Observatório de Dados Públicos de Ciência e Tecnologia da Bahia[7], plataforma que coleta informações de fontes como Currículos Lattes e *OpenAlex*, fundamental na gestão do conhecimento científico do estado.

A metodologia *Design Science Research* (DSR) é adotada como a estrutura condutora deste estudo, orientando a criação do *pipeline* como o artefato principal. O objetivo é transformar a base de dados textual do NPAI/Observatório, atualmente explorada via listas estáticas, em um mapa de conhecimento navegável. Nessa solução, o BERTopic é empregado para extrair padrões temáticos a partir de títulos de publicações, e o WizMap é utilizado para a exploração visual.

1.1 Contribuições do Trabalho

Este estudo caracteriza-se como uma pesquisa de natureza aplicada, enquadrada na DSR como a instanciação de um artefato tecnológico em um contexto específico. Diferentemente de trabalhos que visam a inovação algorítmica pura, a contribuição central deste trabalho é técnica-regional, residindo na orquestração de componentes do estado da arte para oferecer uma arquitetura alternativa para a descoberta de temas no acervo de produção científica da Bahia.

As contribuições específicas podem ser estratificadas em:

1. **Validação de Pipeline em Cenário Controlado (NPAI):** O estudo implementa e valida o *pipeline* utilizando como base de teste o acervo do Núcleo de Pesquisa Aplicada e Inovação (NPAI). Esta aplicação piloto demonstrou a viabilidade técnica de transformar metadados estáticos em mapas de conhecimento navegáveis, servindo como prova de conceito (PoC) para a expansão ao acervo completo do Observatório.
2. **Ferramenta de Orientação e Visibilidade para o Ecossistema Regional:** O artefato atua como um instrumento de direcionamento estratégico para pesquisadores e gestores da Bahia. Como uma alternativa à busca lexical, a visualização interativa amplia a visibilidade de áreas interdisciplinares frequentemente ofuscadas, permitindo que o usuário identifique conexões não óbvias entre diferentes domínios. Essa exploração por vizinhança semântica apoia diretamente o levantamento de novas hipóteses investigativas, revelando a topologia do conhecimento que servirá de norte para o fomento e desenvolvimento da produção científica local.

[6] Disponível em: https://github.com/poloclub/wizmap.
[7] Disponível em: https://simcc.uesc.br/observatorio.

1.2 Organização do Trabalho

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 baseada em DSR. O Capítulo 5 descreve o Projeto de Desenvolvimento e a arquitetura do *pipeline*. Por fim, o Capítulo 6 discute os Resultados e a avaliação da solução proposta.

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. **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).

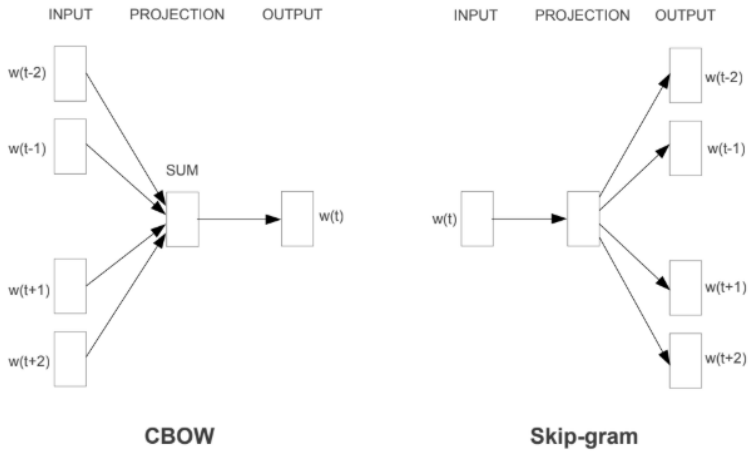

Figura 1 – Arquiteturas CBOW e *Skip-gram*.

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.

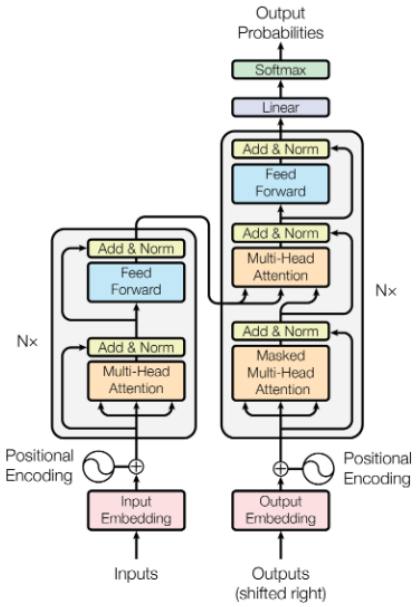

Figura 2 – Arquitetura do modelo Transformador.

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 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).

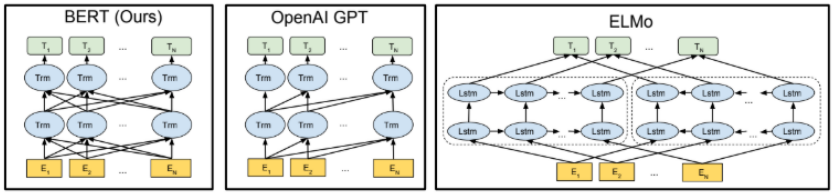

Figura 3 – Diferenças nas arquiteturas de pré-treinamento. BERT é bidirecional, GPT é unidirecional (da esquerda para a direita) e ELMo.

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 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.

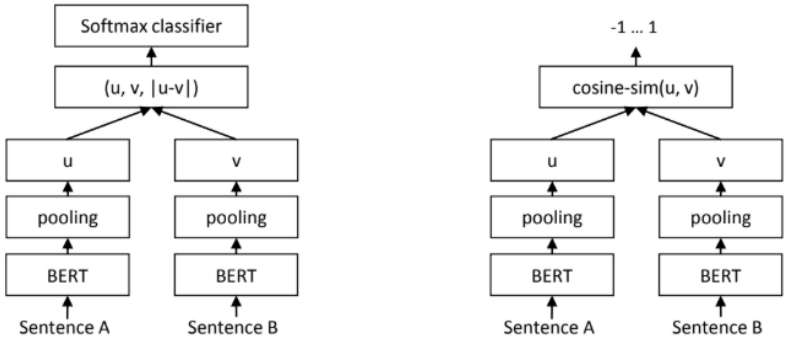

Figura 4 – Arquitetura de inferência do SBERT para computar similaridade.

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.

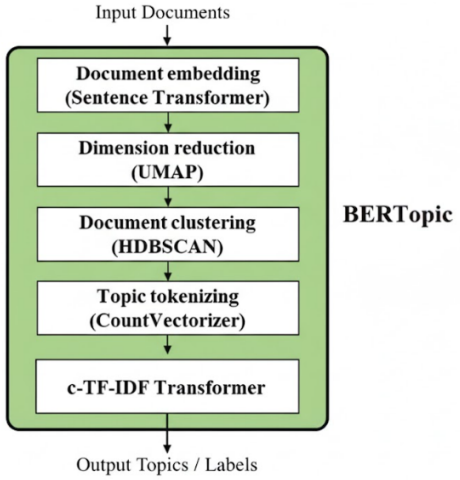

Figura 5 – Diagrama esquemático do *pipeline* BERTopic.

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 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).

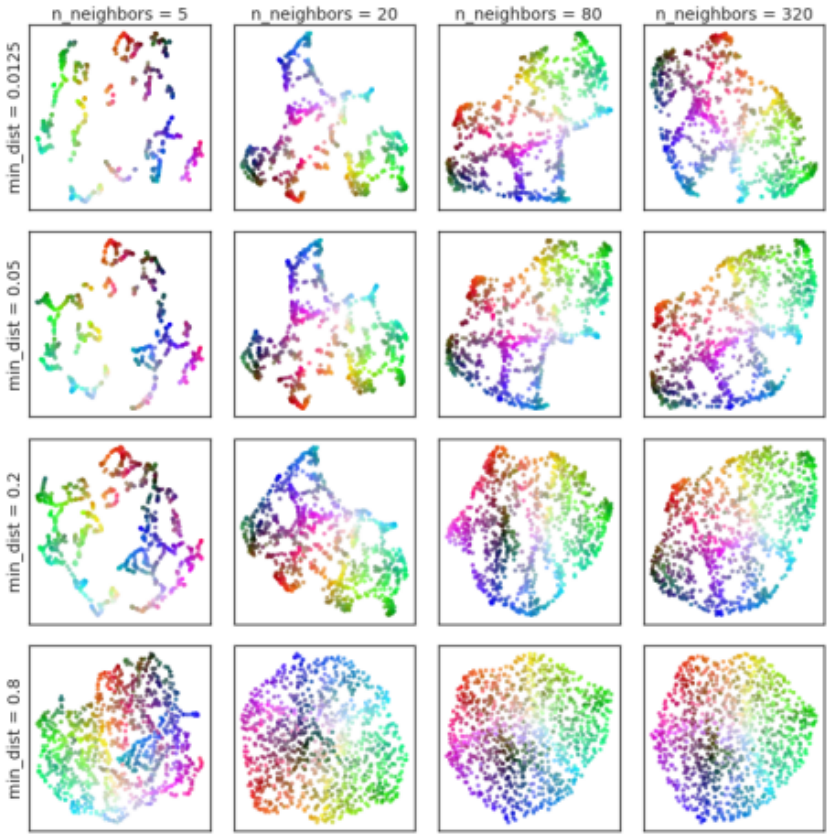

Figura 6 – Diagrama ilustrativo do UMAP, demonstrando a relação entre os hiperparâmetros `n_neighbors` e `min_dist` e a representação visual dos dados.

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.

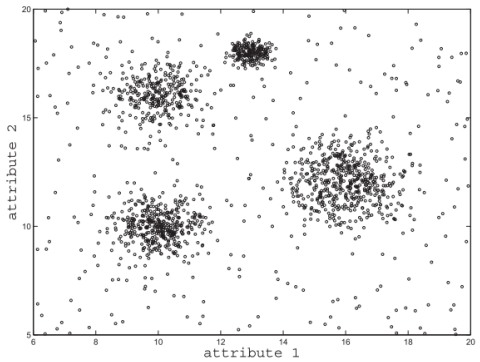

Figura 7 – Figura ilustrativa de um *dataset* sintético com quatro *clusters* e ruído de fundo.

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. Como apontam Wang, Hohman e Chau (2023), esses dados caracterizam-se por sua baixa interpretabilidade humana (opacidade), alta dimensionalidade e pelo grande volume dos conjuntos de dados modernos.

Para tornar esses vetores complexos inteligíveis, aplicam-se técnicas de redução de dimensionalidade, como o UMAP 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 estática em larga escala permanece limitada. Em conjuntos de dados com milhões de pontos, a inspeção visual ponto a ponto torna-se inviável para a compreensão da estrutura global.

Alternativas, como gráficos de contorno (*contour plots*), podem resumir a distribuição global, mas restringem a exploração das estruturas locais de um *embedding*. Para conectar a visão global à exploração local, ferramentas de visualização interativa tornam-se necessárias.

Neste contexto, Wang, Hohman e Chau (2023) desenvolveram o WizMap[15], uma ferramenta de visualização interativa escalável projetada para a interpretação de embeddings massivos. A ferramenta emprega um design de interação familiar semelhante a mapas geográficos (*map-like interaction design*), permitindo a navegação pelo espaço semântico através de operações 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

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

Controle (*Control Panel*), para customização da visualização (WANG; HOHMAN; CHAU, 2023, p. 1).

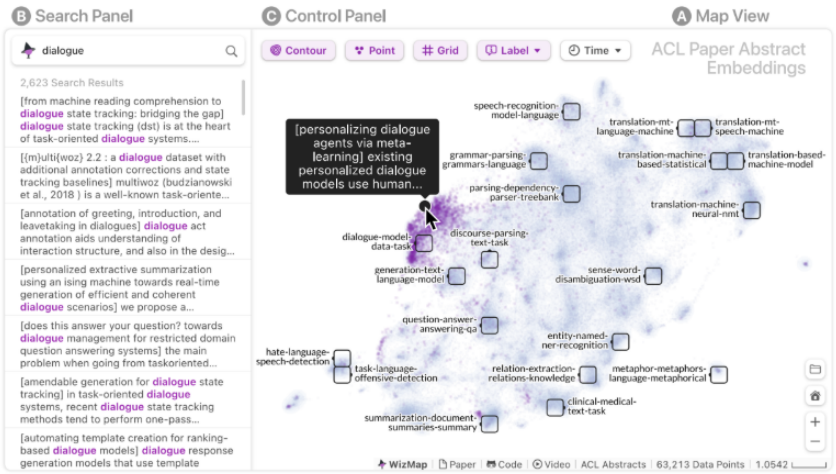

Figura 8 – Interface da ferramenta WizMap e seus componentes principais.

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.

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 agrega as informações de baixo para cima, permitindo que os rótulos textuais

[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.

se ajustem dinamicamente ao nível de zoom do usuário. Essa técnica de agregação multi-resolução permite transitar de uma visão macro (tópicos gerais) para uma visão micro (documentos individuais) de forma fluida. (WANG; HOHMAN; CHAU, 2023, p. 2-3).

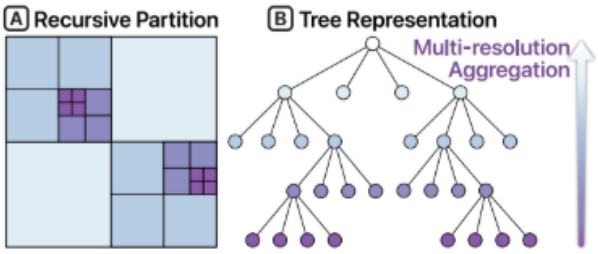

Figura 9 – Estrutura de dados *Quadtree* usada pelo WizMap para agregação multi-resolução. (A) Particionamento recursivo do espaço **2d**. (B) Representação em árvore.

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

A integração de interfaces interativas de multi-resolução com técnicas de redução de dimensionalidade é fundamental para traduzir a saída matemática de modelos como o BERTopic em mapas de conhecimento navegáveis, facilitando a identificação de padrões latentes em grandes bases textuais.

É importante estabelecer uma distinção conceitual sobre o termo "Mapa de Conhecimento" utilizado nesta pesquisa. Diferentemente da Representação de Conhecimento clássica, que se baseia em ontologias formais e relações lógicas explícitas (como grafos de conhecimento), esta abordagem fundamenta-se no paradigma do Mapeamento da Ciência. Segundo Börner (2010), mapas de ciência são representações visuais que utilizam a metáfora espacial para comunicar a estrutura e a evolução de campos científicos.

Neste contexto, a validade do "mapa" não advém de regras ontológicas pré-definidas, mas sim da proximidade semântica. Assume-se a premissa da Hipótese Distribucional, onde a proximidade espacial entre dois pontos (documentos) no gráfico do WizMap representa uma alta similaridade de conteúdo (embeddings). Portanto, operacionalmente, define-se o artefato como um mapa de conhecimento no sentido de navegação e descoberta de estruturas latentes, permitindo ao usuário perceber agrupamentos (clusters) como domínios de conhecimento consolidados.

2.7 Síntese do Referencial Teórico

Em síntese, este capítulo consolidou a fundamentação teórica necessária para o desenvolvimento da pesquisa. Inicialmente, discutiram-se as limitações dos métodos tradicionais de recuperação da informação, evidenciando a necessidade de abordagens baseadas em Processamento de Linguagem Natural, especificamente o uso de embeddings contextuais e da arquitetura de Transformadores. Na sequência, o detalhamento técnico

do modelo BERTopic e da ferramenta WizMap forneceu os subsídios para a construção do artefato computacional proposto. Com o referencial estabelecido, o capítulo seguinte contextualiza este estudo em relação aos trabalhos correlatos e ao estado da arte.

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).	Reporta desempenho superior do BERTopic frente a baselines tradicionais (LDA, NMF) especificamente na análise de resumos (abstracts) e textos noticiosos, validando a arquitetura para documentos com contexto semântico rico.
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).

3.2 Considerações Finais

Conforme evidenciado na Tabela 1, a arquitetura baseada em embeddings apresenta vantagens estruturais sobre abordagens probabilísticas clássicas. A análise comparativa realizada por Jung *et al.* (2024) corrobora os dados apresentados, fundamentando a escolha do BERTopic para este trabalho ao demonstrar sua superioridade na captura de nuances semânticas em detrimento de modelos como o LDA, que dependem de contagens exatas de palavras (*Bag-of-Words*) e apresentam dificuldades com sinonímia e polissemia (GALLI *et al.*, 2024).

No entanto, impõe-se uma distinção crucial de escopo para a presente pesquisa. Enquanto Jung *et al.* (2024) validaram o modelo em *corpora* de textos longos, este estudo submete a técnica a um cenário de maior esparsidade informacional: a modelagem de títulos curtos. Neste contexto, a literatura indica que métricas puramente estatísticas de coocorrência (como o NPMI) podem divergir da interpretabilidade humana. Por essa razão, a validação deste artefato priorizará a coerência interpretativa (C_v) e a capacidade de agrupamento semântico, adaptando os indicadores de sucesso observados em textos longos para a realidade dos dados curtos.

Portanto, alinhado aos princípios da DSR, este trabalho não visa um *benchmarking* exaustivo de algoritmos, mas sim a construção de um artefato de visualização inovador. Adota-se, assim, a arquitetura de estado da arte (*State-of-the-Art*) já validada na literatura, concentrando o esforço experimental na tunagem de hiperparâmetros e na integração com a interface visual (WizMap) para superar os desafios impostos pela brevidade dos títulos.

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.

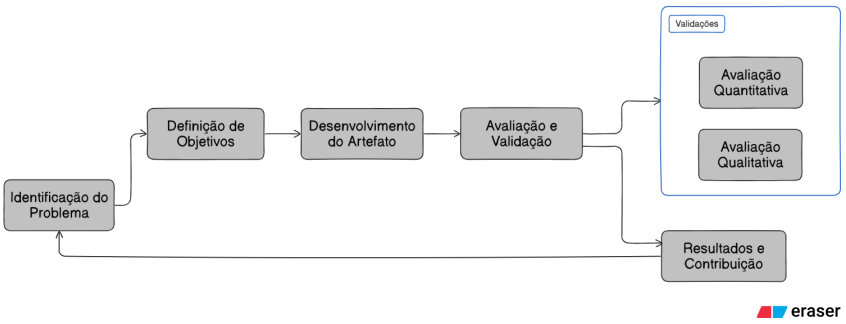

Figura 10 – Adaptação da Design Science Research para este projeto.

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 pesquisas 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, oferecendo uma modalidade de exploração visual e semântica complementar à busca tradicional por palavras-chave, permitindo a identificação de padrões não-lineares no acervo?

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.

Para fins de escopo deste trabalho, adota-se uma definição operacional de Mapa de Conhecimento como: uma interface visual interativa que projeta um espaço vetorial de alta dimensionalidade em um plano 2D, onde a posição relativa dos documentos codifica sua afinidade semântica. O artefato não visa estabelecer relações formais (como "A é causa de B"), mas sim revelar a topologia do acervo (quais temas são vizinhos, quais são isolados), facilitando a exploração exploratória em detrimento da busca exata.

É importante delimitar que o foco desta pesquisa reside na validação da arquitetura do software e na eficácia dos algoritmos de PLN propostos. Portanto, o problema não abrange, nesta etapa do ciclo de Design Science Research, a integração em tempo real com o banco de dados de produção do Observatório, nem a realização de um estudo bibliométrico exaustivo de todo o ecossistema científico da Bahia. O problema restringe-se à prototipagem e validação do pipeline em um ambiente controlado de dados.

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 Estratégia de Avaliação do Artefato

Para aferir a qualidade e a utilidade do artefato desenvolvido, adota-se uma estratégia mista, alinhada aos conceitos de Avaliação e Validação em DSR. Neste trabalho, define-se Avaliação (*Evaluation*) como a mensuração objetiva do desempenho do modelo computacional através de métricas de referência. Por sua vez, a Validação é entendida como a verificação qualitativa da coerência semântica e da aplicabilidade do artefato em representar corretamente os domínios de conhecimento, realizada através da inspeção por especialista.

4.3.1 Avaliação de Desempenho (Métricas)

A avaliação de desempenho (Métricas) foca em avaliar objetivamente a qualidade dos tópicos gerados pelo *pipeline* do BERTopic. Serão utilizadas métricas consolidadas na literatura de PLN para mensurar tanto a consistência interna quanto a distinção entre os grupos:

- **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 formam um conceito semântico lógico (JUNG *et al.*, 2024). Para esta avaliação, serão empregadas duas métricas com abordagens complementares:

  1. Coerência C_v: Esta métrica utiliza janelas deslizantes (*sliding windows*) e similaridade de cosseno para calcular a relação contextual entre as palavras principais do tópico. O escore varia de 0 a 1, onde valores mais altos indicam maior consistência estrutural e interpretabilidade humana.

  2. *Normalized Pointwise Mutual Information* (NPMI) (*Normalized Pointwise Mutual Information*): O cálculo da NPMI foca na probabilidade estatística de coocorrência léxica. Consiste em extrair os N termos mais representativos de cada tópico (ex: $N = 10$) e calcular a probabilidade de sua aparição conjunta nos documentos do *corpus* original, normalizada pela probabilidade de suas ocorrências individuais. A pontuação varia de +1 (coocorrência perfeita) a -1 (nunca aparecem juntas).

- **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 pontuação de coerência) e distintos entre si (alta diversidade). Conforme (GROOTENDORST, 2022, p. 5, Traduzido), a métrica é calculada como a porcentagem de palavras únicas em todos os tópicos 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 Semântica (Inspeção por Especialista)

Dada a natureza não supervisionada da modelagem de tópicos, métricas quantitativas nem sempre refletem a intelegibilidade humana. Portanto, a validação foi complementada por uma análise qualitativa conduzida pelos pesquisadores responsáveis (autor e orientador), atuando como especialistas do domínio técnico.

O protocolo de inspeção adotou uma abordagem heurística baseada na verificação de *Coerência Intratópico*. Foram selecionados aleatoriamente 20% dos tópicos gerados (amostragem estratificada por densidade) para a verificação manual de dois critérios fundamentais:

1. **Alinhamento de Rótulos:** Verificar se as palavras-chave geradas pelo algoritmo (c-TF-IDF/MMR) descrevem adequadamente o conteúdo dos títulos dos documentos agrupados naquele *cluster*.
2. **Detecção de Intrusos:** Identificar a presença de documentos semanticamente desconexos dentro de um agrupamento (ex: um artigo de 'Química' classificado erroneamente em um tópico de 'Direito').

Essa etapa visa assegurar a *validade interna* do artefato, garantindo que o "mapa" gerado possua lógica semântica antes de ser submetido a futuros testes de usabilidade com público externo. Adicionalmente, será conduzida uma comparação direta entre os rótulos gerados pelo método padrão (c-TF-IDF) e os refinados pelo MMR, para aferir o ganho na interpretabilidade humana.

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 segue a abordagem metodológica da DSR, conforme detalhado no 4. O objetivo é a construção e validação de um *pipeline* computacional projetado para complementar a análise de publicações científicas na plataforma do Observatório de dados públicos de ciência e tecnologia da Bahia[1].

Conforme descrito por Santos *et al.* (2024), Jorge *et al.* (2025), o Observatório possui uma arquitetura de dados que integra fontes heterogêneas, como a Plataforma Lattes[2], a Plataforma Sucupira[3], o OpenAlex[4] e o *Journal Citation Reports* (JCR)[5]. O sistema atual utiliza um processo de Extract, Transform and Load (ETL) com a ferramenta Apache Hop[6] para consolidar as informações em um banco de dados PostgreSQL[7], empregando técnicas de recuperação de informações baseadas em termos e palavras-chave.

A busca lexical constitui o mecanismo padrão para a recuperação de informações na plataforma. No entanto, a variação terminológica em domínios científicos pode dificultar a identificação de conexões temáticas não explícitas. Nesse contexto, 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, a construção de um *pipeline* que representa uma evolução para a arquitetura do Observatório. A solução proposta introduz a modelagem de tópicos com o BERTopic, o refinamento de rótulos com MMR e a visualização interativa com o WizMap. O foco é apresentar uma alternativa à lista de resultados tradicional, permitindo a navegação visual pelas áreas de pesquisa e a identificação de relações entre diferentes campos do conhecimento.

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 trabalho.

[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/

5.1.1 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.

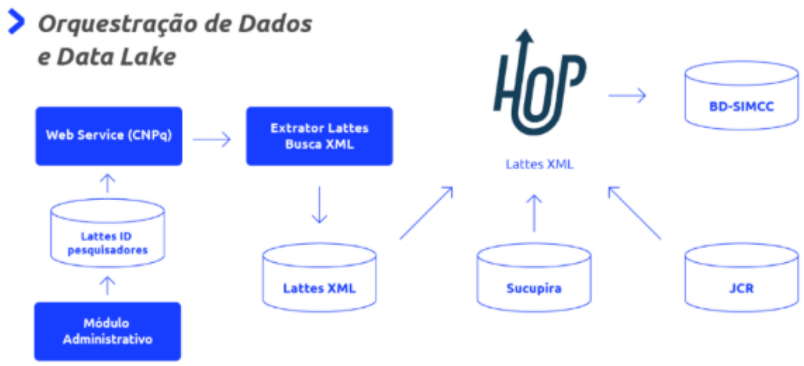

Figura 11 – Arquitetura Geral do Observatório (SIMCC).

Fonte: Jorge *et al.* (2025)

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.

5.1.2 Pipeline de Modelagem e Visualização

Sobre a base conceitual do Observatório, este projeto implementa um *pipeline* voltado à descoberta e análise semântica do conhecimento. Considerando a metodologia DSR adotada, esta etapa corresponde ao desenvolvimento e instanciação do artefato.O pipeline foi construído e validado em um ambiente de prototipagem computacional (ambiente de testes), configurado como uma Prova de Conceito para demonstrar a viabilidade técnica da solução proposta.

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`.
  - **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*.

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

- **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.

Para fins de transparência, reprodutibilidade científica e validação da arquitetura modular descrita, todo o código-fonte do pipeline, incluindo a configuração dos hiperparâmetros das instâncias UMAP e HDBSCAN, foi disponibilizado publicamente[10]. O notebook de execução do pipeline encontra-se acessível via Google Colab[11], e os artefatos de dados resultantes para visualização estão hospedados via GitHub Gist, conforme referenciado nas notas de rodapé deste capítulo.

5.2 Projeto e Implementação da Solução

Esta seção detalha a implementação prática do artefato computacional, detalhando o fluxo de dados e as etapas de processamento do *pipeline*. Em conformidade com a fase do (Capítulo 4) da metodologia DSR, o objetivo da implementação é processar os dados textuais do Observatório para gerar uma representação interativa do conhecimento científico.

Para operacionalizar essa tarefa, a solução foi estruturada em quatro etapas principais, conforme ilustrado na Figura 12:

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.
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.

[9] Disponível em: https://gist.github.com/jeoaraujx/c9a610202e70139054c7e37eab937b93
[10] Disponível em: https://github.com/jeoaraujx/TCC
[11] Disponível em: https://colab.research.google.com/drive/1IdeIJ14TeLEZuJAB176UqALWFRwFXA2p

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.

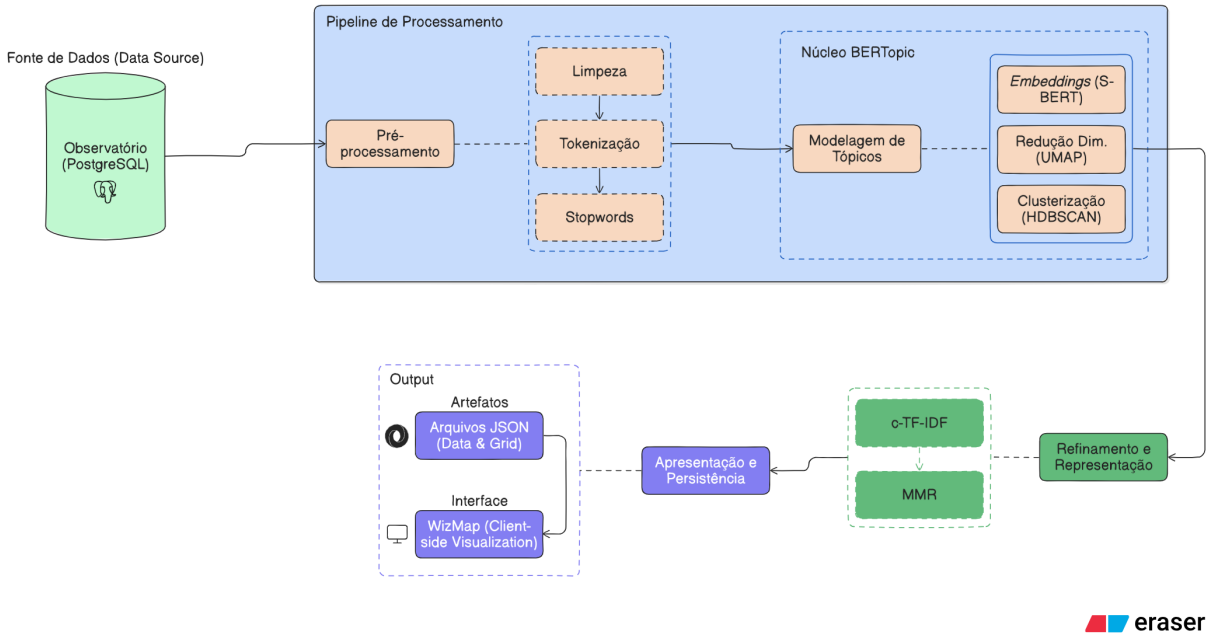

Figura 12 – Arquitetura do Pipeline de Modelagem e Visualização.

Fonte: O Autor

5.2.1 Coleta e Pré-processamento dos Dados

A camada de ingestão de dados foi projetada utilizando uma extração estática (dump) dos metadados do Núcleo de Pesquisa Aplicada e Inovação (NPAI). A opção por um dataset estático e isolado do ambiente de produção foi uma decisão metodológica para garantir a integridade e a reprodutibilidade dos experimentos de otimização de hiperparâmetros. O uso do banco completo e dinâmico nesta fase de desenvolvimento introduziria variáveis de instabilidade e custo computacional desnecessários para a validação da prova de conceito (PoC) da arquitetura proposta.

5.2.1.1 Caracterização do Corpus de Teste

Para a validação técnica do pipeline, utilizou-se um corpus de teste extraído do repositório do NPAI. Este conjunto de dados consiste em uma amostra de *343 registros*, selecionados a partir dos critérios de relevância bibliográfica: publicações em livros, capítulos de livros e artigos completos. A utilização deste volume reduzido foi intencional para viabilizar a execução iterativa dos testes de otimização de hiperparâmetros em ambiente de desenvolvimento, sem comprometer a estabilidade do sistema de produção.

A amostra compreende um intervalo temporal de *26 anos*, com publicações registradas entre *1998 e 2024*. Quanto à distribuição linguística, o corpus é predominantemente bilíngue, composto por *62,39%* de títulos em Português e *35,28%* em Inglês, contendo ainda uma fração de *2,33%* em Espanhol. Essa heterogeneidade linguística foi fundamental para testar a capacidade do modelo *paraphrase-multilingual-MiniLM-L12-v2* em agrupar semanticamente documentos de mesmo tema, independentemente do idioma de origem. Para fins de transparência e reprodutibilidade, o dataset completo encontra-se disponibilizado no repositório GitHub[12] referenciado na nota de rodapé deste capítulo.

5.2.1.2 Etapas de Pré-processamento

O módulo de pré-processamento foi estruturado em dois estágios lógicos: saneamento (limpeza) do *dataset* e normalização linguística.

No estágio de saneamento, implementado via biblioteca Pandas, aplicaram-se filtros para a exclusão de registros inconsistentes, como entradas nulas ou duplicatas exatas. Essa etapa foi crítica para assegurar que a frequência de termos (TF-IDF) não fosse distorcida por redundâncias na base de dados.

No estágio de normalização, cada documento foi submetido a um pipeline de tratamento textual multilíngue, composto pelas seguintes rotinas:

1. **Identificação do Idioma:** Classificação automática do texto via biblioteca *langdetect* para direcionar o processamento específico (Português/Inglês).
2. **Normalização e Limpeza de Caracteres:** Conversão para caixa baixa e remoção de acentuação (normalização Unicode NFKD) e caracteres especiais via expressões regulares, reduzindo a dimensionalidade do vocabulário.
3. **Lematização e Tokenização (spaCy):** Aplicação de lematização via *spaCy*, convertendo termos flexionados à sua forma canônica (lema) para agrupar variações da mesma palavra.
4. **Filtragem Final de Tokens:** Remoção de *stopwords* (termos funcionais sem carga semântica) e de *tokens* com comprimento inferior a três caracteres.

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, consistiu em uma lista de documentos processados, prontos para a etapa de geração de *embeddings*.

[12] Disponível em: https://github.com/jeoaraujx/TCC

5.2.2 Modelagem de Tópicos

A arquitetura do pipeline foi concebida de forma modular, seguindo o padrão de injeção de dependências permitido pela biblioteca BERTopic. Conforme ilustrado na 12, o "Núcleo BERTopic" não atua como um monólito rígido, mas sim como um orquestrador do fluxo de dados. As etapas de Redução de Dimensionalidade e Clusterização são executadas por instâncias customizadas dos algoritmos UMAP e HDBSCAN, respectivamente, que são configuradas externamente (com hiperparâmetros ajustados para o cenário de textos curtos) e injetadas no pipeline principal. Essa abordagem garante a reprodutibilidade e o controle fino sobre cada etapa do processamento, permitindo que o BERTopic gerencie a integração entre os embeddings gerados e os clusters resultantes.

5.2.2.1 Geração de Embeddings

A representação vetorial dos documentos foi realizada através do modelo pré-treinado *paraphrase-multilingual-MiniLM-L12-v2* da biblioteca *Sentence-Transformers*. A escolha deste modelo justifica-se pela sua arquitetura otimizada para aferir similaridade semântica em contextos multilíngues, convertendo cada título em um vetor denso de 384 dimensões.

5.2.2.2 Redução de Dimensionalidade (UMAP)

A configuração do UMAP priorizou a preservação da estrutura local dos dados em detrimento da estrutura global, visando a identificação de micro-tópicos. Os hiperparâmetros foram ajustados experimentalmente para projetar os vetores em um espaço bidimensional (*n_components=2*), requisito necessário para a compatibilidade com a camada de visualização WizMap. A métrica de distância cosseno foi adotada para capturar a orientação semântica dos vetores.

5.2.2.3 Clusterização (HDBSCAN) e Otimização de Hiperparâmetros

A definição dos agrupamentos foi executada pelo algoritmo HDBSCAN. Para determinar o valor ideal do hiperparâmetro *min_cluster_size*, que controla o tamanho mínimo para a formação de um tópico, foi conduzida uma análise de sensibilidade quantitativa.

O experimento consistiu no treinamento iterativo do modelo variando o parâmetro *min_cluster_size* em um intervalo de 4 a 25. Para cada iteração, foram coletadas duas métricas de desempenho:

1. **Coerência Semântica (C_v):** Utilizando a medida c_v (baseada em janelas deslizantes) para avaliar o grau de associação semântica entre as palavras principais de cada tópico.

2. **Taxa de Ruído (*Outliers*):** Percentual de documentos classificados como -1 (sem tópico atribuído) pelo algoritmo.

O critério de seleção adotado buscou o ponto de equilíbrio que maximizasse a coerência dos tópicos gerados, mantendo a taxa de descarte de documentos (perda de dados) abaixo de 10%.

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. Os resultados, incluindo os IDs dos tópicos para cada documento e suas probabilidades, foram então obtidos. O modelo treinado foi salvo persistentemente para facilitar sua reutilização.

5.2.3 Refinamento e Representação dos Tópicos (MMR)

Para avaliar a qualidade semântica e a interpretabilidade dos tópicos gerados, conduziu-se um experimento comparativo focado na etapa de representação de palavras-chave do pipeline. O objetivo foi contrastar a abordagem padrão baseada em frequência (c-TF-IDF) com a abordagem baseada em diversidade (Maximal Marginal Relevance - MMR).

Para garantir a validade interna da comparação, as etapas antecedentes à representação foram mantidas constantes em ambos os cenários de teste:

- **Embeddings:** Utilizou-se o modelo pré-treinado *paraphrase-multilingual-MiniLM-L12-v2*.
- **Redução e Clusterização:** O algoritmo UMAP foi configurado com 5 vizinhos e 2 componentes. O algoritmo HDBSCAN foi fixado com *min_cluster_size=8*, valor determinado previamente na análise de sensibilidade (Seção 6.1) por apresentar o maior índice de coerência ($C_v \approx 0.43$) e taxa de outliers reduzida ($< 9\%$).

Foram estabelecidos dois cenários de representação:

1. **Cenário de Controle (Padrão):** Utilização exclusiva do c-TF-IDF. Este método prioriza palavras frequentes dentro de um cluster específico, mas raras no restante do

corpus. A hipótese é que este método gere rótulos fiéis estatisticamente, porém com tendência à redundância morfológica.

2. **Cenário Experimental (MMR):** Aplicação do MMR sobre os resultados do c-TF-IDF, com fator de diversidade (λ) configurado em *0.3*. A hipótese é que o reranqueamento reduza a sinonímia e revele subtemas latentes ofuscados por termos genéricos de alta frequência.

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 camada final da arquitetura é responsável pela interface com o usuário. A implementação utilizou a biblioteca wizmap para processar os artefatos gerados pelo núcleo (coordenadas UMAP, metadados e rótulos refinados) e gerar os arquivos de visualização.

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. Conforme definido na metodologia, a validação do artefato visa verificar se a integração do BERTopic com a ferramenta WizMap permite a identificação coerente de temas e a exploração estruturada do conhecimento, utilizando como base de teste os dados extraídos do Observatório.

A análise dos resultados está estruturada em três dimensões:

- **Resultados da Avaliação de Desempenho (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 Semântica e Discussão (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 Resultados da Avaliação de Desempenho

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 Análise de Sensibilidade e Definição de Tópicos

A avaliação de desempenho dos hiperparâmetros, conforme metodologia descrita na Seção 5.2.2.3, resultou nos indicadores apresentados na Figura 13. O gráfico relaciona o Escore de Coerência (eixo esquerdo, linha azul) e o Número de Tópicos Gerados (eixo direito, linha vermelha) em função do tamanho do cluster.

A análise dos dados demonstra que o valor 8 para o *min_cluster_size* apresentou o desempenho superior no conjunto de testes. Observam-se os seguintes comportamentos:

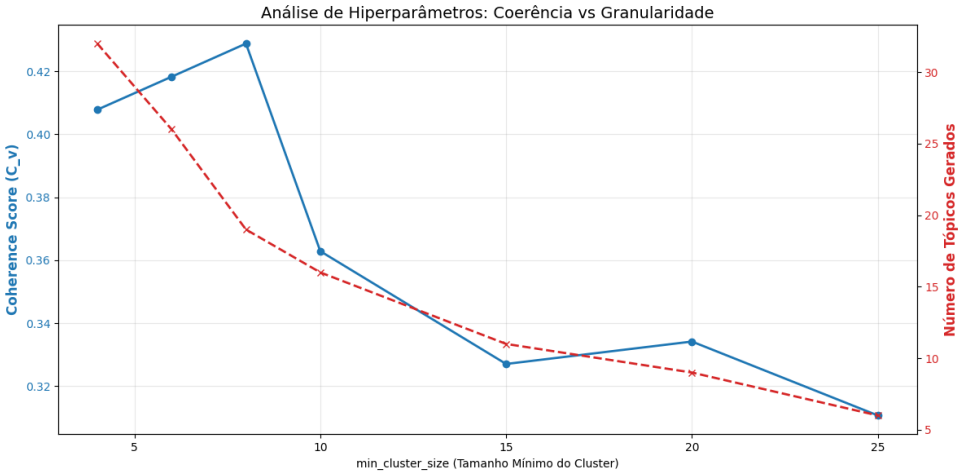

Figura 13 – Análise de Hiperparâmetros: Coerência (C_v) vs. Granularidade.

Fonte: O Autor

- **Maximização da Coerência:** A curva de coerência apresenta uma trajetória ascendente a partir do valor 4, atingindo seu pico máximo ($C_v \approx 0,429$) em 8. Nota-se uma degradação acentuada da métrica a partir do valor 10, indicando que a imposição de agrupamentos maiores forçou a fusão de temas semanticamente distintos, reduzindo a qualidade interna dos tópicos.

- **Controle de Ruído:** O aumento do parâmetro de 6 para 8 resultou em uma redução na taxa de *outliers* de 13,7% para 8,75%. Isso indica que o modelo conseguiu agregar documentos anteriormente dispersos em tópicos consistentes, diminuindo a perda de informação.

- **Granularidade:** A configuração selecionada resultou na identificação de 19 tópicos. Essa granularidade mostrou-se adequada para a análise, evitando a fragmentação excessiva observada nos valores menores (32 tópicos em tamanho 4) e a generalização excessiva dos valores maiores.

Com base nesses resultados, o modelo final foi instanciado com *min_cluster_size=8*, garantindo robustez semântica e cobertura de dados.

6.1.2 Diversidade de Tópicos

A métrica de diversidade 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 uma baixa sobreposição de vocabulário entre os tópicos

gerados. O alto índice sugere que a configuração do *BERTopic*, aliada ao refinamento por *MMR*, foi eficaz na segregação dos documentos em *clusters* temáticos linguisticamente distintos, minimizando a redundância.

6.1.3 Avaliação de Coerência e Trade-off Semântico

Para a avaliação de coerência dos tópicos, foram empregadas duas métricas com objetivos distintos: a Coerência C_v, focada na estabilidade estrutural, e o NPMI, focado na coocorrência léxica.

A métrica C_v foi estabelecida como o indicador principal durante a otimização de hiperparâmetros (conforme Seção 6.1.1). O modelo atingiu seu ponto ótimo ($C_v \approx 0.43$) com clusters de tamanho 8, evidenciando que os agrupamentos formados possuem forte consistência interna.

Posteriormente, para avaliar o impacto da estratégia de diversificação de rótulos, comparou-se o desempenho do modelo Padrão (*Standard c-TF-IDF*) contra o modelo com refinamento MMR utilizando a métrica NPMI. A Figura 14 ilustra os resultados obtidos.

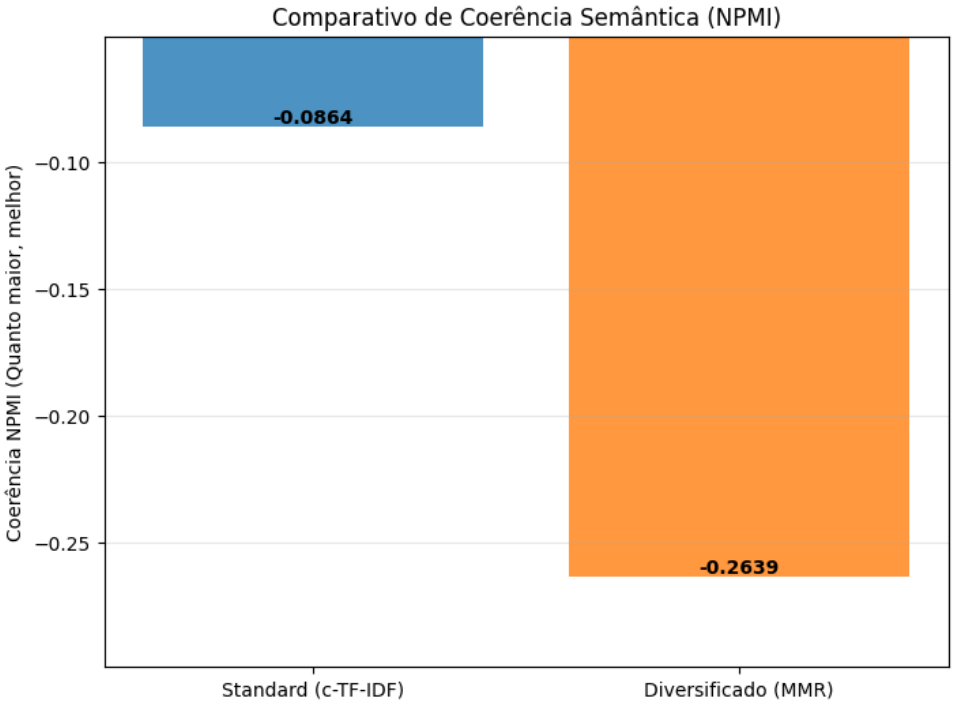

Figura 14 – Comparativo de Coerência Semântica (NPMI) entre os modelos.

Fonte: O Autor

A análise dos dados revela que o Modelo Padrão obteve um índice NPMI superior (-0.0864) em comparação ao Modelo com MMR (-0.2639). Embora métricas de coerência mais altas sejam usualmente preferíveis, neste cenário específico, a divergência corrobora a eficácia do algoritmo MMR em sua função de diversificação semântica.

O Modelo Padrão tende a maximizar o NPMI ao selecionar termos de alta coocorrência imediata, frequentemente resultando em pares redundantes ou sinônimos (ex: "doença" e "vírus"). Em contrapartida, o algoritmo MMR penaliza intencionalmente essa redundância para introduzir termos distintos (ex: "doença" e "transporte"). Como a probabilidade estatística de um termo diverso aparecer no mesmo contexto curto é menor do que a de um sinônimo, a redução no índice NPMI é um reflexo natural da maior variabilidade vocabular.

Portanto, a variação no NPMI representa um *trade-off* metodológico: o modelo prioriza a informatividade e a cobertura temática em detrimento da redundância estatística.

Adicionalmente, deve-se observar que a magnitude negativa dos valores em ambos os cenários é influenciada por fatores intrínsecos ao corpus utilizado:

1. **Natureza dos Dados (Textos Curtos):** O modelo foi treinado exclusivamente com **títulos** de publicações. A brevidade dos títulos reduz a probabilidade de coocorrência de termos correlatos, penalizando métricas baseadas em contagem estrita como o NPMI, que foi originalmente projetada para documentos longos.
2. **Volume de Dados:** A validação ocorreu em um subconjunto de dados de teste, cujo volume reduzido pode limitar a estabilização de métricas estatísticas globais de coocorrência.

Diante dessas limitações de desempenho intrínsecas, a validação semântica e discussão, apresentada a seguir, torna-se indispensável para aferir a utilidade real e a interpretabilidade dos tópicos gerados.

6.2 Validação Semântica e Discussão

A validação semântica do artefato é crucial para complementar as métricas de desempenho (Seção 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 Experimento Comparativo de Representação (c-TF-IDF vs. MMR)

A avaliação da qualidade dos rótulos baseou-se na inspeção semântica de Nuvens de Palavras (*WordClouds*) pareadas, observando-se critérios de especificidade semântica e redução de redundância.

Inicialmente, a análise do Tópico 9 (Arboviroses e Saúde Pública), apresentada na Figura 15, ilustra o impacto do refinamento na contextualização do tema. No modelo padrão (à esquerda), a representação é dominada por termos de alta ocorrência como "dengue", "case" (caso) e "fever" (febre), centrando a descrição apenas na doença principal e seus sintomas.

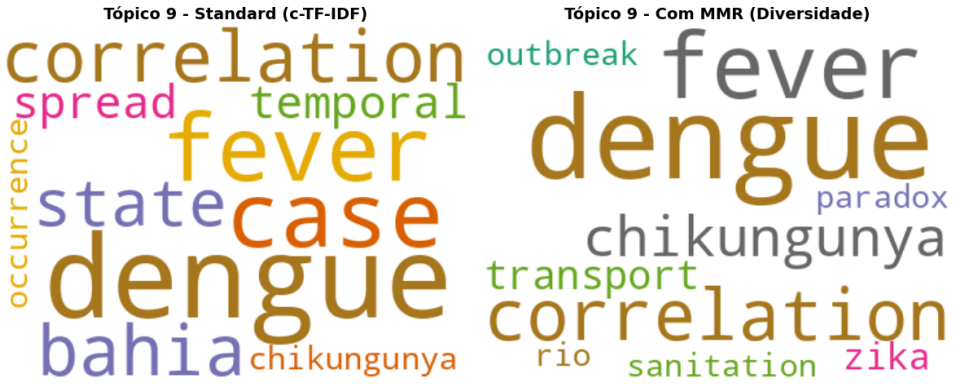

Figura 15 – Nuvens de Palavras para Arboviroses e Saúde Pública

Fonte: O Autor

Ao aplicar o MMR (à direita), observa-se que o modelo penalizou a redundância de termos genéricos, permitindo a emergência de vetores correlatos como "chikungunya" e "zika", além de fatores ambientais e sociais como "outbreak" (surto), "transport" e "sanitation" (saneamento). O uso da diversidade transformou uma descrição genérica em uma representação que abarca o ecossistema das arboviroses.

De forma análoga, a análise do Tópico 3 (Figura 16) demonstra a capacidade do MMR em revelar nichos específicos que, no modelo padrão, apareciam diluídos em um tema amplo de tecnologia educacional (com termos como "aprendizagem" e "ensino").

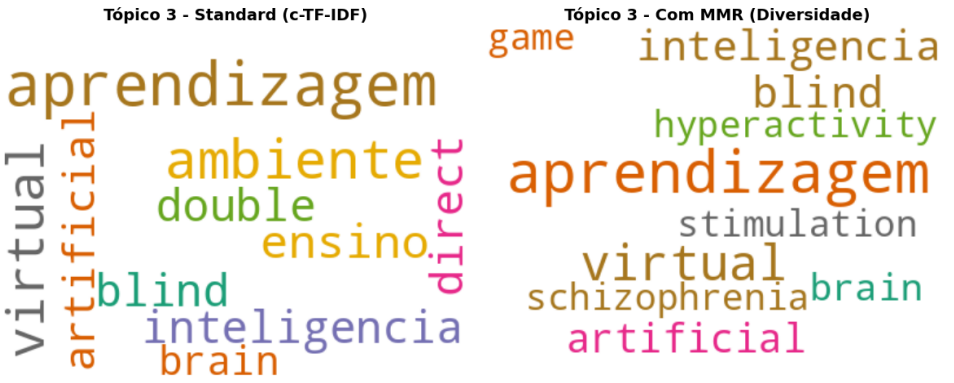

Figura 16 – Nuvens de Palavras para tecnologia educacional

Fonte: O Autor

O refinamento evidenciou que o foco real deste agrupamento recai sobre aplicações para especificidades cognitivas. Termos como "schizophrenia" (esquizofrenia), "hyperacti-

vity" (hiperatividade) e "brain" (cérebro) ganharam relevância sobre os termos genéricos, permitindo identificar a intersecção entre inteligência artificial e neurodiversidade que não estava clara na representação baseada apenas em frequência.

Por fim, o Tópico 0 (Figura 17) ilustra a distinção entre conceitos macroeconômicos e componentes físicos.

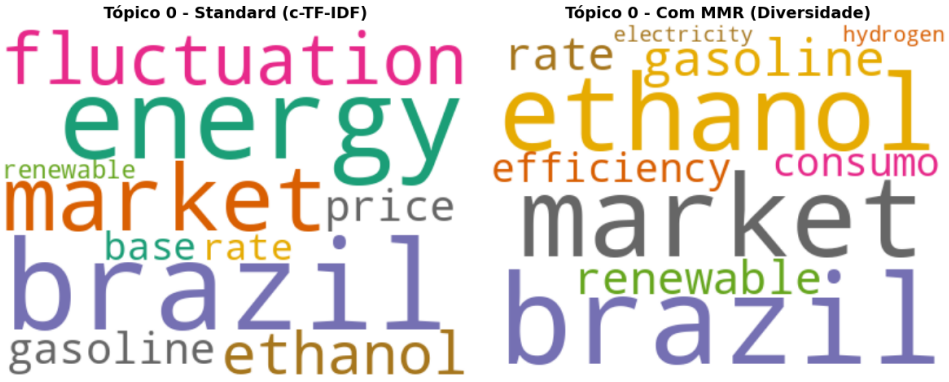

Figura 17 – Nuvens de Palavras para conceitos macroeconômicos e componentes físicos

Fonte: O Autor

Enquanto a abordagem c-TF-IDF priorizou termos abstratos como "market", "price" e "fluctuation", o MMR diversificou a representação para incluir os componentes específicos da matriz energética analisada: "ethanol", "gasoline" e "hydrogen". A representação diversificada oferece, portanto, uma descrição mais concreta e imediata do objeto de estudo das publicações agrupadas.

6.2.2 Análise dos Tópicos Identificados

Para superar a limitação de uma análise puramente descritiva, foi conduzida uma auditoria qualitativa nos tópicos identificados pelo pipeline. Os resultados principais, contendo os tópicos mais populosos e suas palavras-chave representativas, estão sumarizados na Tabela 2. Adicionalmente, a Figura 14 apresenta as nuvens de palavras dos agrupamentos selecionados, permitindo a inspeção visual da coesão semântica.

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

ID Tópico	Palavras-Chave	Nº Documentos
0	brazil, market, ethanol, gasoline, rate	31
1	inovacao, empreendedorismo, pesquisa, aplicacoes, intelectual	30
2	soil, fractal, oil, microrrelieve, detection	27
3	aprendizagem, virtual, blind, inteligencia, artificial	26
4	conhecimento, academico, validacao, gerencia, decision	26
5	computador, sistema, computacao, conhecimento, organizacoes	24
6	network, synchronization, semantic, sincronizacao, protein	17
7	robotica, escola, makers, facial, protetor	16
8	pain, chronic, fibromyalgia, postural, vibration	13
9	dengue, fever, correlation, chikungunya, transport	12

Fonte: O Autor

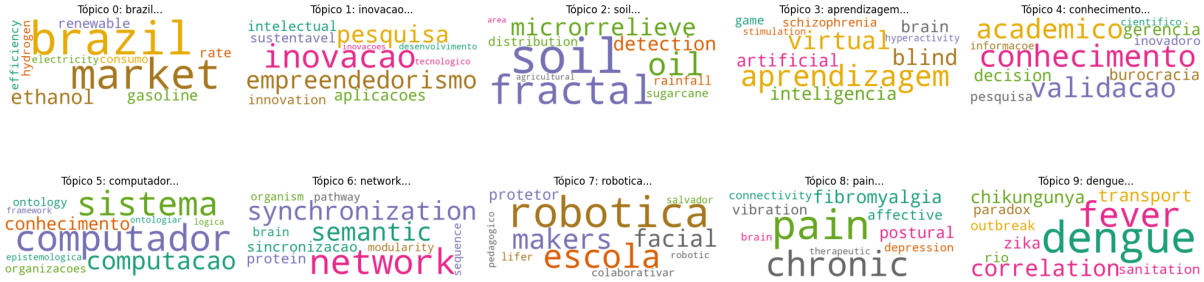

Figura 18 – Nuvens de Palavras para Tópicos Selecionados.

Fonte: O Autor

Como critério de validação externa, os tópicos gerados foram confrontados com as Grandes Áreas do Conhecimento do CNPq. A análise revelou que os agrupamentos semânticos refletem espontaneamente as divisões disciplinares formais, conforme a distribuição observada na amostra auditada:

- **Ciências da Saúde:** Representada por tópicos de alta coesão como o Tópico 8 (Dor/Fibromialgia), Tópico 9 (Arboviroses/Dengue) e Tópico 16 (Pandemia/COVID-19), que alinham-se diretamente com as subáreas da Medicina e Saúde Coletiva.
- **Ciências Exatas e da Terra:** Identificou-se forte presença de temas de Física e Engenharias, exemplificados pelo Tópico 0 (Energia/Combustíveis), Tópico 12 (Engenharia de Software) e Tópico 13 (Física/Fractais e Wavelets).

- **Ciências Humanas e Sociais:** A modelagem capturou nuances de áreas como Educação (Tópico 3 - Aprendizagem/IA), Sociologia Política (Tópico 18 - Fake News/Participação) e Turismo (Tópico 17 - Mobilidade Urbana e Turismo).

Este alinhamento entre os clusters matemáticos e a taxonomia oficial sugere que o modelo foi capaz de capturar a estrutura disciplinar subjacente ao acervo sem supervisão prévia.

Para aferir a consistência global, os tópicos foram classificados quanto à sua qualidade interpretativa. Na amostra dos tópicos mais representativos, estima-se que acima de 85% apresentaram Alta Especificidade, definindo um tema central claro e sem ambiguidade.

No entanto, é necessário reportar a existência de agrupamentos de menor qualidade, que constituem uma limitação do ajuste não supervisionado:

1. **Tópicos Genéricos/Mistos:** O Tópico 10, por exemplo, apresentou uma mistura de termos metodológicos ("bibliografico", "colaborativo") com temas desconexos ("cancer", "dengue"). Isso indica um agrupamento baseado mais na estrutura de escrita científica (termos comuns em resumos) do que em um domínio de conhecimento específico.
2. **Ruído (Outliers):** O grupo de ruído (Tópico -1), gerado nativamente pelo algoritmo HDBSCAN, conteve documentos que misturavam termos sem relação semântica aparente, como "espiritualidade", "chagas" e "judicial". A segregação eficaz desses documentos neste cluster de descarte valida a capacidade do modelo de limpar os tópicos principais, evitando que documentos desconexos contaminem os clusters de alta qualidade.

Essa auditoria confirma que, embora o modelo tenha sucesso em identificar os grandes domínios científicos, a presença residual de tópicos genéricos (como o Tópico 10) reforça a necessidade de curadoria humana ou refinamento de stopwords para a aplicação bibliométrica definitiva.

6.2.3 Discussão do Mapa de Conhecimento Interativo

O artefato final, visualizado através da ferramenta WizMap na Figura 19, atua como uma interface de navegação sobre o espaço semântico calculado. Embora não constitua uma ontologia formal, a projeção atua funcionalmente como um mapa de conhecimento ao permitir que o pesquisador identifique visualmente as fronteiras e intersecções entre os domínios científicos. As áreas de maior densidade (*clusters*) representam a consolidação de tópicos, enquanto a distância euclidiana entre os pontos serve como proxy para a dissimilariedade temática.

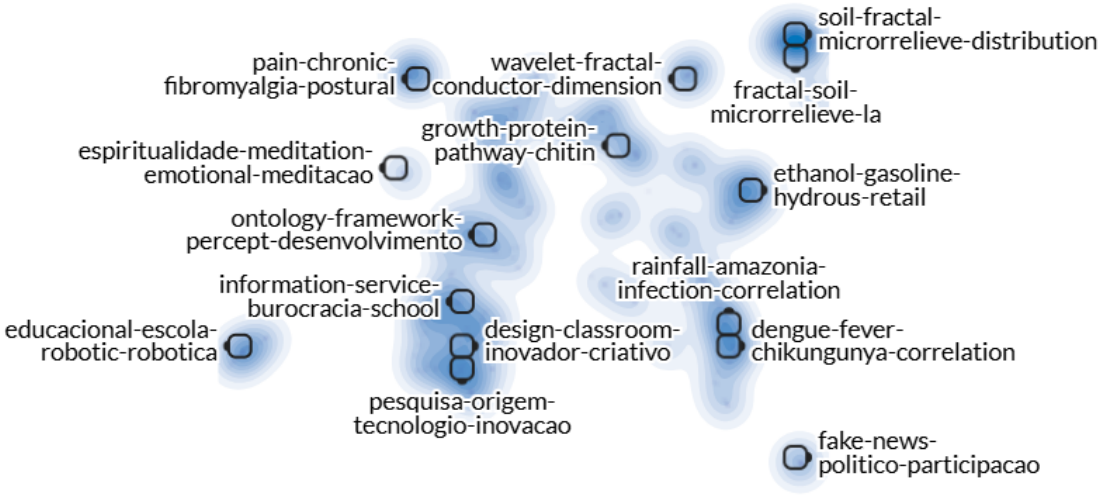

Figura 19 – Visão Geral do Mapa de Conhecimento Interativo (WizMap).

Fonte: O Autor

Na Figura 19, 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.

6.2.4 Comparativo Qualitativo: Busca Lexical e Exploração Semântica

Para estabelecer uma linha de base de comparação, foi realizado um procedimento de recuperação de informação no sistema atual do Observatório utilizando o descritor "Dengue". A Figura 20 apresenta o resultado retornado pela plataforma.

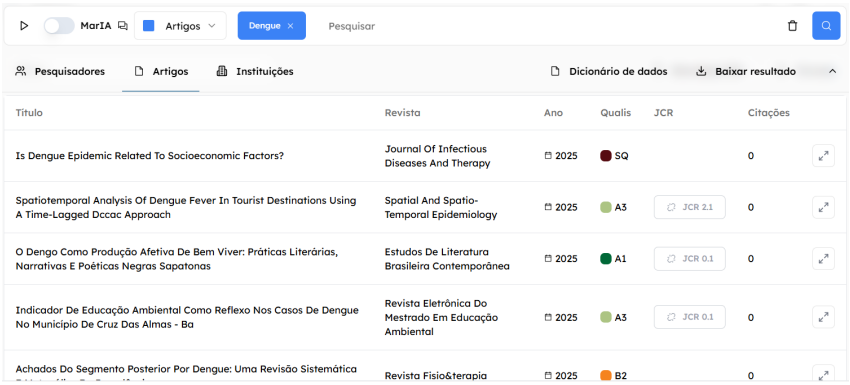

Figura 20 – Resultado da busca lexical pelo termo "Dengue" no Observatório.

Fonte: O Autor

Conforme observado na Figura 20, o mecanismo de busca opera sob um paradigma estritamente lexical. O algoritmo filtra e exibe apenas os registros em que a cadeia de caracteres "Dengue" (ou variações morfológicas próximas) consta explicitamente no título da publicação. Nota-se que a dependência da correspondência exata do termo pode, inclusive, gerar ruído na recuperação, trazendo resultados semanticamente distintos do objetivo da pesquisa, como observado no terceiro item da lista, recuperado devido à similaridade gráfica com o termo de busca.

Essa restrição terminológica resulta em um problema de recuperação conhecido como "silêncio" ou falsos negativos. Publicações relevantes que abordam o domínio das arboviroses, controle vetorial ou vetores (ex: "Aedes aegypti", "Zika", "Chikungunya") mas que não contêm o token específico "Dengue" no título, são excluídas da lista de resultados.

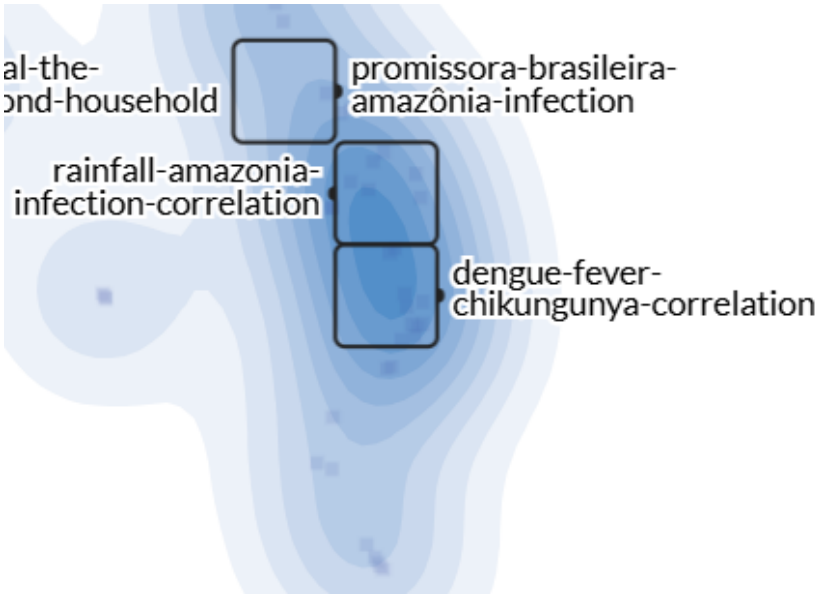

Figura 21 – Cluster de correlação de termos relacionados a arboviroses no mapa de conhecimento interativo.

Fonte: O Autor

Em contraste, ao analisar o Tópico 4 gerado pelo pipeline proposto (conforme Tabela 2) e a figura 21, verifica-se o agrupamento de termos como “febre” e “chikungunya” no mesmo cluster de “dengue”. Isso indica que o modelo vetorial foi capaz de conectar documentos pelo contexto semântico compartilhado e mitigar o silêncio na recuperação identificado no sistema atual.

No que tange à validação do objetivo central de “melhorar a capacidade de exploração”, os resultados da inspeção visual (demonstrados na Seção 6.2.2) demonstram que o artefato é eficaz em agrupar documentos por afinidade semântica, uma funcionalidade inexistente na busca lexical atual (demonstrada na Seção 6.2.4). A validação por especialista confirma que os agrupamentos (clusters) correspondem a domínios reais de conhecimento. Portanto, considera-se que o artefato resolve o problema da ausência de ferramentas de navegação semântica, validando sua utilidade estrutural como um complemento à busca tradicional, embora a mensuração de ganhos de eficiência (tempo de busca por usuário final) permaneça como uma proposta para trabalhos futuros.

6.3 Limitações do Estudo

A interpretação dos resultados deve considerar as limitações inerentes ao escopo experimental e à natureza dos dados utilizados.

- **Conjunto de Dados (Prototipagem vs. Produção):** Uma característica deste ciclo de pesquisa foi a utilização de um subconjunto de dados estático (NPAI) em detrimento da conexão direta com o acervo completo do Observatório. Embora

isso impeça a generalização estatística dos tópicos encontrados para toda a ciência baiana (validade externa bibliométrica), essa restrição foi necessária para viabilizar a validação técnica do artefato (validade interna). O uso de um ambiente controlado permitiu o ajuste fino dos algoritmos sem as latências e restrições de segurança inerentes ao acesso a bancos de dados de produção governamentais. Portanto, os resultados validam a tecnologia desenvolvida, cumprindo o objetivo de engenharia proposto, deixando a aplicação em larga escala como uma etapa subsequente de implantação.

- **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` (UMAP) e `min_cluster_size` (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 6.1.3), 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 6.1.3, 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 de desempenho ideal para modelos baseados em similaridade semântica (como o BERTopic). A Validação Semântica e Discussão (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 experimento é demonstrar a *viabilidade* da metodologia, e não apresentar uma solução de *software* em produção.

- **Rigidez na Hierarquia de Visualização (WizMap):** Embora a ferramenta WizMap seja eficiente para a renderização escalável de embeddings, observou-se uma limitação na integração de seus mecanismos de resumo automático com os

tópicos pré-calculados pelo BERTopic. O WizMap utiliza algoritmos internos baseados em frequência de termos para gerar os rótulos dinâmicos nos níveis de zoom mais afastados (visão macro). Esse comportamento pode sobrepor a classificação semântica refinada do BERTopic, resultando, por vezes, na exibição de termos de alta frequência mas baixa relevância semântica (como nomes de autores ou termos institucionais comuns no corpus) em detrimento dos descritores temáticos. Essa característica exigiria uma etapa adicional de limpeza agressiva dos metadados de entrada especificamente para a visualização, ou o desenvolvimento de uma camada de visualização customizada que respeitasse estritamente os rótulos do c-TF-IDF em todas as escalas de resolução.

7 CONCLUSÃO

Este trabalho abordou o desafio de exploração de grandes acervos de publicações científicas, onde as abordagens tradicionais de busca lexical (palavras-chave) limitam a descoberta de conhecimento diante da sobrecarga de informação. O objetivo central consistiu no projeto, desenvolvimento e avaliação de um artefato computacional capaz de processar dados estruturados do Núcleo de Pesquisa Aplicada e Inovação (NPAI), servindo como prova de conceito para o Observatório de dados públicos da Bahia, apresentando-os na forma de um mapa de conhecimento interativo.

Para atingir esse objetivo, foi instanciado um *pipeline* modular fundamentado em técnicas de PLN. Embora a orquestração de componentes como BERTopic e WizMap possua precedentes na literatura, a contribuição deste estudo reside na adaptação metodológica necessária para operar em um *corpus* de alta restrição (apenas títulos) e no idioma português. A avaliação demonstrou que a aplicação direta dessas ferramentas resultaria em ruído excessivo, sendo necessária a definição de uma heurística de configuração (ajuste de hiperparâmetros como *min_cluster_size* e MMR) capaz de extrair inteligência de dados esparsos.

A adoção da metodologia DSR guiou a estrutura iterativa do trabalho, permitindo transitar entre a fundamentação teórica e o desenvolvimento prático. Isso garantiu que o artefato não fosse apenas uma implementação de software, mas uma proposta arquitetural validada tecnicamente para a lacuna de exploração semântica regional.

A avaliação do artefato (Capítulo 6) demonstrou a viabilidade técnica da solução. A avaliação de desempenho indicou uma alta diversidade de tópicos (**0.9214**), sugerindo baixa sobreposição lexical. O escore de coerência NPMI (**-0.2639**) apresentou-se negativo, resultado compreendido como um *trade-off* entre a coocorrência estatística (penalizada por títulos curtos) e a diversidade semântica buscada pelo algoritmo. A validação qualitativa, por meio de auditoria taxonômica e inspeção visual, corroborou a eficácia da solução, evidenciando que os tópicos gerados possuem coerência interpretativa e alinhamento com as grandes áreas do conhecimento.

Quanto às limitações, destaca-se que o uso exclusivo de títulos restringiu a janela de contexto para o modelo. Além disso, o artefato foi executado como um protótipo em ambiente controlado (NPAI). Portanto, os resultados atuais validam a tecnologia, mas não refletem toda a complexidade de um ambiente de produção em tempo real.

Conclui-se, portanto, que o estudo atingiu seu objetivo ao demonstrar a *viabilidade técnica* da modelagem de tópicos baseada em embeddings como uma camada complementar à arquitetura de dados científicos. Os resultados obtidos na Prova de Conceito indicam que

o pipeline é capaz de revelar agrupamentos semânticos latentes que não são capturados pela busca exata. Desta forma, o artefato apresenta-se como uma arquitetura funcional validada, instrumentalizando gestores e pesquisadores com uma ferramenta de apoio à decisão baseada em evidências, deixando a mensuração de eficiência operacional para trabalhos futuros.

7.1 Trabalhos Futuros

As limitações identificadas durante o desenvolvimento e a natureza de protótipo deste estudo abrem caminho para diversas frentes de pesquisa que não foram abordadas nesta iteração:

- **Enriquecimento Semântico (Full-Text):** Uma limitação central foi o uso exclusivo de títulos. Trabalhos futuros devem integrar o *pipeline* aos resumos (*abstracts*) das publicações. A hipótese é que textos mais longos forneçam contexto adicional, impactando positivamente as métricas estatísticas de coocorrência (NPMI) e refinando a densidade dos clusters.
- **Análise Temporal:** O estudo atual apresenta um "retrato estático" do acervo. Uma extensão promissora seria aplicar a variante dinâmica do BERTopic para visualizar a evolução dos temas ao longo do tempo (ex: o surgimento e declínio de termos como "COVID-19"), permitindo aos gestores identificar tendências emergentes ou obsoletas.
- **Validação com Usuários Reais:** Embora validado tecnicamente, o artefato carece de testes de usabilidade com os gestores do Observatório. Pesquisas futuras devem mensurar métricas de eficiência (tempo de descoberta) e satisfação do usuário para quantificar o ganho real em comparação ao sistema de busca tradicional.
- **Implementação de Busca Híbrida:** O desenvolvimento de um sistema que combine a precisão da busca lexical (palavras-chave) com a exploração semântica (mapa visual). Isso permitiria que o usuário filtrasse o mapa por um termo específico, unindo o melhor dos dois paradigmas de recuperação da informação.
- **Validação em Escala de Produção:** Aplicação do *pipeline* sobre a totalidade do acervo do Observatório. Esta etapa exigirá o reajuste (*tuning*) dos hiperparâmetros de modelagem para adequá-los ao volume massivo de dados, garantindo a escalabilidade da solução proposta.

REFERÊNCIAS

ANGELOV, D. **Top2Vec: Distributed Representations of Topics**. 2020. Disponível em: https://arxiv.org/abs/2008.09470. Citado nas páginas 26 e 30.

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 25.

BÖRNER, K. **Atlas of Science: Visualizing What We Know**. Cambridge, Massachusetts: The MIT Press, 2010. ISBN 9780262014458. Citado na página 33.

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 29.

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 30.

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 14, 25 e 26.

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 24.

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 14, 22 e 23.

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 39 e 42.

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 14, 18, 19, 35, 36, 37 e 38.

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 25.

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 35, 36 e 37.

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 14, 26, 27, 30, 36 e 42.

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 25.

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 25.

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 30.

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 44 e 45.

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 27, 35, 36, 37, 38, 41 e 42.

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 18.

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 17, 35, 36 e 37.

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 31.

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 28.

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 19.

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 27 e 28.

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 36 e 37.

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 25.

MIKOLOV, T. *et al.* **Efficient Estimation of Word Representations in Vector Space**. 2013. Disponível em: https://arxiv.org/abs/1301.3781. Citado nas páginas 19 e 20.

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 14.

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 19.

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 25.

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 22.

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019. Disponível em: https://arxiv.org/abs/1908.10084. Citado nas páginas 23, 24 e 27.

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 44 e 45.

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 26.

TAYLOR, W. L. Cloze procedure: A new tool for measuring readability. **Journalism Quarterly**, v. 30, p. 415–433, 1953. Citado na página 22.

VASWANI, A. *et al.* **Attention Is All You Need**. 2017. Disponível em: https://arxiv.org/abs/1706.03762. Citado nas páginas 14 e 21.

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 15, 31, 32 e 33.

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 36.

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 14, 17, 18, 20 e 25.