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

JEOSTON ARAUJO DA CRUZ JÚNIOR

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

SALVADOR
2024

JEOSTON ARAUJO DA CRUZ JÚNIOR

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

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

Orientador: Prof. Dr. Eduardo Manuel de Freitas Jorge

SALVADOR
2024

TERMO DE ANUÊNCIA DO ORIENTADOR

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

Prof. Dr. Eduardo Manuel de Freitas Jorge
Orientador

JEOSTON ARAUJO DA CRUZ JÚNIOR

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

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

Aprovada em: .

BANCA EXAMINADORA

Prof. Dr. Eduardo Manuel de Freitas Jorge
Orientador

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

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

AGRADECIMENTOS

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

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

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

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

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

RESUMO

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

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

ABSTRACT

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

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

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

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

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

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

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

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

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

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

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

2 REFERENCIAL TEÓRICO

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

2.1 Ciência da Informação e Análise de Publicações Científicas

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

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

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

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

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

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

2.2 Transformadores e Embeddings no Contexto do PLN

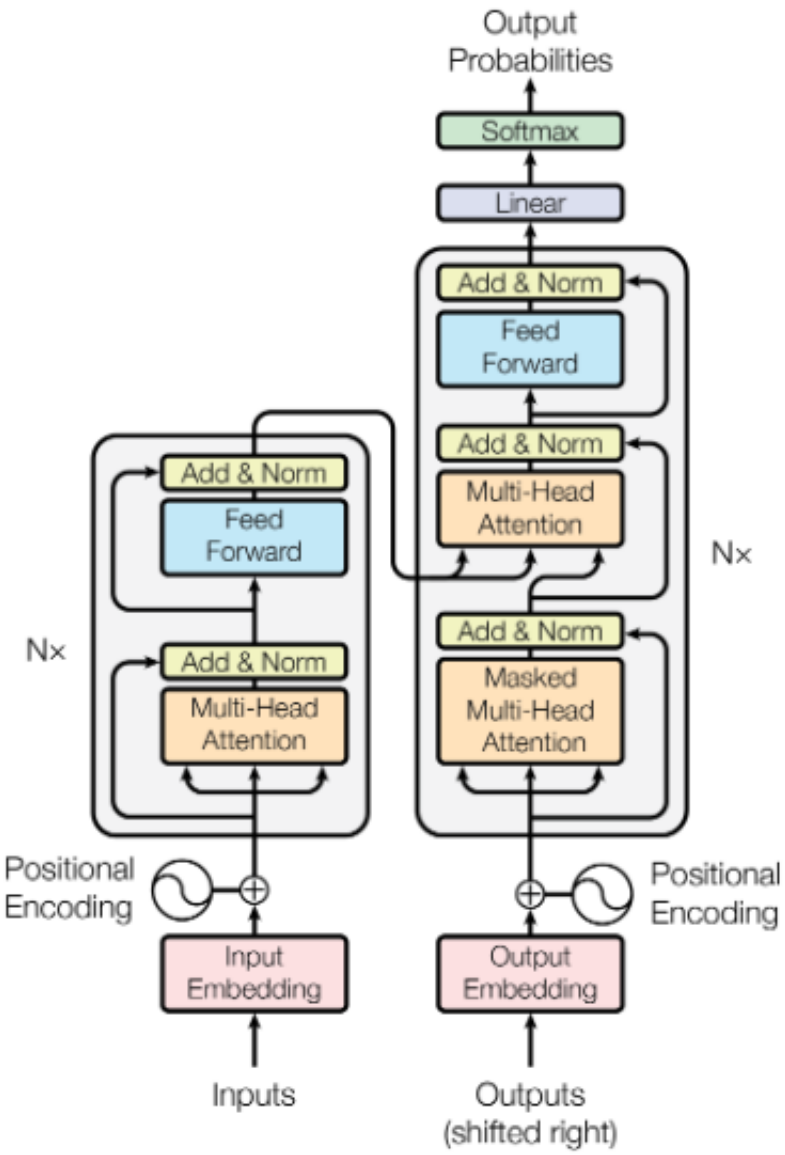
O avanço no campo do PLN tem sido marcado pela busca por representações vetoriais que capturem não apenas informações sintáticas, mas também aspectos semânticos e contextuais dos textos. As primeiras abordagens, como o *Word2Vec* de Mikolov *et al.* (2013) e o *GloVe* de Pennington, Socher e Manning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

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

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

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

Figura 1 – Transformador - modelo arquitetural.



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

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

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

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

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

2.3 Abordagens Tradicionais de Modelagem de Tópicos

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

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

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

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

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

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

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

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

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

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

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

2.4 BERTopic: Uma Abordagem Moderna

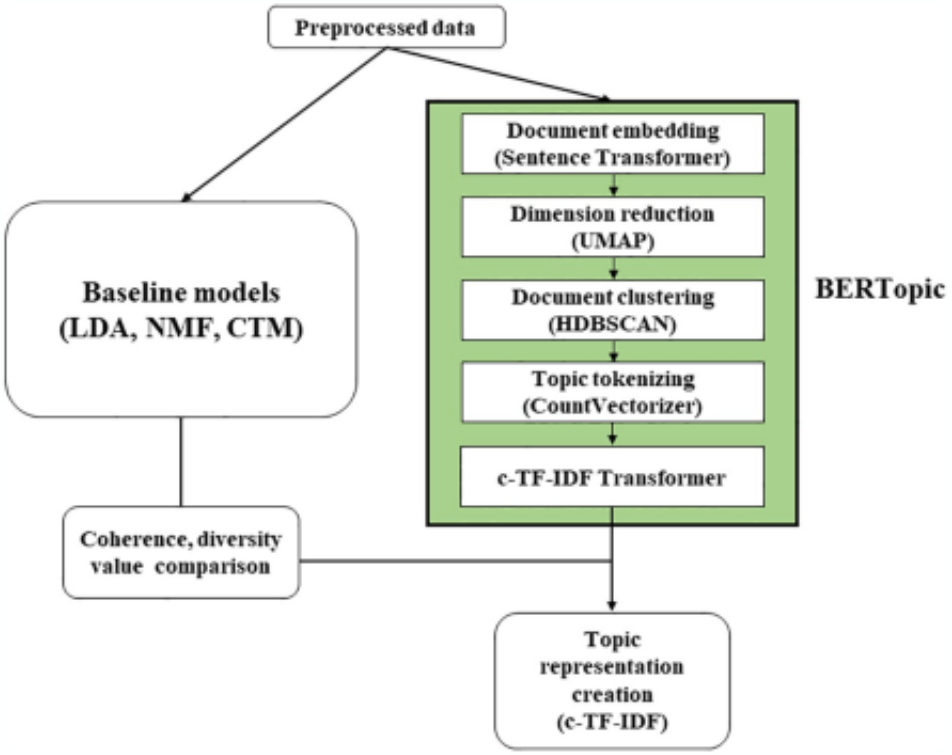
O *Bidirectional Encoder Representations from Transformers* BERT, introduzido por Devlin *et al.* (2019), marcou um avanço significativo no campo do *Natural Language Processing* NLP. Baseado na arquitetura de *Transformers* (Vaswani *et al.* (2017)), o BERT emprega o mecanismo de *self-attention* para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

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

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

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

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



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

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

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

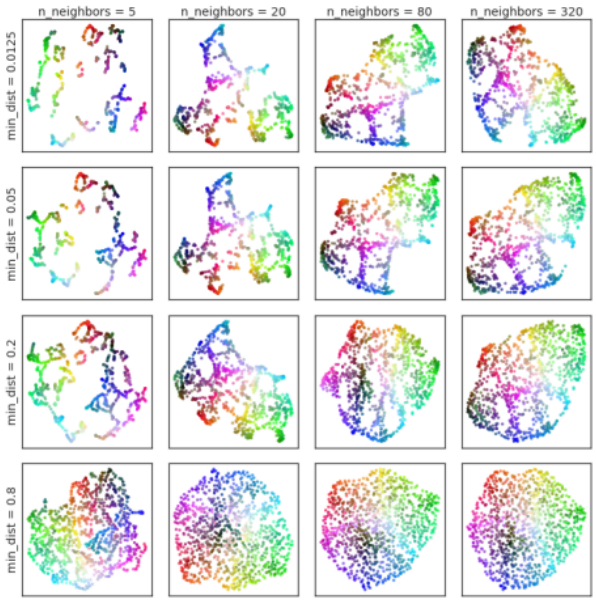


Figura 3 – Diagrama ilustrativo do UMAP, que demonstra a relação entre os hiperparâmetros *n_neighbors* e *min_dist* e a representação visual dos dados. O parâmetro *n_neighbors* controla a balança entre a preservação da estrutura global (valores altos) e local (valores baixos), enquanto *min_dist* ajusta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

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

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

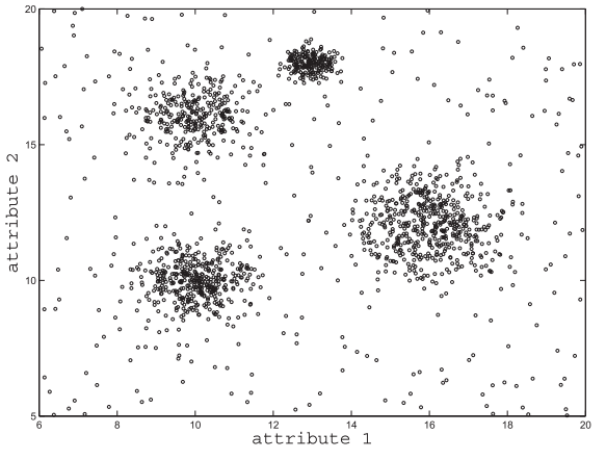


Figura 4 – Figura ilustrativa de um *dataset* sintético com quatro *clusters* e ruído de fundo. A imagem demonstra o tipo de desafio que o algoritmo HDBSCAN é capaz de superar, como a identificação de agrupamentos de densidades e formas variadas, além de tratar *outliers* de forma eficiente. Este comportamento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

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

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

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

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

2.5 Modelos de Linguagem de Grande Escala (LLMs)

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

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

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

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

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

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

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

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

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

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

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

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

REFERÊNCIAS

ACHIAM, J. *et al.* Gpts are gpts: An early look at the labor market impact potential of large language models. **arXiv preprint**, 2023. Disponível em: https://arxiv.org/abs/2304.02142. Citado na página 24.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, MIT Press, v. 3, n. Jan, p. 993–1022, 2003. ISSN 1532-4435. Disponível em: http://jmlr.org/papers/v3/blei03a.html. Citado na página 18.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J. *et al.* (Ed.). **Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 160–172. Citado nas páginas 21 e 22.

DATCHANAMOORTHY, K.; S, A. M. G.; B, P. Text mining: Clustering using bert and probabilistic topic modeling. **Social Informatics Journal**, 2023. Disponível em: https://api.semanticscholar.org/CorpusID:267122800. Citado nas páginas 12, 18, 19, 20 e 23.

DEERWESTER, S. *et al.* Indexing by latent semantic analysis. **Journal of the American Society for Information Science**, v. 41, n. 6, p. 391–407, 1990. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9. Citado na página 18.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: https://arxiv.org/abs/1810.04805. Citado nas páginas 12, 16, 19 e 23.

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

GALLI, C. *et al.* Topic modeling for faster literature screening using transformer-based embeddings. **Metrics**, v. 1, n. 1, 2024. ISSN 3042-5042. Disponível em: https://www.mdpi.com/3042-5042/1/1/2. Citado nas páginas 12, 14, 15 e 24.

GANA, B. *et al.* Leveraging llms for efficient topic reviews. **Applied Sciences**, v. 14, n. 17, 2024. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/14/17/7675. Citado nas páginas 22 e 23.

GEORGE, L.; SUMATHY, P. An integrated clustering and bert framework for improved topic modeling. **International Journal of Information Technology**, v. 15, n. 4, p. 2187–2195, 2023. Disponível em: https://doi.org/10.1007/s41870-023-01268-w. Citado nas páginas 18 e 20.

GROOTENDORST, M. **BERTopic: Neural topic modeling with a class-based TF-IDF procedure**. 2022. Disponível em: https://arxiv.org/abs/2203.05794. Citado nas páginas 12, 13, 17, 19, 22 e 24.

HOFMANN, T. Probabilistic latent semantic indexing. In: **Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. Berkeley, CA, USA: ACM Press, 1999. p. 50–57. ISBN 1-58113-096-1. Citado na página 18.

HOFMANN, T. Hofmann, t.: Unsupervised learning by probabilistic latent semantic analysis. machine learning 42(1-2), 177-196. **Machine Learning**, v. 42, p. 177–196, 01 2001. Citado na página 18.

JUNG, H. S. *et al.* Expansive data, extensive model: Investigating discussion topics around llm through unsupervised machine learning in academic papers and news. **PLOS ONE**, Public Library of Science, v. 19, n. 5, p. 1–18, 05 2024. Disponível em: https://doi.org/10.1371/journal.pone.0304680. Citado nas páginas 17, 20 e 23.

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 na página 14.

LIU, J.; ZHENG, C. Summary of the gpt-4 technical report. **arXiv preprint**, 2023. Disponível em: https://arxiv.org/abs/2303.08774. Citado na página 24.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008. Disponível em: http://jmlr.org/papers/v9/vandermaaten08a.html. Citado na página 20.

MCINNES, L.; HEALY, J.; MELVILLE, J. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2018. Disponível em: https://arxiv.org/abs/1802.03426. Citado nas páginas 20 e 21.

MENG, F. *et al.* Demand-side energy management reimagined: A comprehensive literature analysis leveraging large language models. **Energy**, v. 291, p. 130303, 2024. ISSN 0360-5442. Disponível em: https://www.sciencedirect.com/science/article/pii/S0360544224000744. Citado nas páginas 23 e 24.

MIFRAH, S.; BENLAHMAR, E. H. Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. **International Journal of Advanced Trends in Computer Science and Engineering**, 08 2020. Citado na página 18.

MIKOLOV, T. *et al.* **Efficient Estimation of Word Representations in Vector Space**. 2013. Disponível em: https://arxiv.org/abs/1301.3781. Citado na página 15.

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

OPENAI *et al.* **GPT-4 Technical Report**. 2024. Disponível em: https://arxiv.org/abs/2303.08774. Citado na página 24.

OUYANG, L. *et al.* Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, v. 35, p. 27730–27744, 2022. Disponível em: https://arxiv.org/abs/2203.02155. Citado na página 24.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: https://aclanthology.org/D14-1162. Citado na página 15.

POLYZOS, E.; WANG, F. Twitter and market efficiency in energy markets: Evidence using lda clustered topic extraction. **Energy Economics**, v. 114, p. 106264, 2022. ISSN 0140-9883. Disponível em: https://www.sciencedirect.com/science/article/pii/S0140988322004017. Citado na página 18.

RADFORD, A.; NARASIMHAN, K. Improving language understanding by generative pre-training. In: . [s.n.], 2018. Disponível em: https://api.semanticscholar.org/CorpusID:49313245. Citado nas páginas 16, 19, 23 e 24.

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019. Disponível em: https://arxiv.org/abs/1908.10084. Citado nas páginas 17 e 19.

VASWANI, A. *et al.* **Attention Is All You Need**. 2017. Disponível em: https://arxiv.org/abs/1706.03762. Citado nas páginas 12, 15, 16, 19, 23 e 24.

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

WENG, M.-H.; WU, S.; DYER, M. Identification and visualization of key topics in scientific publications with transformer-based language models and document clustering methods. **Applied Sciences**, v. 12, n. 21, 2022. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/12/21/11220. Citado nas páginas 13 e 17.

XIE, Q. *et al.* Monolingual and multilingual topic analysis using lda and bert embeddings. **Journal of Informetrics**, v. 14, n. 3, p. 101055, 2020. ISSN 1751-1577. Disponível em: https://www.sciencedirect.com/science/article/pii/S1751157719305127. Citado nas páginas 12, 14 e 18.