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
2024

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

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

Aprovada em: .

BANCA EXAMINADORA

Prof. Dr. Eduardo Manuel de Freitas Jorge
Orientador

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

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

AGRADECIMENTOS

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

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

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

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

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

RESUMO

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

SUMÁRIO

1 INTRODUÇÃO

O cenário da pesquisa científica global tem testemunhado um crescimento exponencial da produção científica nas últimas décadas, resultando em um vasto volume de dados que desafia os métodos tradicionais de organização e análise desse acervo. Para navegar nessa imensidão de informações, pesquisadores confiam em plataformas de busca, como *Web of Science*[1], *Scopus*[2] e *IEEE Xplore*[3], 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).

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

Essa arquitetura de modelos, os Transformadores, introduzida por Vaswani *et al.* (2017), revolucionou o campo da PLN com seu mecanismo de autoatenção (*self-attention*). Modelos subsequentes, como o Representações Bidirecionais Codificadas de Transformadores (BERT)[4], proposto por Devlin *et al.* (2019), passaram a capturar relações contextuais em textos com alta eficiência. 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 Latent Dirichlet Allocation (Alocação de Dirichlet Latente) (LDA) (GALLI *et al.*, 2024).

Nesse contexto, a técnica de Modelagem de tópicos baseada em embeddings BERT (BERTopic)[5], proposto por Grootendorst (2022), surge como uma abordagem moderna. Seu diferencial reside na utilização dos *embeddings* contextuais de modelos como o BERT para a modelagem de tópicos. Esta técnica permite identificar tópicos de forma dinâmica e mais

[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.

coesa, superando as deficiências de modelos tradicionais ao capturar nuances semânticas e lidar com a complexidade de textos interdisciplinares.

Este projeto de pesquisa foca no desenvolvimento de um *pipeline* computacional para o mapeamento interativo de publicações científicas. A proposta central é construir um artefato que combina a modelagem de tópicos do BERTopic Grootendorst (2022) com a ferramenta de visualização Ferramenta de visualização para mapas de conhecimento (WizMap)[6] (WANG; HOHMAN; CHAU, 2023). O estudo de caso é aplicado ao acervo do Observatório de dados públicos de ciência e tecnologia da Bahia, que coleta informações de fontes como Currículos Lattes, Plataforma Sucupira e *OpenAlex*, e tem um papel fundamental na gestão do conhecimento científico regional.

A metodologia Design Science Research (DSR) é adotada como a estrutura principal deste estudo, orientando a criação deste artefato. O objetivo é transformar a base de dados textual do Observatório em um mapa de conhecimento navegável. Nessa solução, o BERTopic é empregado para extrair os padrões temáticos e o WizMap é utilizado para a exploração visual e interativa desses tópicos. Essa integração permite a identificação de temas emergentes e a compreensão da estrutura do conhecimento científico da plataforma, indo além das análises estáticas tradicionais. Espera-se que este mecanismo otimize a experiência dos usuários e contribua para a gestão estratégica da pesquisa na plataforma.

Para articular o desenvolvimento deste estudo, a monografia segue uma progressão lógica. O Capítulo 2 estabelece o Referencial Teórico, fundamentando os conceitos de Ciência da Informação, a arquitetura dos Transformadores e as técnicas de modelagem de tópicos, com foco nos componentes do BERTopic. A seguir, o Capítulo 3 analisa os Trabalhos Correlatos, contextualizando esta pesquisa frente ao estado da arte. O Capítulo 4 detalha a Metodologia (DSR), que fornece o rigor científico para a construção do artefato proposto. O Capítulo 5, cerne deste trabalho, apresenta o Projeto de Desenvolvimento, descrevendo a arquitetura completa do *pipeline*: desde a ingestão dos dados do Observatório e a modelagem com BERTopic, até a integração final com a ferramenta de visualização interativa WizMap. Por fim, o Capítulo 6 discute os Resultados Esperados e os métodos de validação aplicados a essa solução de mapeamento de conhecimento.

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

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. 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, Tradução nossa, p. 1)

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, uma parcela significativa da produção ocorre em outros idiomas. Metodologias convencionais de análise revelam-se insuficientes para o tratamento multilíngue, o que pode restringir a circulação global do conhecimento e reduzir a visibilidade de estudos relevantes.

> 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, Tradução nossa, p. 1)

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.

> 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, Tradução nossa, p. 2)

Essa capacidade favorece a identificação de padrões temáticos em documentos que não compartilham necessariamente o mesmo vocabulário. Métodos modernos, como o BERTopic, oferecem uma estrutura metodológica para a extração de tópicos a partir dessas representações vetoriais densas. A literatura aponta que a aplicação dessas ferramentas é particularmente relevante em textos científicos heterogêneos e multilíngues, como os encontrados em grandes repositórios de publicações científicas, dada a sua robustez em capturar 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 seu objetivo central é desenvolver métodos computacionais capazes de processar, analisar, compreender e gerar a linguagem humana — seja em formato de texto ou voz — de maneira útil e análoga à humana (JURAFSKY; MARTIN, 2009)[1].

Historicamente, o PLN dependia de abordagens estatísticas e regras linguísticas manuais para modelar a linguagem (MANNING; SCHUTZE, 1999)[2]. As técnicas de PLN são projetadas para extrair significado e estrutura de dados textuais, que são inerentemente não estruturados. Isso envolve uma série de tarefas complexas, desde a análise sintática (a estrutura gramatical) até a análise semântica (o significado por trás das palavras). Tarefas comuns incluem a classificação de textos, a tradução automática, a sumarização de documentos e, de relevância particular para este referencial, a Modelagem de Tópicos (*Topic Modeling*).

A evolução recente do campo foi impulsionada pelo *Deep Learning* (Aprendizado Profundo), que permitiu a criação de representações vetoriais de alta qualidade. Como

[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.

[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.

destaca Galli *et al.* (2024), o PLN moderno depende fundamentalmente da capacidade de capturar o significado contextual.

> 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, Tradução nossa, p. 2)

Dessa forma, o PLN evoluiu de uma análise baseada em contagem de palavras para uma abordagem focada na compreensão semântica, o que viabilizou os avanços em modelagem de tópicos discutidos nas seções seguintes.

2.2.1 A Evolução das Representações Vetoriais em 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.

2.2.1.1 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 a noção de *embeddings*. Estes são vetores em espaços de alta dimensionalidade capazes de representar o significado aproximado de uma palavra.

O avanço conceitual desses modelos foi o de permitir que o significado semântico fosse quantificado. 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 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".

Galli *et al.* (2024) destacam que esta capacidade de representação numérica é o alicerce da compreensão semântica no PLN moderno:

> 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

[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.

> relacionamentos complexos entre palavras e frases usando arquiteturas especiais [...]. Galli *et al.* (2024, Tradução nossa, p. 2)

O aprendizado desses vetores ocorre através do treinamento de uma rede neural rasa em uma tarefa de previsão de contexto, conforme ilustrado na Figura 1. O artigo seminal de Mikolov *et al.* (2013) propôs duas arquiteturas principais:

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

É importante notar 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).

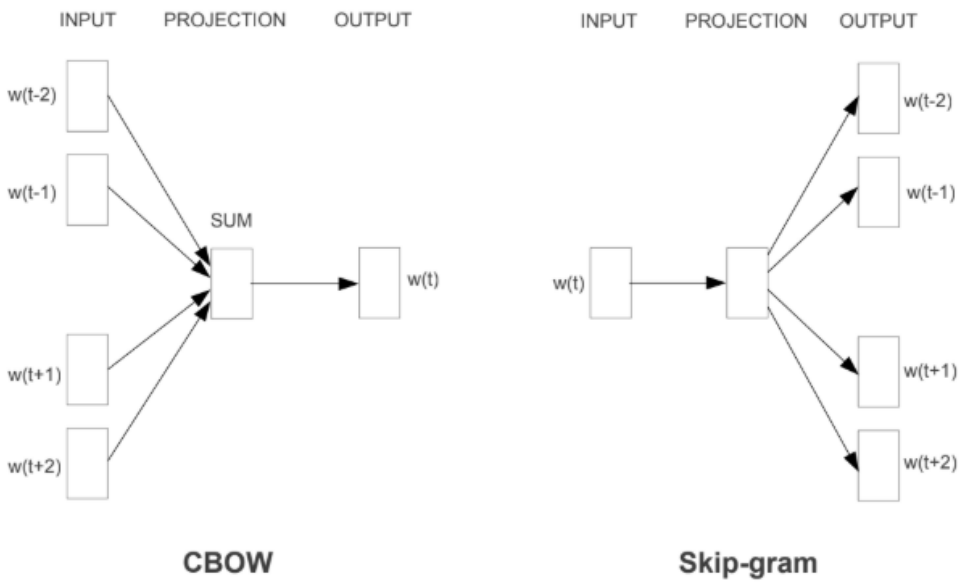

Figura 1 – Arquiteturas CBOW e *Skip-gram*.

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

Embora inovadores, 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.

2.2.1.2 A Revolução dos Transformadores e o Mecanismo de Atenção

A verdadeira virada de 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 rompeu com o paradigma das RNN e convolucionais, fundamentando-se inteiramente no mecanismo de autoatenção (*self-attention*).

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 modelos baseados em Transformadores a capacidade de gerar representações contextuais, um avanço significativo em relação às técnicas anteriores.

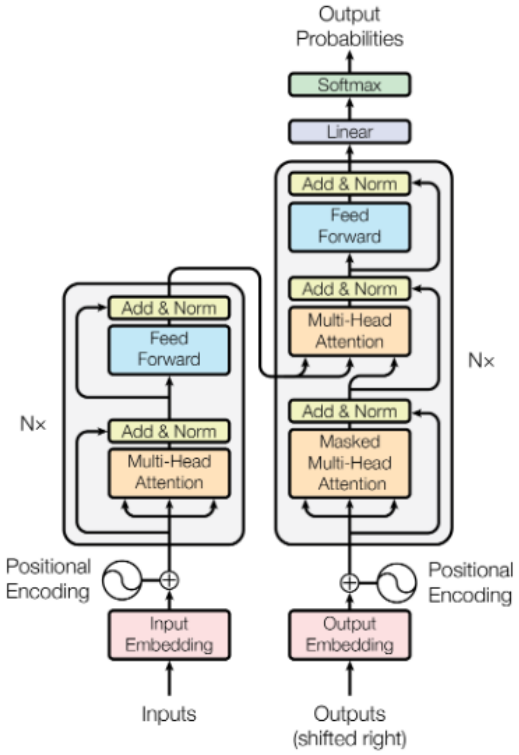

Figura 2 – Arquitetura do modelo Transformador.

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

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.

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

[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 (Rede Neural Recorrente (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).

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.

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.

2.2.1.3 Embeddings Contextuais: BERT e SBERT

Sobre essa base arquitetônica, 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 Transformador Pré-treinado Generativo (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, Tradução nossa).

Para alcançar essa bidirecionalidade sem que o modelo "visse a resposta", Devlin *et al.* (2019) introduziu o objetivo do Modelo de Linguagem Mascarado (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).

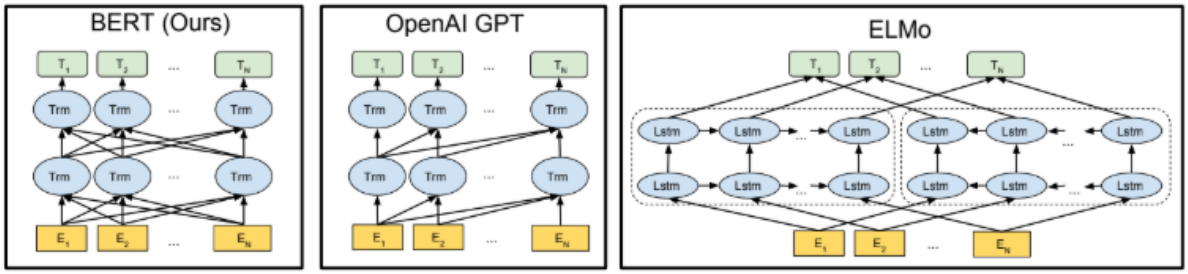

Figura 3 – Diferenças nas arquiteturas de pré-treinamento. BERT é bidirecional, GPT é unidirecional (da esquerda para a direita) e Representações de Linguagem por Modelos de Linguagem (ELMo) é uma concatenação rasa de LSTMs.

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

Apesar de sua eficácia em tarefas de classificação, a arquitetura do BERT puro mostrou-se inadequada para tarefas de busca de similaridade semântica ou *clustering*. Como Reimers e Gurevych (2019) explicam, o 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 *corpora*.

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 resolver essa limitação, Reimers e Gurevych (2019) propuseram o BERT para Sentenças (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.

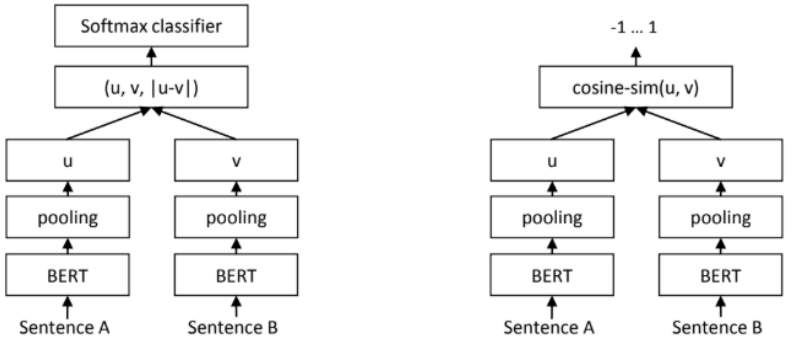

Figura 4 – Arquitetura de inferência do SBERT para computar similaridade.

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.

2.2.2 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[9] — 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.

Entre as abordagens iniciais destacam-se três marcos históricos: a Latent Semantic Analysis (Análise Semântica Latente) (LSA), a Probabilistic Latent Semantic Analysis (Análise Semântica Latente Probabilística) (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.

[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*.

A LSA, proposta por Deerwester *et al.* (1990), parte da decomposição de matrizes termo-documento por meio da técnica de Singular Value Decomposition (Decomposição em Valores Singulares) (SVD)[10]. 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; 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; 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),

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

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 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, Tradução nossa).

Essas restrições evidenciam que, embora fundamentais, tais técnicas falham em capturar o significado contextual profundo e a ordem das palavras. Essa limitação estrutural torna-os insuficientes para tarefas que exigem uma compreensão semântica robusta, especialmente em bases textuais heterogêneos ou multilíngues onde a ambiguidade lexical é alta, destacando a necessidade de abordagens que superem o paradigma *bag-of-words*.

estrutura gramatical das palavras, tratando um documento apenas como um conjunto (ou multiconjunto) de suas palavras e suas frequências.

REFERÊNCIAS

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 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, 22 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 22.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: https://arxiv.org/abs/1810.04805. Citado nas páginas 12, 19 e 20.

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, 16 e 17.

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 22 e 23.

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

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 22.

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 22.

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 15.

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.

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 15.

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 22.

MIKOLOV, T. *et al.* **Efficient Estimation of Word Representations in Vector Space**. 2013. Disponível em: https://arxiv.org/abs/1301.3781. Citado nas páginas 16 e 17.

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.

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 16.

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 22.

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 19.

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019. Disponível em: https://arxiv.org/abs/1908.10084. Citado nas páginas 20 e 21.

TAYLOR, W. L. Cloze procedure: A new tool for measuring readability. **Journalism Quarterly**, v. 30, p. 415–433, 1953. Citado na página 19.

VASWANI, A. *et al.* **Attention Is All You Need**. 2017. Disponível em: https://arxiv.org/abs/1706.03762. Citado nas páginas 12 e 18.

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.

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, 15, 16, 22 e 23.