Eixo 6 – O mundo digital: apropriações e desafios

Pesquisa vetorial em língua portuguesa: avaliação comparativa de modelos de *embeddings*

Vector Search in Portuguese: Comparative Evaluation of Embedding Models

Ernesto Carlos Bode – Biblioteca da Câmara dos Deputados – ernesto.bode@camara.leg.br

Janice de Oliveira e Silva Silveira – Biblioteca da Câmara dos Deputados – janice.silveira@camara.leg.br

Resumo: Este estudo analisa, de forma comparativa, modelos de **embeddings** aplicados à recuperação vetorial em língua portuguesa. Utilizando o *corpus* bilíngue **Pirá** — composto por pares equivalentes em português e inglês —, a pesquisa avalia a proximidade semântica entre vetores gerados por modelos multilíngues frente a um modelo focado em inglês. A metodologia emprega a **similaridade de cosseno** para mensurar o alinhamento de respostas traduzidas. Os resultados evidenciam a superioridade dos modelos multilíngues e reiteram a importância de testes criteriosos antes da implementação em sistemas de recuperação de informações. Conclui-se que a seleção do modelo impacta decisivamente a precisão de buscas em sistemas de recuperação da informação em bibliotecas e bibliotecas digitais.

Palavras-chave: Recuperação da informação. Busca semântica. *Embeddings*. Similaridade de cosseno. Língua portuguesa.

Abstract: *This study presents a comparative analysis of embedding models applied to vector retrieval in the Portuguese language. Utilizing the Pirá bilingual corpus, consisting of equivalent Portuguese-English pairs, the research evaluates the semantic proximity between vectors generated by multilingual models versus an English-centric model. The methodology employs cosine similarity to measure the alignment of translated responses. The findings demonstrate the superiority of multilingual models and emphasize the necessity of rigorous evaluation before implementation in information retrieval systems. Ultimately, model selection is shown to decisively impact search precision within library and digital library systems.*

Keywords: *Information retrieval. Vector search. Embeddings. Cosine similarity. Portuguese language*.

1 INTRODUÇÃO

A pesquisa vetorial é um recurso relativamente recente, baseado em tecnologias de inteligência artificial, que permite pesquisa semântica em documentos textuais. Comparada à forma tradicional de pesquisa para recuperação da informação, fortemente lastreada em termos e ferramentas de apoio ao uso de termos como tesauros e ontologias, a pesquisa semântica apresenta vantagens e, em muitos casos, pode substituir a pesquisa tradicional, pois possibilitam também perguntas de pesquisa em linguagem natural, tipo ”quais livros tratam de álgebra linear?”.

Um modelo de *embeddings* é um sistema de inteligência artificial treinado para converter textos em vetores numéricos — listas de números que representam o significado semântico de palavras, frases ou documentos em um espaço multidimensional. A partir da leitura de grandes volumes de texto, esses modelos aprendem a posicionar conceitos semanticamente semelhantes em regiões geometricamente próximas desse espaço, de modo que a proximidade de significado passe a ser mensurável matematicamente. A qualidade dessa representação depende, entre outros fatores, dos idiomas presentes nos dados de treinamento do modelo.

O objetivo deste artigo é apresentar um método simples de avaliação multilíngue capaz de comparar o desempenho de modelos de *embeddings* na representação semântica de textos em língua portuguesa, por meio da comparação de desempenho. Busca-se verificar em que medida modelos multilíngues preservam a proximidade semântica entre frases traduzidas e implicações práticas nos sistemas de recuperação da informação largamente utilizados em bibliotecas tradicionais e digitais.

1.1 A pesquisa vetorial semântica

O termo “vetorial” aqui tem significado estritamente matemático e geométrico. Um vetor é um artifício (geométrico, ou algébrico) para representar coisas como forças (em física) ou meramente abstrações matemáticas de elementos físicos e químicos. Mas de fato, no âmbito da inteligência artificial, podem representar algebricamente pontos de imagens fotográficas, sons, além de texto (termos em separado, frases, artigos e até livros inteiros). Nesse fórum, não é o caso de aprofundamento matemático, mas o ponto importante é compreender como um elemento de uso matemático pode ser usado para

representações semânticas de conceitos como psicologia, engenharia, sexualidade e etc.?

O primeiro elemento da resposta a essa pergunta é compreender um espaço de vetores representando um determinado universo de conceitos. Como exemplo, uma classe de livros de biblioteca, como história/geografia/biografia.

É possível associar um vetor (um ente matemático com direção, sentido e valor de comprimento) a cada livro em um conjunto dado de livros, nesse caso livros de história. O que cada vetor associado a cada livro teria de diferente são os dados referentes à direção, sentido e valor, mas seguindo uma regra específica: livros com temas e assuntos próximos estarão próximos no mesmo universo geométrico definido. Considere que 'história da Europa' e 'Roma antiga' teriam vetores matematicamente parecidos, mas diferentes de livros sobre 'história moderna da América do Sul'. Assim, teríamos agrupamentos de vetores próximos (geometricamente) com base nos assuntos próximos nos livros. Para uma compreensão ampla do uso de vetores e conceitos (GUZMÁN, 2026).

Em um espaço vetorial, livros com temas semanticamente próximos tendem a possuir vetores geometricamente semelhantes. Assim, uma consulta como "história de Roma" pode ser transformada em vetor e comparada com vetores próximos gerados a partir dos documentos do acervo. Esse processo, executado por modelos de IA, permite recuperação semântica sem depender exclusivamente de termos previamente atribuídos por catalogação ou indexação. Trata-se, portanto, de uma abordagem complementar aos métodos tradicionais de organização e recuperação da informação. Mais sobre modelos semânticos (NHACUONGUE, 2025; MANIKANTA, 2025).

1.2 Modelos para vetorização

A vetorização pode ser feita para qualquer tipo de conteúdo digital, mas para modelos de vetorização de textos (incluindo caracteres numéricos e especiais), normalmente usa-se o termo *embeddings*. Matematicamente, um "*embeddings*" é apenas uma lista de números reais. Quando uma palavra ou frase é processada por um modelo, ele entrega algo assim:

$$\text{Vetor}_{(\text{telefone})} = [0.25, -1.82, 0.03, 4.11, -0.76, ...]$$

Cada um desses números representa uma coordenada em um **espaço multidimensional**. Enquanto nós conseguimos visualizar até 3 dimensões (X, Y, Z), os *embeddings* operam em espaços de centenas ou milhares de dimensões (por exemplo, 300, 768 ou 1536). As dimensões extras nos *embeddings* são fundamentais para a precisão da busca semântica porque permitem capturar mais nuances, o contexto e o significado real das frases, superando as limitações da busca baseada apenas na correspondência exata de palavras-chave. Sobre representações em mais de três dimensões, vide (MARIUTTI, 2025). O número de dimensões de cada modelo é definido no momento da criação e treinamento dos modelos.

Para simplificação didática, imagine um espaço de apenas três dimensões. Considere que o eixo X (horizontal) represente o quão "imagético" um livro é (ilustrado), o eixo Y (vertical) represente o quão "textual" um livro é, e o eixo z (profundidade) represente o quão científico um livro é (Quadro 1):

Quadro 1 – Vetores hipotéticos

Vetores	Sua representação hipotética
Um livro de fotografia seria vetor [0,9; 0,4; 0,2]	(muito imagético, pouco textual, muito pouco científico)
Um livro religioso seria vetor [0,3;0,9;0,2]	(pouco imagético, muito textual, muito pouco científico)
O livro cientifico seria vetor [0,4;0,9;0,9]	(pouco imagético, muito textual, muito científico)

Fonte: elaborado pelos autores

Em aplicações reais, com um modelo de centenas de dimensões, os eixos não têm rótulos humanos tão claros ("imagético", "textual", "científico"). Eles representam características que a rede neural aprendeu sozinha ao ler bilhões de textos. A premissa principal é a hipótese distribucional (*distributional vector space models*): *palavras que estão nos mesmos contextos tendem a ter significados semelhantes* e estar nos mesmos espaços (SOGAARD; FARUQUI; VULIC, 2019).

Como o modelo é treinado (criado) lendo textos reais, ele ajusta as coordenadas (os números que compõem os vetores) de forma que conceitos que são usados de maneira semelhante na língua sejam "empurrados" para a mesma região geométrica desse espaço multidimensional. Pelo menos é isso que idealmente um modelo de *embeddings* deveria fazer.

É por isso que conceitos semelhantes resultam em vetores semelhantes. Se mapearmos geometricamente:

- Os vetores de "Roma Antiga" e "Césares" estarão muito próximos um do outro.
- A palavra "Brasil" estará num aglomerado de vetores totalmente diferente, junto com "Bolívia" e "Argentina", por exemplo.

1.3 Comparação e escolha de modelos de *embeddings*

Um dos principais desafios dos modelos de embeddings está relacionado ao idioma utilizado em seu treinamento e ao idioma em que serão aplicados. Como a maioria desses modelos foi treinada predominantemente com textos em inglês, seu desempenho pode variar em outros idiomas. Embora fatores como precisão e velocidade sejam relevantes, interessa-nos analisar a capacidade dos modelos de representar conceitos equivalentes em diferentes línguas. Idealmente, textos semanticamente equivalentes deveriam gerar vetores geometricamente próximos, o que, como demonstrado nos testes, não ocorre de forma consistente em todos os modelos avaliados.

Mas, “Como comparar modelos de *embeddings* em diferentes idiomas?”. A solução adotada aqui é a comparação do ângulo entre os vetores na **língua A** e os vetores gerados na **língua B**. A similaridade angular é frequentemente utilizada como aproximação da proximidade semântica. Assim, “Direito” e “Direito Constitucional” são dois conceitos que teriam uma representação geométrica com ângulos muito parecidos. Mas direito e psicologia infantil, por sua vez, teriam diferenças de ângulos maiores. Sobre similaridade de ângulos e cosseno e comparação com outros métodos (LIU, 2026). Há outros métodos matemáticos para comparar vetores, como a distância euclidiana (MACÊDO, 2018). A seguir, a fim de melhor subsidiar os conceitos, vamos expor um exemplo matemático simplificado e com valores arredondados.

O cálculo matemático para a similaridade de cosseno é feito pela fórmula: $\cos(\theta) = A \cdot B / \|A\| . \|B\|$, $A \cdot B$ é o produto escalar dos componentes de cada vetor, para a par, e $\|A\| . \|B\|$ é o produto da norma de cada vetor. Considere este exemplo:

A = [0,134, -0,221, 0,587, 0,091, -0,332... (até 384 dimensões)]

B = [0,129, -0,210, 0,601, 0,080, -0,340]... (até 384 dimensões)]

Produto escalar par a par:

A·B = (0,134×0,129) + (-0,221×-0,210) + ... todas as dimensões)

> 0,0173 + 0,0464 + 0,3528 + 0,0073 + 0,1129 = ***0,537***

Normas:

‖A‖*= sqrt(0.134² + (-0,221)² + 0,587² + 0,091² ....)*

> sqrt(0,529) = **0,727**

‖B‖ *= sqrt(0,129² + (-0,210)² + 0,601² + 0,080² + (-0,340)²...)*

sqrt(0,543) = ***0,737***

Similaridade:

cos = 0,5367 / (0,727 × 0,737) aproximadamente ***0,970*** *(para o par A e B considerado)*

Esse valor praticamente igual a 1 indica vetores quase paralelos com significados equivalentes.

2 METODOLOGIA

A metodologia proposta compara modelos de *embeddings* para identificar sua adequação ao uso com textos em língua portuguesa. O método utiliza pares equivalentes inglês-português e compara a similaridade de cosseno entre os vetores gerados em cada idioma. Espera-se que textos semanticamente equivalentes produzam vetores geometricamente próximos. Ao final, as médias das similaridades obtidas permitem comparar o desempenho dos modelos avaliados.

1.4 O *corpus* utilizado

Para as tarefas de geração e comparação dos *embeddings* utilizamos o corpus *Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean, the Brazilian coast, and climate change*. Esse *dataset* está disponível publicamente em repositórios, utilizamos a versão no repositório GitHub[1]. Trata-se de um projeto brasileiro levado a cabo pelo *Center for Artificial Intelligence*[2] ligado à Universidade de São Paulo.

Utilizamos especificamente o arquivo Pirá-F-Total.csv[3] com 1346 linhas em inglês e outras 1346 linhas em tradução para o português. As linhas presentes no mesmo

[1] <https://github.com/C4AI/Pira/tree/main>
[2] <https://c4ai.inova.usp.br/>
[3] <https://github.com/C4AI/Pira/blob/main/Pira1.0/CSV/Pir%C3%A1-F-Total.csv>

arquivo acima citado foram utilizadas para criar dois *dataframes*[4]: corpus_answer_en_origin (inglês) e *corpus_answer_pt_origin* (português).

2.2. Ciclo de comparação dos modelos

O ciclo de tarefas executadas em linguagem *python* é de baixa complexidade e está disponível publicamente em repositório GitHub[5]. As etapas executadas foram:

1. **Importação do *Dataset* Pirá**. A importação dos dados compreendeu especificamente um dos arquivos disponíveis dentro do *Dataset* (*formato CSV, Pirá-F-Total.csv*)[6].
2. **Criação de dois *corpus* a partir do *dataset*** (inglês e português). Cada corpus criado compreende 1.346 linhas com versões do mesmo texto em inglês e português. Esses idiomas são ideais para essa comparação, pois os modelos seguramente utilizaram predominantemente inglês em sua fase de treino e a língua portuguesa é justamente a variável que queremos analisar nesse contexto de modelos do tipo *embeddings*.
3. **Definição e importação de modelos de *embeddings* para avaliação**. Foram escolhidos três modelos de *embeddings*, dois multilíngues e um treinado apenas em inglês (no qual espera-se performance ruim na língua portuguesa). Os modelos escolhidos foram:

```
model_names = {
   'Multilingual-MiniLM': 'paraphrase-multilingual-MiniLM-L12-v2',
   'Multilingual-DistilUSE': 'distiluse-base-multilingual-cased-v1',
   'English-MiniLM': 'all-MiniLM-L6-v2' # Este não é multilíngue
}
```

4. **Criação dos *embeddings* nas duas línguas consideradas**. Apenas uma única linha de código foi necessária para criar os *embeddings* de cada corpus:

```
embeddings_pt=
model.encode(corpus_answer_pt_origin, convert_to_tensor=True)
```

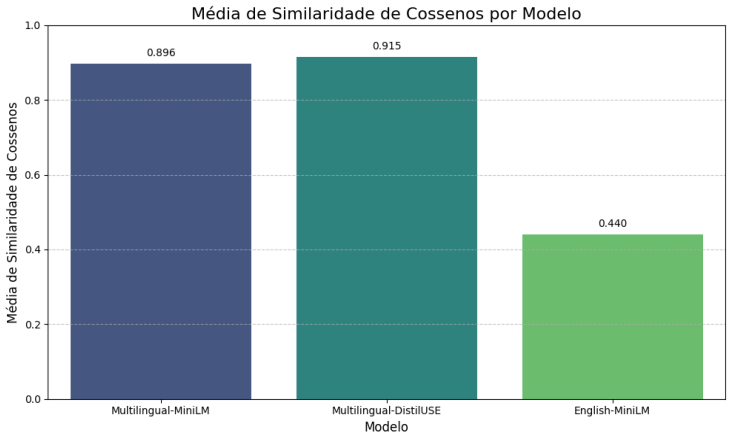
[4] Objetos da biblioteca *python* Pandas para manipulação de dados.

[5] <https://github.com/ErnyBSB/*embeddings*ModelComparing/blob/main/Assess*Embeddings*sModels_04.ipynb>

[6] <https://github.com/C4AI/Pira/blob/main/Pira1.0/CSV/Pir%C3%A1-F-Total.csv>

5. **Comparação dos cossenos obtidos**. Utilizamos a biblioteca *python pytorch*[7] para essa tarefa, que na verdade é a tarefa central no processamento do corpus nas duas línguas.
6. **Cálculo de Médias** (com desvio padrão e valores máximo e mínimo). Finalmente, a partir dos valores obtidos (*paired_cosine_scores*) são calculados média, respectivo desvio padrão e valores máximo e mínimo considerados na análise, para cada par equivalente que foi comparado.
7. **Exibição em formato gráfico.** Por último, como facilitador de visualização dos resultados, é apresentado um gráfico com a comparação dos três modelos e a média das similaridades dos cossenos nos vetores comparados.

Gráfico 1 – Comparação modelos testados



Fonte: elaborado pelos autores
Descrição: *comparação pela média de similaridade cossenos*

O código utilizado, inclusive com os resultados em visualização *Google Colab*[8], está disponível no repositório com exaustivos comentários e podem ser livremente reproduzidos.

3 RESULTADOS E DISCUSSÕES

Os resultados obtidos são coerentes com os modelos escolhidos, ou seja, a comparação de similaridade de cossenos entre vetores em língua inglesa e portuguesa são altos e similares entre os modelos multilíngues considerados. E baixos no modelo

[7] <https://pytorch.org/>
[8] <https://colab.research.google.com/>

treinado exclusivamente em inglês. Isso não significa, necessariamente, que todo modelo multilíngue ofereça bons resultados, ou mesmo que todo modelo treinado apenas em inglês ofereça um resultado ruim em português (Quadro 2).

Isso nos leva ao segundo resultado importante. Trata-se de um método para teste prévio do modelo de *embeddings* que poderá ser adotado num sistema de recuperação de informações em língua portuguesa.

É fato também que nessa área novos modelos surgem constantemente e a unidade de informação que utilize a língua portuguesa (bibliotecas e outras unidades de informação) que tenha interesse em utilizar um sistema de pesquisa vetorial podem e devem checar quais modelos estão disponíveis e avaliar tais modelos em relação ao seu vernáculo.

Quadro 2 – Modelos avaliados

Modelo	Tipo	Média	Desvio	Mín.	Máx.	Interpretação
paraphrase-multilingual-MiniLM-L12-v2	Multilíngue	0,8965	0,1011	0,1680	1,0000	Alinhado PT-BR
distiluse-base-multilingual-cased-v1	Multilíngue	0,9148	0,0709	0,4368	1,0000	Alinhado PT-BR
all-MiniLM-L6-v2	Inglês	0,4403	0,2168	-0,0144	1,0000	Não alinhado PT-BR

Fonte: dados gerados por código *python*, compilado pelos autores
Descrição: resultados estatísticos da comparação

4 CONSIDERAÇÕES FINAIS

Há vários outros modelos que poderiam ser utilizados, inclusive mais recentes como o BGE-M3 ou multilingual-e5. Como o código utilizado, bem como o *dataset* Pirá estão disponíveis em *Creative Commons*, as adaptações devem ser de fácil implementação.

Este artigo busca dialogar com a área de Ciência da Informação, mais especificamente com o campo da Recuperação da Informação, apresentando e esclarecendo uma técnica emergente cuja adoção em sistemas de bibliotecas é crescente e cujos impactos sobre a qualidade da busca merecem atenção crítica por parte dos profissionais da área. Entendemos que a pesquisa vetorial não é apenas uma inovação tecnológica, mas uma mudança de paradigma na forma como documentos são representados e recuperados — o que a torna objeto legítimo de estudo e reflexão no âmbito da Ciência da Informação.

Contudo, dado que o presente artigo também conversa com as áreas de Ciência da Computação e Matemática, optou-se por dedicar parte relevante do espaço disponível ao esclarecimento dos termos e conceitos técnicos empregados — tarefa necessária para que o texto seja acessível ao público da área de informação sem formação específica naquelas disciplinas. Nos limites de páginas estabelecidos pelo formato do evento, não seria possível realizar uma revisão de literatura aprofundada sem ultrapassar consideravelmente o espaço disponibilizado. Essa escolha é, portanto, consciente e justificada pela natureza interdisciplinar do tema.

Algumas limitações desta metodologia são: (1) não houve testes e validação humana; (2) utilizamos apenas três modelos, relativamente antigos; (3) a técnica de similaridade de cosseno não verifica a recuperação real em sistemas, o que necessita de outros elementos tecnológicos e matemáticos.

As perspectivas de aplicação são imediatas em sistemas de recuperação da informação em bibliotecas. Atualmente, vivemos um momento de transição na adoção da pesquisa vetorial, e os modelos utilizados impactam a qualidade da recuperação em língua portuguesa, visto que tais modelos utilizam porções menores de nossa língua em seus treinamentos.

REFERÊNCIAS

GUZMÁN, Fernando Martínez. Embeddings y modelos conversacionales: una nueva forma de explorar, comprender, describir y conectar la información en bibliotecas. **Dossier Anuario SEDIC**, [S. l.], 2026. Não paginado. Disponível em: https://edicionsedic.es/dossier/article/view/192. Acesso em: 1 maio 2026.

LIU, Xiqiu; WANG, Canrong. **Semantic Information Modeling with Retrieval-Augmented Generation for Academic Digital Libraries. Research Square**, [S. l.], mar. 2026. Preprint. Disponível em: https://www.researchgate.net/publication/402851229_Semantic_Information_Modeling_with_Retrieval-Augmented_Generation_for_Academic_Digital_Libraries. Acesso em: 1 maio 2026.

MACÊDO, Lincoln. Distância Euclidiana vs Similaridade de Cossenos. **Código (Blog),** [S. l.], 2018. Disponível em: https://demacdolincoln.github.io/anotacoes-nlp/posts/distancia-euclidiama-vs-similaridade-de-cossenos/. Acesso em: 1 maio 2026.

MANIKANTA, Nihanth. Building a Hybrid Information Retrieval System using BM25 and Sentence-BERT. **Medium (Revista e Blog on line)**, [S. l.], out. 2025. Disponível em:

https://medium.com/@nm.pattabhi/building-a-hybrid-information-retrieval-system-using-bm25-and-sentence-bert-ef0ae924b1da. Acesso em: 1 maio 2026.

MARIUTTI, Eduardo Barros. Tecnologias da percepção. **Texto para discussão**, Campinas, n. 486, set. 2025. ISSN 0103-9466. Disponível em: https://www.eco.unicamp.br/images/arquivos/artigos/TD/TD486.pdf. Acesso em: 1 maio 2026.

NHACUONGUE, Januário Albino. Inteligência Artificial e Recuperação da Informação: desafios e implicações éticas emergentes. In: ISKO Brasil, 2025, Canela. **Anais...** [S.l: s.n.], 2025. DOI: 10.22477/ISKO25.91. Disponível em: https://isko.org.br/ojs/index.php/iskobrasil/article/view/91. Acesso em: 1 maio 2026.

SOGAARD, Anders; FARUQUI, Manaal; VULIC, Ivan. **Cross-Lingual Word Embeddings.** [S. l.]: Morgan & Claypool, 2019. ISBN 978-1-68173-063-9. Disponível em: https://dl.acm.org/doi/abs/10.5555/3360269. Acesso em: 1 maio 2026.