Classificação de Textos com Matrizes de Similaridade: Um Estudo de Caso em um Corpus com Nomes de Produtos

Antonio Ferreira de Castro Barbosa[1], Eduardo Corrêa Gonçalves[2]

[1]Bacharelado em Sistemas de Informação – Universidade Federal do Estado do Rio de Janeiro (BSI/UNIRIO)
Avenida Pasteur, 458 – 22.290-255 – Rio de Janeiro – RJ – Brasil

[2]Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
Rua André Cavalcanti, 106 – 20.231-050 – Rio de Janeiro – RJ – Brasil

antoniofdcbarbosa@gmail.com, eduardo.correa@ibge.gov.br

Abstract. *This paper addresses the problem of semantic comparison of pairs of short strings in Portuguese. More specifically, this work evaluates an unsupervised approach based on similarity matrices for classifying product names that may contain different specifications. The matrices were evaluated individually and combined in pairs. In preliminary experiments on a database of 1,000 instances, combining the TF-IDF matrix with matrices built with pre-trained BERT embeddings achieved the best accuracy.*

Resumo. *Este trabalho aborda o problema da comparação semântica de pares de strings curtas em português. Mais especificamente, o trabalho avalia uma abordagem não supervisionada baseada em matrizes de similaridade para classificar nomes de produtos contendo diferentes tipos de especificações. As matrizes foram utilizadas de forma isolada e combinadas aos pares. Em experimentos realizados sobre um corpus composto por 1.000 instâncias, a melhor acurácia foi obtida através da combinação da matriz TF-IDF com matrizes de embeddings pré-treinados da família BERT.*

1. Introdução

Plataformas digitais usualmente empregam técnicas de busca semântica para aprimorar a experiência de busca em suas aplicações [Huang et al. 2020]. Por exemplo, em uma aplicação de busca de produtos em um supermercado, é possível contornar o problema da falta de resultados com duas possibilidades diferentes: (i) apresentar o mesmo produto, mas de marcas alternativas (ex.: o cliente procura por manteiga da marca A, que não existe no estoque, então a aplicação retorna as manteigas das marcas B e C); (ii) apresentar produtos similares, independente da marca (nesse caso, a aplicação poderia retornar uma margarina). Para que o sistema seja capaz de apresentar esses resultados similares, é necessário que haja uma forma de classificar automaticamente os nomes comerciais dos produtos para nomes padronizados [Köpcke and Rahm, 2010].

Nesse contexto, o presente trabalho utiliza-se de um corpus de 1.000 pares relacionando nomes de produtos alimentícios vendidos em supermercados com o seu nome padronizado no Índice Nacional de Preços ao Consumidor (INPC) do IBGE [IBGE 2016]. A partir desse *dataset*, matrizes de similaridade de cosseno para diferentes representações dos textos (TF-IDF [Manning et al. 20028], BERTimbau Base

[Souza et al. 2020], mBERT [Google 2020] e Universal Sentence Encoder (USE) [Cer et al. 2018]) foram utilizadas para classificar o nome INPC correto de cada produto.

2. Trabalhos Relacionados e Metodologia

2.1. Trabalhos Relacionados

Romualdo et al. (2021) abordam a tarefa de mensurar o grau de similaridade entre títulos de produtos comercializados em um grande site de *e-commerce*. Tais títulos são formados não apenas pelo nome do produto como também por uma série de especificações (ex.: "Cartucho Epson 196 Preto 5ml T196120"). Em testes com um dataset de 400 pares de títulos, *embeddings* genéricos mBERT e BERTimbau superaram os estáticos treinados com textos do domínio (FastText, Word2Vec e GloVe [Jurafsky e Martin 2026]) ao distinguir produtos similares de dissimilares.

Araujo et al. (2023) apresentam um estudo de caso que investigou o emprego de matrizes de similaridade de cosseno [Manning et al. 2008] para identificar boletins policiais semelhantes. Diferentes representações foram avaliadas, tais como TF-IDF e *embeddings* de sentenças USE e SBert [Reimers and Gurevych 2019] pré-treinados e treinados com textos do domínio. Em uma coleção de 1.089 boletins anotados, o melhor resultado foi obtido pela representação TF-IDF (77,38% de acurácia), seguida pelos embeddings de sentença pré-treinados USE (70,47%).

Meirelles et al. (2022) propõem o uso combinado de matrizes de similaridade para considerar simultaneamente medidas que atuam em diferentes subáreas do sistema linguístico [Caseli et al. 2026]. A avaliação utilizou uma base com 3.910 pares casados de nomes de produtos e serviços de duas pesquisas do IBGE (POF e INPC) [IBGE 2026]. O estudo constatou que combinar medidas que atuam no nível morfológico, como Jaro-Winkler [Winkler at al.1990], com uma medida semântica (cosseno de *embeddings* Word2Vec pré-treinados do NILC [Hartmann et al. 2017]) aumentou a eficácia do processo de pareamento de nomes equivalentes.

2.2. Base de Dados

O presente estudo utiliza um corpus com 1.000 pares (x, y) de nomes de produtos alimentícios [de Lima and Gonçalves, 2022], em que x representa o nome de um produto vendido em um supermercado, enquanto y representa a classe (nome padronizado) do produto no INPC [IBGE, 2016]. Há um total de 470 possíveis classes. Alguns exemplos de pares (x, y) do corpus são: {(*Repolho Verde (unidade) – orgânico*, *Repolho orgânico*), (*Patê de Atum Gomes da Costa*, *Patê*), (*Carre Suino Congelado Pc 3,5 Kg*, *Carne de porco*)}. Os nomes dos produtos nos supermercados são mais longos (34,4 caracteres e 6,5 tokens em média) e podem conter erros ortográficos (ex.: "Carre" em vez de "Carré"), além de abreviações, nomes de marcas, unidades de medida e outras especificações. Por sua vez, os nomes do INPC são padronizados e concisos (média de 13,5 caracteres e 2,2 tokens).

2.3. Matrizes de Similaridade

A partir do corpus descrito acima, torna-se possível empregar uma abordagem supervisionada para treinar um modelo que classifica o nome de um produto de supermercado em um nome do INPC. Contudo, uma dificuldade para esta abordagem é que o *dataset* contém, em média, apenas 2,13 instâncias por classe (são 1.000 instâncias, cada qual anotada com uma dentre 470 classes possíveis). Por esta razão, decidiu-se

tratar o problema utilizando uma estratégia não supervisionada baseada em matrizes de similaridade. Cada matriz de similaridade *M* avaliada possui 1.000 linhas (nomes dos produtos nos supermercados) e 470 colunas (nomes do INPC, que representam as classes do problema). As células representam o valor da similaridade de cosseno entre cada nome das linhas e cada nome das colunas de acordo com alguma representação vetorial. Para os experimentos preliminares apresentados na próxima seção, foram escolhidas as representações com melhor desempenho nos trabalhos de Araujo et al. (2023) e Romualdo et al. (2021): TF-IDF, embeddings pré-treinados de palavras BERTimbau Base e mBERT e embeddings pré-treinados de sentença USE. Foram avaliadas matrizes de similaridade individuais dessas representações e combinadas aos pares, de acordo com a proposta de Meirelles et al. (2022) (detalhada na Seção 3.2).

3. Experimentos

Nesta seção, são apresentados os resultados obtidos nos experimentos. O pacote *scikit-learn* [Pedregosa et al. 2011] foi utilizado para obter a representação TF-IDF dos textos, enquanto as implementações disponibilizadas no *Hugging Face* [Hugging Face 2026] foram utilizadas para os modelos BERTimbau, mBERT e USE, que não foram submetidos a ajuste fino. Os modelos foram testados com e sem o pré-processamento das strings, que consistiu na remoção de stop words e lematização com a ferramenta *spaCy* [Honnibal and Montani 2017]. Para TF-IDF, BERTimbau e mBERT, os melhores resultados ocorreram com pré-processamento; já para o USE, o desempenho foi superior sem essa operação. Dois tipos de métricas de avaliação de desempenho foram adotadas. A primeira é a Acurácia Estrita [Meirelles et al. 2022], que consiste na proporção de vezes em que a matriz de similaridade mapeou única e corretamente a classe INPC para um nome de produto. A segunda é a Acurácia *Top-k*, que considera a classificação correta caso o valor de similaridade da classe real do produto esteja entre os *k* maiores da matriz.

3.1. Desempenho Isolado das Matrizes

A Tabela 1 apresenta os resultados do primeiro experimento, que comparou o desempenho individual das matrizes. A representação TF-IDF obteve Acurácia Estrita de 61,90%, quase 25% superior ao do segundo melhor resultado, obtido pelo BERTimbau. O TF-IDF também alcançou maiores valores para a Acurácia *Top-k*, mas dessa vez com menor diferença para os embeddings (ex.: diferença de 7,2% para o BERTimbau na Acurácia *Top*-10). É possível ainda observar que o crescimento no valor da Acurácia Estrita para a Acurácia *Top-k* é menos acentuado no TF-IDF do que nos embeddings. Isso pode ser justificado pelo fato de a matriz TF-IDF ser muito esparsa (média de apenas 7,1 células com valores diferentes de zero por linha).

Tabela 1. Desempenho das Matrizes Individuais

Matriz de Similaridade	Acurácia Estrita	Acurácia *Top*-5	Acurácia *Top*-10
TF-IDF	**0,6190**	**0,6820**	**0,6890**
BERTimbau	0,3700	0,5640	0,6170
mBERT	0,3530	0,5190	0,5870
USE	0,2650	0,4380	0,5250

O resultado dos *embeddings* pode estar relacionado ao tamanho muito curto das strings e a consequente falta de contexto. Por exemplo, a representação USE (pior valor de Acurácia Estrita, com 26,50%) foi projetada para codificar sentenças. Porém, no experimento foi empregada para representar vetorialmente nomes muito curtos, frequentemente formados apenas por um substantivo principal seguido de um reduzido número de adjetivos ou locuções adjetivas (ex.: "Rabada bovina congelada 1Kg").

3.2. Desempenho de Pares de Matrizes Combinadas

O segundo experimento avaliou pares de matrizes combinadas, utilizando a abordagem proposta em Meirelles et al. (2022). Nesta abordagem, uma matriz híbrida M_H é obtida utilizando a Equação 1, em que M_1, M_2, ..., M_n são *n* matrizes de similaridade previamente geradas, enquanto $\alpha \in \mathbb{R}+$ atua como ponderação, possibilitando dar maior peso para valores de similaridade maiores. Para o experimento, foi utilizado $\alpha = 1$.

$$M_H = \frac{(M_1^\alpha + M_2^\alpha + \cdots + M_n^\alpha)}{n} \quad (1)$$

Foram geradas 3 matrizes híbridas M_{H1}, M_{H2} e M_{H3} representando, respectivamente, a média da matriz TF-IDF (melhor resultado no experimento anterior) com as matrizes BERTimbau, mBERT e USE. Os resultados são apresentados na Tabela 2. Com relação à Acurácia Estrita, as matrizes híbridas MH_1 e MH_2 (TF-IDF e BERTimbau e TF-IDF e mBERT) obtiveram resultados ligeiramente superiores a melhor matriz do experimento anterior (TF-IDF). Contudo, o mais interessante foi a grande melhora nos valores da Acurácias *Top-k* das matrizes combinadas. Considerando a aplicação descrita na Introdução, um sistema de busca de produtos em supermercados, o resultado de Acurácia *Top*-5 acima de 81% obtido pelas matrizes híbridas indica que, na maioria das vezes, o nome INPC correto estaria entre as primeiras sugestões apresentadas ao usuário. Esse resultado é especialmente relevante em cenários em que o sistema não necessita de uma classificação única e objetiva, mas sim de uma seleção de alternativas razoáveis, como na substituição de produtos fora de estoque.

Tabela 2. Desempenho das Matrizes Combinadas

Matriz de Similaridade	Acurácia Estrita	Acurácia Top-5	Acurácia Top-10
MH_1 (TF-IDF e BERTimbau)	**0,6290**	0,8180	0,8480
MH_2 (TF-IDF e mBERT)	0,6280	0,8190	0,8450
MH_3 (TF-IDF e USE)	0,5980	**0,8230**	**0,8610**

4. Conclusões

Os resultados dos experimentos preliminares sugerem que a combinação de matrizes de similaridade pode aprimorar o processo de classificação e busca semântica de textos curtos, como nomes de produtos. Entretanto, no corpus testado, os resultados dos *embeddings* pré-treinados isoladamente foram bem inferiores aos do TF-IDF. Sendo assim, no futuro pretende-se realizar o ajuste fino desses modelos utilizando bases de dados públicas que contenham nomes de produtos. Intenciona-se ainda avaliar o desempenho de modelos *decoder-only* (como em Pellegrini et al. (2025) e Yepez et al. (2024)) e empregar outras estratégias para combinar as matrizes de similaridade.

5. Limitações do Estudo

Os resultados apresentados nesse estudo são preliminares e envolvem um único conjunto de dados composto por 1.000 instâncias. Foram utilizadas versões pré-treinadas dos embeddings avaliados, sem a realização de ajuste fino com textos do domínio.

6. Declaração de Ética

O corpus utilizado foi construído exclusivamente a partir de dados de acesso público, sendo eles os nomes dos produtos do INPC do IBGE e nomes de produtos anunciados em sites de supermercados. Desta forma, não foi necessária a aprovação por parte de comitê de ética. O método proposto pode ser visto como uma ferramenta de apoio que visa aprimorar o desempenho de sistemas de busca semântica.

7. Uso de IA Generativa

Os autores declaram que houve utilização da IA generativa somente como apoio à implementação de trechos do código utilizado nos experimentos. Não houve a utilização de IA generativa para produzir ou modificar o texto do artigo.

Referências

Araújo, J., Silva, T., Rocha, A., and Lira, V. (2023). Police Report Similarity Search: A Case Study. In *Proceedings of the XII Brazilian Conference on Intelligent Systems (BRACIS 2023)*, pages 394-409, Belo Horizonte, Springer.

Caseli, H. M., Nunes, M. G. V., and Pagano, A. (2026). O que é pln? In Caseli, H. M. and Nunes, M. G. V., editors, Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português – Volume 1, book chapter 1. BPLN, 4 ed.

Cer, D., et al. (2018). Universal sequence encoder. arXiv preprint arXiv:1803.11175. Disponível em: https://arxiv.org/abs/1803.11175.

de Lima, L. S. G. and Gonçalves, E. C. (2022). Similaridade Semântica de Nomes de Produtos Alimentícios Utilizando Wordnets do Português. In *Proceedings of the 15th Seminar on Ontology Research in Brazil (ONTOBRAS) and 6th Doctoral and Masters Consortium on Ontologies (WTDO)*, Online, CEUR-WS.org.

Google (2020). BERT. Disponível em: https://github.com/google-research/bert

Hartmann, N. S., et al. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL 2017)*, pages 122–131, Uberlândia, SBC.

Honnibal, M. and Montani, I. (2017). "spaCy: Industrial-strength Natural Language Processing in Python", https://spacy.io.

Huang, J.-T., Sharma, A., Sun, S., Xia, X., Zhang, D., et al. (2020). Embedding-based Retrieval in Facebook Search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–256, Online. ACM.

Hugging Face. (2026). "Models", https://huggingface.co/

IBGE (2016). Para compreender o INPC (um texto simplificado), 7a. ed, IBGE.

IBGE (2026). “Inflação”, https://www.ibge.gov.br/explica/inflacao.php.

Jurafsky, D. and Martin, J. H. (2026) “Embeddings”. In: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models. 3rd ed. Stanford University, 2026. Cap. 5. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 14 mai. 2026.

Köpcke, H. and Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2): 197–210. Elsevier

Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval, Cambridge University Press.

Meirelles, T. P., Gonçalves, E. C., and Gomes, D. T. (2022). Pareamento de Nomes de Produtos e Serviços Utilizando Medidas de Similaridade Textual nos Níveis Alfabético, Léxico e Semântico. *Cadernos Do IME - Série Informática*, 46:104–117. IME / UERJ.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830. MIT Press.

Pellegrini, L., Santos, F., and Cantarino, F. (2025). Classificação de Notícias Falsas na Língua Portuguesa Utilizando Modelos baseados na Arquitetura Transformer. In *Proceedings of the 16th Brazilian Symposium in Information and Human Language Technology (STIL 2025)*, pages 549-556, Fortaleza. SBC.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. Disponível em: https://arxiv.org/abs/1908.10084.

Romualdo, A., Real, L., and Caseli, H. (2021). Measuring Brazilian Portuguese Product Titles Similarity Using Embeddings. In *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology (STIL 2021)*, pages 121-132, Online. ACL.

Souza, F., Nogueira, R. and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the 2020 Brazilian Conference on Intelligent Systems (BRACIS 2020)*, pages 403–417, Online. Springer.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage”. In *Proceedings of the Sect. on Surv. Research*, pages 354–359, ERIC.

Yepez, J., Tavares, B., Peres, F., and Becker, K. (2024). Na Batida do Funk: Modelagem de Tópicos Combinando LLM, Engenharia de Prompt e BERTopic. *In Anais do XXXIX Simpósio Brasileiro de Bancos de Dados* (SBBD 2024), pages 613–625, Florianópolis, SBC.