

Algoritmo inteligente para otimização de buscas em ferramentas de e-commerce

Wallacy Sebastian Aparecido Jeronimo de Almeida¹,
Helen Cristina de Mattos Senefonte¹

¹Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

wallacy.sebastian@uel.br, helen@uel.br

Abstract. Enterprise Resource Planning (ERP) system was created in 1990s, and since their creation, the use of them has enhanced the productivity in many organizations, ranging from small to large companies, even those with a vast amount of data to manage, and there is a constant growth in their adoption. Related to this, many retail's entrepreneurs take advantage from ERPs to manage and increase their sales through online platforms, such as e-commerce websites and marketplaces. However, to be successful in their sales, the product information needs to be shown properly and clear to the end user. From this perspective, this study aims to provide an automated solution to improve and rebuild product names, enhancing the search for product categories at Mercado Livre marketplace.

Resumo. A utilização de sistemas ERPs (Enterprise Resource Planning), desde a sua criação, tem favorecido o aumento da produtividade nas empresas, desde microempresas até empresas de grande porte que possuem um grande número de dados para gerir, e há um aumento constante em sua busca. Atrelado a isso, muitos empresários do setor de varejo recorreram a esta solução para gerir melhor as suas vendas e aumentar a comercialização de seus produtos através de plataformas online, como e-commerces e marketplaces. No entanto, para que suas vendas sejam bem-sucedidas, é necessário que as informações dos produtos sejam claras e adequadas ao expô-las para o usuário final. Nesta perspectiva, este estudo tem o objetivo de proporcionar uma solução automatizada para a correção e alteração de nomes de produtos, visando facilitar a busca de categorias de produtos no marketplace Mercado Livre.

1. Introdução

A Inteligência Artificial (IA) ganhou importância neste último ano perante o público em geral, através de produtos como ChatGPT da OpenAI e Bard do Google. O surgimento de produtos como estes são resultados do incentivo dado pelas empresas das maiores áreas econômicas do mundo, como as áreas da saúde, educação e varejo, para que reduzissem custos e melhorassem o trabalho remoto durante a pandemia. Após este período, ainda há um crescente uso de IAs no dia-a-dia das pessoas, não só através dos produtos mencionados como em outros produtos do mercado, para reconhecimento de imagens e vídeos, processamento

de textos, automatização de tarefas, análises no mercado financeiro, dentre muitas outras utilizações [5].

Quando observamos mais de perto a área do varejo, vemos o quanto a pandemia influenciou no crescimento dos e-commerces (termo usado para definir um comércio eletrônico). Durante este período, as ruas estavam com circulação restrita, os estabelecimentos possuíam limite de pessoas e muitas empresas do varejo necessitaram de uma mudança estratégica de venda, levando estas a vender pela Internet, seja através de marketplaces - modelo de negócio que possibilita a venda através de uma plataforma utilizada por vários lojistas -, de lojas virtuais, de redes sociais ou de aplicativos como WhatsApp [3]. E com o aumento do uso de e-commerces, naturalmente aumentaram também os problemas enfrentados pelos lojistas e pelos usuários que compram.

Desde que os sistemas ERP (Enterprise Resource Planning, ou Sistema de Gestão Integrado) surgiram na década de 90, as empresas têm recebido bem a solução para gerir os negócios [1], e com os e-commerces não foi diferente. Muitos e-commerces recorreram aos ERPs para gerenciar seus negócios, buscando não só ferramentas básicas (como controle de estoque, preços, entre outros) para os seus produtos, mas também uma forma de potencializar suas vendas através de integrações com outros sistemas [8]. Algumas das integrações de interesse foram os marketplaces: a empresa coloca seu produto à venda em seu e-commerce, e ao mesmo tempo em algum marketplace, sendo da responsabilidade do ERP gerenciar automaticamente os estoques das duas plataformas.

Junto às necessidades de automação devido ao grande volume de produtos de algumas empresas, surgiu também a necessidade de ferramentas que fizessem o seu produto ter mais visibilidade e praticidade ao integrá-los em marketplaces. Este estudo será focado na empresa MKX Lojas Virtuais - a qual desenvolve e-commerces para outras empresas [12] - e visa facilitar a correção e alteração de nomes de produtos, baseado em seus nomes e descrições atuais, para que a integração com o seu ERP e o marketplace Mercado Livre [6] obtenha aumento na precisão da busca de categorias de produtos do Mercado Livre.

As próximas seções descrevem como este estudo acontecerá e quais recursos serão utilizados. A seção 2, que vem a seguir, apresenta uma série de métodos para alcançar o objetivo proposto neste estudo, como a Similaridade de Cosseno e os algoritmos word2vec, TF-IDF e BERT, para previsão de palavras. Já a seção 3 apresenta os objetivos buscados através deste trabalho. E a seção 4 mostra como serão utilizados os algoritmos descritos na seção 2, para atingir os objetivos propostos.

2. Fundamentação Teórico-Metodológica e Estado da Arte

Para que o estudo seja realizado, será necessário uma pesquisa aprofundada em alguns modelos para Processamento de Linguagem Natural (NLP - Natural Language Processing), sendo esta uma potencial solução para problemas que exigem a aquisição e transformação de conhecimento através de inteligência artificial [4].

Uma maneira de representar textos é o método Word Embedding, ou Word Vector, que representa textos em vetores de palavras, atribuindo valores numéricos

para cada palavra de acordo com o significado dela [11]. Como mostra a Figura 1, existem diversos algoritmos que se utilizam desta representação, e dentre eles foram escolhidos um de cada tipo para este estudo, descritos brevemente a seguir.

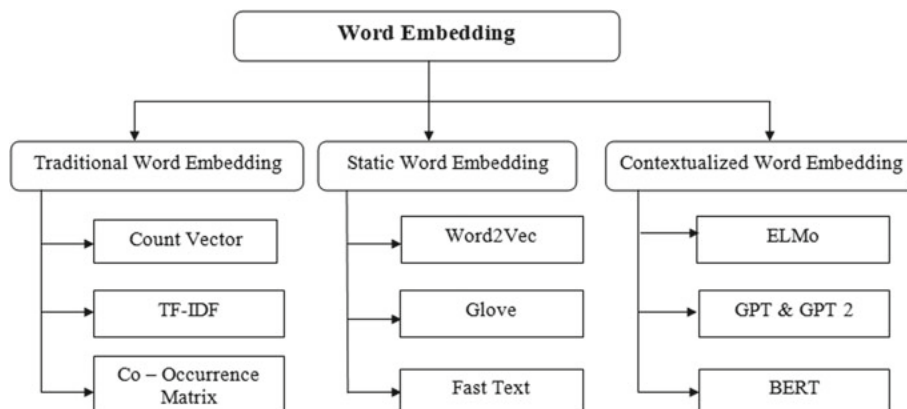


Figura 1. Imagem retirada do artigo A Review on Word Embedding Techniques for Text Classification, Birunda et. al: Tipos de Word Embedding [10]

2.1. word2vec

Foi desenvolvido pela Google em 2013, utiliza como métrica a similaridade de cossenos e possui duas variantes: CBOW (Continuous Bag of Words) e Skip-Gram. Na Figura 2, vemos que a variante CBOW tem como entrada as palavras ao redor da palavra que se quer prever, enquanto a variante Skip-gram recebe como entrada uma palavra e prevê as palavras ao redor desta.

A Similaridade de Cosseno (Cosine Similarity) é uma métrica amplamente usada para identificar a similaridade entre dois textos, transformando estes textos em vetores e obtendo um valor que varia entre -1 e 1, onde -1 significa que um texto é totalmente oposto ao outro, e 1 significa que um texto é totalmente similar ao outro. Pode-se obter também o valor 0, onde não há similaridade [4]. Dado que os nomes dos produtos que serão utilizados neste estudo podem ser substituídos por nomes mais adequados para busca, a similaridade de cosseno pode ser útil não só como métrica mas também na verificação de palavras semelhantes ao substituir palavras escritas incorretamente por palavras gramaticalmente corretas.

2.2. TF-IDF

É um método relativamente antigo, mas muito utilizado e simples de ser implementado. Do ponto de vista matemático, é o produto da frequência que uma palavra aparece em um documento pela frequência inversa das aparições desta palavra nos documentos. O resultado é o quão relevante esta palavra é em um determinado documento [9].

2.3. BERT

Apresentado pela Google em 2019, BERT (Bidirectional Encoder Representations from Transformers) foca em seu pré-treinamento, onde recebe uma entrada, esconde palavras aleatórias e o objetivo é tentar prever a palavra original da palavra escondida baseando-se somente em seu contexto [2].

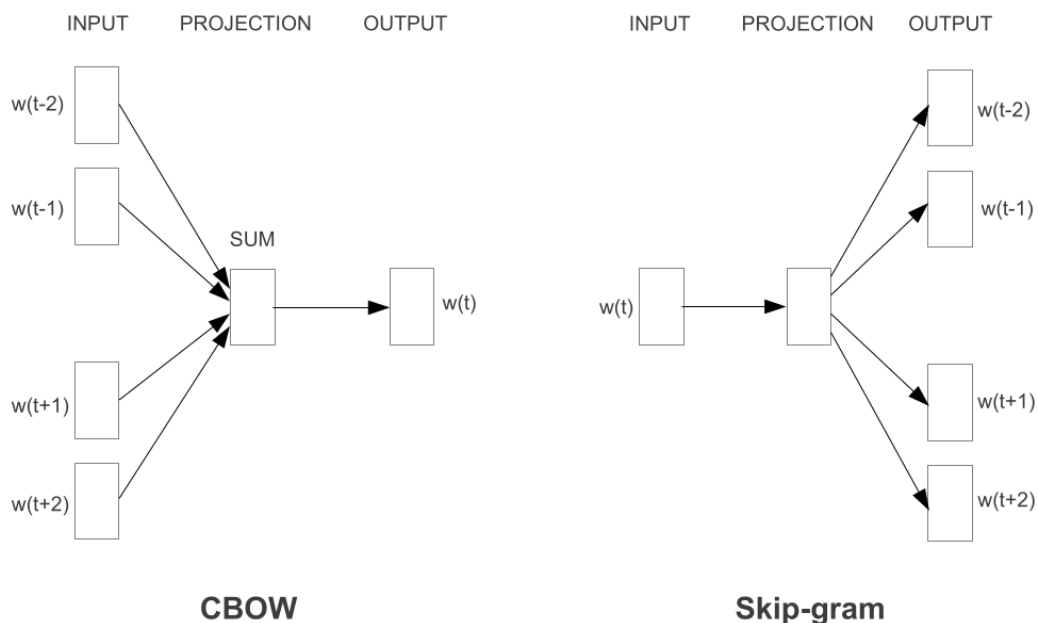


Figura 2. Imagem retirada do artigo Efficient Estimation of Word Representations in Vector Space, Mikolov et. al: Comparação entre CBOW e Skip-gram [7]

3. Objetivos

O objetivo deste estudo é otimizar a precisão da busca em ferramentas de integração em e-commerces, tendo como foco a integração MKX Lojas Virtuais e Mercado Livre.

Os objetivos específicos são:

- Estudar os algoritmos propostos (word2vec, TF-IDF e BERT), bem como suas restrições para lidar com textos de tamanhos variados;
- Propor estratégias para reescrever nomes de produtos com foco na otimização da busca de categorias no Mercado Livre;
- Implementar soluções e testar o comportamento de cada algoritmo diante de diferentes nomes e tipos de produtos;
- Realizar comparações entre os algoritmos propostos;
- Divulgar os resultados obtidos com os experimentos realizados através das soluções criadas, bem como as estratégias e algoritmos propostos.

4. Procedimentos metodológicos/Métodos e técnicas

Para entender sobre como os algoritmos de NLP funcionam, o livro Natural Language Processing [4] e os artigos BERT: Pre-training of deep bidirectional transformers for language understanding [2] e Efficient estimation of word representations in vector space [7] propiciarão a base necessária para o entendimento.

Os dados utilizados, como nomes e descrição de produtos, serão restritos e provenientes da base de dados da empresa MKX Lojas Virtuais [12], sem expor dados adicionais desnecessários a este estudo.

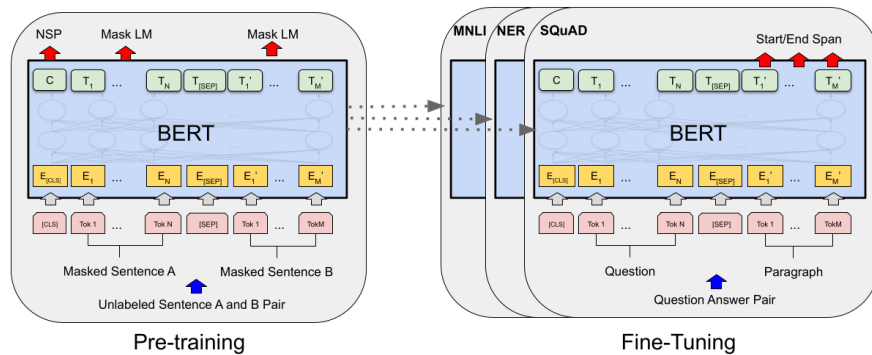


Figura 3. Imagem retirada do artigo BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et. al: Procedimentos realizados no BERT para treiná-lo e ajustá-lo [2]

Estes dados irão passar por uma análise exploratória, a qual servirá para o entendimento das particularidades de cada grupo de produtos e de cada loja, e além disso incluirá a normalização dos dados.

Após a análise, os algoritmos escolhidos irão receber estes dados como treinamento e teste, sendo 80% dos dados para treinamento e 20% para teste.

Ao final, o algoritmo será incrementado para atingir o objetivo deste estudo, e então será feita uma análise dos resultados.

5. Cronograma de Execução

Atividades:

1. Revisão bibliográfica dos conceitos em relação aos algoritmos de NLP;
2. Análise exploratória dos dados obtidos;
3. Treinamentos e testes preliminares dos algoritmos;
4. Desenvolvimento da ferramenta para reescrever os nomes dos produtos;
5. Análises e discussões preliminares;
6. Entrega da versão preliminar;
7. Revisão dos dados utilizados e realização de nova análise exploratória;
8. Treinamentos e testes dos algoritmos;
9. Desenvolvimento da ferramenta para atingir os objetivos propostos;
10. Análise e discussão final;
11. Divulgação dos resultados;
12. Entrega da versão final;

6. Contribuições e/ou Resultados esperados

Espera-se que o desenvolvimento deste estudo contribua para o aprofundamento do conhecimento em relação aos algoritmos utilizados, bem como para as etapas de obtenção do resultado esperado.

As conclusões obtidas com este estudo devem ser utilizadas para melhorar a solução proposta e incluir outras finalidades de uso com base nesta solução.

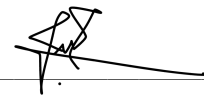
A etapa final do trabalho contempla uma ferramenta de busca que proporcionará boa precisão para a recomendação de categorias do Mercado Livre adequadas para cada produto buscado.

Tabela 1. Cronograma de Execução

	set	out	nov	dez	jan	fev	mar	abr
Atividade 1	•	•	•	•				
Atividade 2		•						
Atividade 3		•						
Atividade 4		•	•					
Atividade 5			•					
Atividade 6			•					
Atividade 7			•	•				
Atividade 8				•	•			
Atividade 9					•	•	•	
Atividade 10								•
Atividade 11								•
Atividade 12								•

7. Espaço para assinaturas

Londrina, 18 de setembro de 2023.



Aluno

Orientador

Referências

- [1] Majed Al-Mashari. Enterprise resource planning (erp) systems: a research agenda. *Industrial Management & Data Systems*, 102(3):165–170, Jan 2002.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [3] E-commerce Brasil. Evolução do e-commerce: cinco fatos para entender melhor o mercado. <https://www.ecommercebrasil.com.br/artigos/evolucao-do-e-commerce>. [Online; acessado em 07-Setembro-2023].
- [4] Jacob Eisenstein. *Natural Language Processing*. The MIT Press, October 2019.
- [5] World Economic Forum. COVID-19 increased the use of AI. Here’s why it’s here to stay. <https://www.weforum.org/agenda/2021/02/covid-19-increased-use-of-ai-here-s-why-its-here-to-stay/>. [Online; acessado em 07-Setembro-2023].
- [6] Mercado Livre. Documentação do Mercado Livre. https://developers.mercadolivre.com.br/pt_br/guia-para-produtos. [Online; acessado em 15-Setembro-2023].
- [7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

- [8] Oracle. Your complete guide to modern erp. <https://www.oracle.com/webfolder/s/assets/ebook/modern-erp/index.html>, 2017. [Online; acessado em 12-Setembro-2023].
- [9] Juan Ramos. Using tf-idf to determine word relevance in document queries, 01 2003.
- [10] S. Selva Birunda and R. Kanniga Devi. A review on word embedding techniques for text classification. In Jennifer S. Raj, Abdullah M. Ilyasu, Robert Bestak, and Zubair A. Baig, editors, Innovative Data Communication Technologies and Application, pages 267–281, Singapore, 2021. Springer Singapore.
- [11] Turing. A Guide on Word Embeddings in NLP. <https://www.turing.com/kb/guide-on-word-embeddings-in-nlp>. [Online; acessado em 08-Setembro-2023].
- [12] MKX Lojas Virtuais. Sobre a MKX e-commerce - Desenvolvimento de Lojas Virtuais e Sites. <https://www.mkx.com.br/empresa/>. [Online; acessado em 15-Setembro-2023].