Melhorando a Experiência do Cliente Online: Um Estudo de Caso com Utilização da Medida de Similaridade de Cosseno

Fabrícia de Jesus Santos[1], Layla Joana Santos[1], Methanias Colaço Júnior[1,2]

[1]Departamento de Sistemas de Informação – Universidade Federal de Sergipe (UFS) Itabaiana – SE – Brazil

[2]Programa de Pós Graduação em Ciências da Computação (PROCC) – Universidade Federal de Sergipe (UFS) – São Cristóvão – SE – Brasil

{fabriciacooper,layla.joana}@gmail.com, mjrse@hotmail.com

Abstract. *E-commerce has become increasingly popular, but users with writing problems may face difficulties when searching for products. Text mining can help improve the experience and deliver better results for these users. This article presents a case study that analyzed the performance of the cosine similarity algorithm in product search, especially for users with low writing skills. The results showed a precision of 77,1%, indicating that the algorithm is capable of correctly classifying most relevant products.*

Resumo. *O Comércio eletrônico tem se popularizado cada vez mais, mas usuários com problemas de escrita podem enfrentar dificuldades na busca de produtos. A mineração de texto pode ajudar a melhorar a experiência e devolver melhores resultados para esses usuários. Este artigo apresenta um estudo de caso que analisou o desempenho do algoritmo de similaridade do cosseno, na busca de produtos, especialmente para usuários com baixo nível de escrita. Os resultados apresentaram uma precisão de 77,1%, o que indica que o algoritmo é capaz de classificar corretamente a maioria dos produtos relevantes.*

1. Introdução

Com o crescente uso das compras *online*, a busca por produtos em aplicativos de comércio eletrônico se tornou uma atividade comum para muitos consumidores (CNDL, 2021). No entanto, para aqueles com dificuldades na escrita, essa atividade pode ser desafiadora.

A dificuldade de escrita varia amplamente entre os brasileiros devido a diversos fatores, incluindo nível de educação, acesso à educação de qualidade, contexto socioeconômico e regional. De acordo com avaliações nacionais e internacionais, como o Programa Internacional de Avaliações de Estudantes (PISA), muitos estudantes brasileiros em 2018 apresentou um desempenho abaixo da média em competência de leitura e escrita (PISA, 2018). Isso pode ser atribuído a desafios no sistema educacional, falta de recursos em algumas regiões, além de questões socioeconômicas.

De acordo com Gupta & Tomar (2021), a inteligência artificial (IA) tem se tornado uma ferramenta importante no comércio eletrônico, permitindo que empresas

ofereçam uma experiência de compra personalizada e precisa (Monteiro et al., 2022). No entanto, para usuários com baixo nível de escrita a mineração de texto também pode ser uma solução útil, ajudando a melhorar a experiência do usuário na busca por produtos em aplicativos de compras *online*.

Neste artigo, vamos analisar o desempenho do algoritmo de similaridade do cosseno na busca por produtos em aplicativos de compra *online*. A análise será especialmente voltada para usuários com deficiência na escrita, com o objetivo de avaliar a eficácia em melhorar a experiência de compra desses usuários. Logo, serão apresentados os resultados dessa análise.

2. Métricas de Qualidade

Para avaliar a efetividade do algoritmo, foram utilizadas as métricas: acurácia, revocação, precisão e medida-F1. Sendo assim, essas medidas são mensuradas a partir das seguintes frequências:

- True Positive (TP): Total de instâncias de produtos presentes na base que foram retornados corretamente entre os 10 primeiros produtos;
- True Negative (TN): Total de instâncias de produtos que não estão presentes na base e não são retornados entre os 10 primeiros produtos;
- False Positive (FP): Total de instâncias de produtos que foram retornados produtos aleatórios entre os 10 primeiros produtos;
- False Negative (FN): Total de instâncias de produtos que não são retornados corretamente nos 10 primeiros.

3.1. Acurácia

A acurácia representa o percentual de instâncias que foram retornadas corretamente, sendo definida por:

$$acurracy = \frac{TP+TN}{TP+TN+FP+FN}$$

3.2. Revocação

A revocação, também conhecida como a taxa de verdadeiros positivos, sensibilidade ou cobertura real de amostragem positiva, é o percentual de instâncias positivas que foram retornados corretamente:

$$recall = \frac{TP}{TP+FN}$$

3.3. Precisão

A precisão é a razão entre as instâncias retornadas como verdadeiro positivo e todas as instâncias retornadas como positivas:

$$precision = \frac{TP}{TP+FP}$$

3.4. Medida-F1

A mediada-F1 é a métrica que combina dois indicadores de desempenho, sendo a expressão da média harmônica da precisão e da revocação:

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

3. Trabalhos Relacionados

Foram realizadas pesquisas nas plataformas IEEE Xplore, Google Scholar e ScienceDirect nos quais foram encontrados poucos trabalhos relacionados com o tema. No entanto, no Research Gate, foi possível encontrar alguns estudos mais próximo do tema.

O trabalho de Sintia et al. (2021) relata a necessidade de aprimorar a precisão do processo de busca de codificação de produtos do aplicativo SiPaGa, uma vez que, verificou-se erros frequentes devido à imprecisão na seleção de códigos das mercadorias. Para solucionar isso, foi proposta uma solução utilizando similaridade do cosseno e TF-IDF, visando melhorar a exatidão da busca. O estudo é baseado em uma amostra de 14.416 dados de codificação de produtos obtidos do banco de dados do aplicativo SiPaGa.

Tem-se também o trabalho de Fontes (2022), que avaliou a eficácia de um classificador de notas fiscais eletrônicas baseado na mineração de textos desestruturados dessas notas para combater a corrupção. O classificador utilizou vários algoritmos, incluindo o algoritmo de cosseno. O objetivo desse trabalho era automatizar a classificação e subtotalização das notas fiscais considerando códigos e descrições únicas, importantes para identificação completa dos produtos adquiridos.

4. Metodologia

A metodologia adotada para o trabalho envolveu, inicialmente, uma pesquisa exploratória para análise da literatura, realizada por meio de um levantamento bibliográfico filtrando os principais conceitos relacionados ao tema.

Posteriormente, foi realizado um estudo de caso, que consistiu em investigar o desempenho do algoritmo de cosseno, sendo criado um ambiente para testes utilizando uma base de dados de produtos utilizados em lojas.

Desta forma, o trabalho seguiu as seguintes etapas: a) Definição e Planejamento do Estudo de Caso; b) Operação do Estudo de Caso; c) Resultados; d) Conclusão.

5. Definição e Planejamento do Estudo de Caso

Nesta e na próxima seção, o estudo de caso é apresentado. Esta seção tem como foco a definição do objetivo e o planejamento do referido estudo.

5.1. Definição do Objetivo

O objetivo deste estudo é avaliar o algoritmo de similaridade do cosseno para fornecer aos usuários, especialmente aqueles com habilidades de escrita deficientes, uma melhor experiência ao buscar produtos em plataformas de comércio eletrônico.

5.2. Planejamento

Para guiar o estudo de caso, foi elaborado a seguinte questão principal, cuja resposta visa cumprir o objetivo: o algoritmo de similaridade do cosseno é capaz de melhorar os resultados de busca de produtos em plataformas de comércio eletrônico para usuários com deficiência na escrita?

Serão utilizados cinco métricas para avaliar estas questões: (1) Acurácia; (2) Revocação; (3) Precisão; (4) Medida-F1.

5.2.1. Conjunto de Dados

Para compor ao conjunto de dados para o estudo de caso, foram escolhidos os produtos que são vendidos no site da Amazon Markteplace Brasil. E para isso, foi utilizada a ferramenta Octoparse, que possibilitou definir critérios de extração para coletar as informações valiosas, como preços, descrições e nomes de produtos. No contexto desse estudo de caso, a Octoparse foi empregada para extrair, especificamente, informações dos produtos. Dessa forma, foram coletados cerca de 15 mil produtos de diversas categorias, incluindo perfumaria, farmácia, eletrônicos, roupas, smartphones e material de construção.

Para tal, foram considerados os atributos: nome, descrição e preço do produto.

5.2.2. Instrumentação

Para instrumentação dos testes realizados, foram utilizadas algumas ferramentas e recursos:

- Octoparse para realizar web scraping do site da Amazon e obter os produtos;
- SQL Server como sistema gerenciador de banco de dados;
- Linguagem de programação C# para desenvolver a API de busca de produtos e integrá-la ao banco de dados.

Adicionalmente, para execução dos testes, utilizou-se o Postman para enviar as requisições. Vale acrescentar, também, que a API continha a lógica do algoritmo de similaridade para permitir a análise do estudo de caso.

6. Operação do Estudo de Caso

6.1. Preparação

Em síntese, foi preparado o ambiente para a realização do estudo de caso, ou seja, o carregamento dos produtos que foram obtidos pelo site da Amazon no SQL Server e a criação de uma API contendo o algoritmo para retornar os produtos de acordo com o que é desejado pelo usuário.

6.1.1. Pré-Processamento dos dados

Durante o processo de coleta de dados do site Amazon Martekplace Brasil, foi identificado que os dados coletados apresentavam um número de caracteres muito elevado. Esse fato tornava difícil a comparação de uma única palavra com o nome de um produto grande, o que comprometia a eficácia do algoritmo de similaridade do cosseno.

Para amenizar esse problema, foi necessário um pré-processamento dos textos para melhorar a eficácia do algoritmo. Esses pré-processamento consistiu na remoção de *stopwords*, espaços duplos, símbolos, números e acentos. Dessa forma, foi possível tornar os dados mais uniformes e tratáveis para o algoritmo.

Logo abaixo, temos uma exemplificação de como funcionou esse processo:

- Exemplo 01 para produto: Novo Kindle 11ª Geração (lançamento 2022) – Mais leve, com resolução de 300 ppi e o dobro de armazenamento - Cor Preta

  - Com processamento: NOVO KINDLE GERACAO LANCAMENTO LEVE RESOLUCAO PPI DOBRO ARMAZENAMENTO COR PRETA

6.1.2. Vetorização

Vetorização é um processo fundamental para representar textos em formato numérico, permitindo calcular a similaridade entre eles usando o algoritmo de similaridade do cosseno. Esse processo transforma os textos em vetores numéricos, onde cada elemento do vetor representa uma característica do texto. Os dados são transformados em vetores e a similaridade é calculada através do cosseno do ângulo entre os vetores.

Para o estudo de caso, foi necessário tratar o problema de vetores de tamanhos diferentes. Para solucionar essa questão, os vetores menores foram preenchidos com zeros, permitindo que todos os vetores tivessem o mesmo tamanho e facilitando, dessa forma, a aplicação do algoritmo para calcular a similaridade.

Conforme é ilustrado na Tabela 1, tem-se um exemplo de como fica o vetor com dimensões diferentes e a frequência de palavras. Neste, cada palavra única no texto é representada por uma dimensão nos vetores.

- Produto 1: Smartphone Xiaomi Redmi Note 11S 6GB RAM 128GB Graphite Gray (Cinza)
- Palavra pesquisada: Redmi Note

Tabela 1 – Exemplo de vetorização

smartphone	xiaomi	redmi	note	11s	6gb	ram	128gb	graphite	gray	cinza
1	1	1	1	1	1	1	1	1	1	1
0	0	1	1	0	0	0	0	0	0	0

6.2. Execução

Consistiu na realização do processo classificatório nos dados da base coletada. Para cada classificação realizada, verificava-se, se os produtos retornados eram de fato o produto desejado, garantindo a precisão e a confiabilidade dos resultados obtidos.

Ao término da execução, os resultados das classificações foram obtidos e os indicadores foram gerados referente às métricas previamente definidas.

7. Resultados

Durante o processo de desenvolvimento do algoritmo, foram realizados diversos testes com objetivo de encontrar o limiar ideal para garantir uma boa precisão na identificação de produtos semelhantes.

Com o propósito de otimizar a experiência do usuário, foi estabelecido um conjunto inicial de dez itens como os principais resultados a serem retornados, visando permitir que o usuário localize facilmente o produto desejado. Diferenciando-se de abordagens de classificação convencionais que fixam uma classe específica, esta abordagem visa antecipar a intenção do usuário.

Para determinar o limiar ideal, um processo sistemático foi empregado, iniciando com a análise dos maiores valores até os menores, seguido pelo exame dos menores para os maiores. Neste percurso, foram realizadas diversas avaliações manuais para ajustar o valor do limiar e tornou-se evidente que o limiar de 0,6 se revelou o mais eficaz, uma vez que proporcionou resultados mais alinhados.

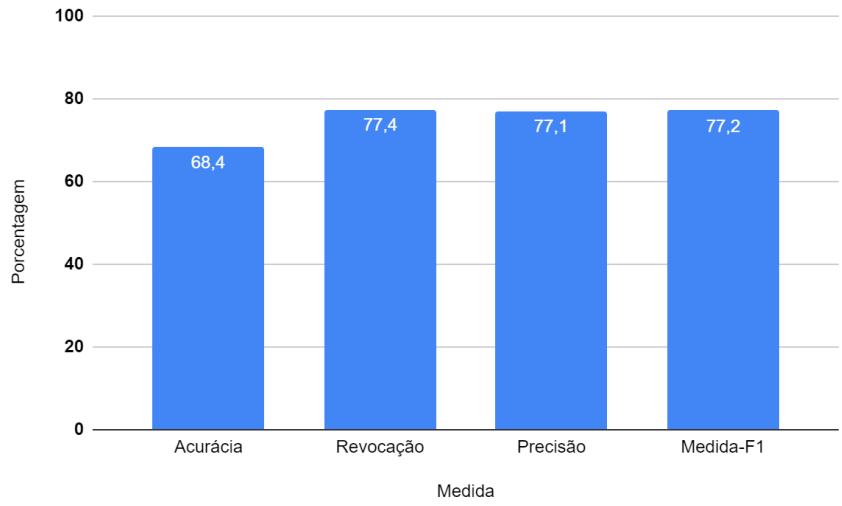

Figura 1. Resultado das medidas do algoritmo de similaridade do cosseno

Conforme ilustrado na Figura 1, os resultados mostraram que acurácia do algoritmo foi de 0,684, o que significa que ele classificou corretamente 68,4% dos produtos relevantes. Além disso, a revocação e a precisão do algoritmo foram 0,774 e 0,771 respectivamente, indicando que ele recuperou 77,4% dos produtos relevantes e que 77,1% dos produtos que ele classificou como relevantes eram de fato relevantes.

A medida-F1, que é uma métrica que combina a precisão e a revocação, foi calculada em 0,772. Isso significa que o algoritmo obteve um bom equilíbrio entre a precisão e a revocação.

Os resultados mostraram um bom desempenho do algoritmo na busca de produtos semelhantes. No entanto, é importante considerar que a eficácia da busca pode variar com os dados, critérios de relevância e configurações do algoritmo. O teste incluiu textos com escritas corretas e incorretas, e o algoritmo demonstrou eficiência ao lidar com erros ortográficos. Isso é benéfico para usuários com dificuldade de escrita, permitindo que encontrem produtos desejados sem preocupações com a correção gramatical.

8. Conclusão

Com base nos resultados, o algoritmo de similaridade do cosseno se mostrou promissor para a busca de produtos *e-commerce*, alcançando uma acurácia de 68,4%. Sua medida-F1 equilibrada entre a precisão e revocação, torna-o uma ferramenta valiosa para aprimorar a experiência do usuário e impulsionar as vendas. É necessário avaliar sua eficácia em diferentes contextos, considerando usuários com diferentes níveis de escrita.

Em resumo, no experimento realizado, o algoritmo de similaridade do cosseno se mostrou eficiente e útil para melhorar a busca de produtos. Seus resultados positivos indicam uma boa capacidade de classificar corretamente produtos relevantes e recuperar uma porcentagem significativa de produtos.

Embora o trabalho ainda esteja no início da fase da Avaliação, a natureza do trabalho permite que sejam apontadas ameaças à validade da solução proposta. Seriam necessários outros Estudos de Casos, replicando a utilização da solução, para confirmar os benefícios esperados.

Por fim, como trabalhos futuros, será realizada a integração com a biblioteca de correções ortográficas e verificar a frequência de uso dos termos, visando obter resultados mais precisos. Isso possibilitará avaliar se existirá diferença entre os resultados de ambas as abordagens.

9. Referências

Confederação Nacional dos Dirigentes Lojistas [CNDL]. 2021. 91% dos internautas realizaram compras pela internet nos últimos 12 meses. Disponível em:<https://materiais.cndl.org.br/pesquisa-consumo-online-no-brasil>. Acesso em: 28 fev. 2023.

Gupta, A., & Tomar, K. (2021). Enhancing Marketing Strategies and Analytics Through ArtificialIntelligence. 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), 174–179. https://doi.org/10.1109/ICCAKM50778.2021.9357763

Monteiro, Mariana & Azevedo, Ana & Pereira, Inês. (2022). A aplicação da inteligência artificial no comércio eletrónico: cross-sell e up-sell. Cadernos de Investigação do Mestrado em Negócio Eletrónico CIMNE. 2. 10.56002/ceos.0063_cimne_1_2.

PISA. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Brasil no Pisa 2018 (Programme for Internacional Student Assessment) [recurso eletrônico].-Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020. 185 p. :il. ISBN 978-65-5801-039-5. Disponível em:< https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/relatorio_brasil_no_pisa_2018.pdf>. Acesso em: 12 ago. 2023.

FONTES, Raphael Silva. Avaliação experimental de um classificador para apoiar a detecção de fraudes em compras públicas. 2022. 68 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Sergipe, São Cristóvão, 2022.

Sintia, Sintia & Defit, Sarjon & Nurcahyo, Gunadi. (2021). Product Codefication Accuracy With Cosine Similarity And Weighted Term Frequency And Inverse Document Frequency (TF-IDF). Journal of Applied Engineering and Technological Science (JAETS). 2. 62-69. 10.37385/jaets.v2i2.210.