10° EBBC
Encontro Brasileiro de Bibliometria e Cientometria

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

RECOMENDAÇÃO AUTOMÁTICA DE PERIÓDICOS CIENTÍFICOS POR SIMILARIDADE SEMÂNTICA EM RESUMOS DE ARTIGOS

Denise Fukumi Tsunoda

https://orcid.org/0000-0002-5663-4534
dtsunoda@ufpr.br
Universidade Federal do Paraná (UFPR)

Alex Sebastião Constâncio

https://orcid.org/0000-0001-8725-4481
alex.constancio@ufpr.br
Universidade Federal do Paraná (UFPR)

Fábio Castro Gouveia

https://orcid.org/0000-0002-0082-2392
fabiogouveia@ibict.br
Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Henrique Denes Hilgenberg Fernandes

https://orcid.org/0000-0003-1537-8762
denes@ibict.br
Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Resumo:
Este estudo propõe um sistema de recomendação automática de periódicos científicos fundamentado em similaridade semântica. Utilizando embeddings e a base BRAPCI, a pesquisa processou 47 mil resumos para identificar títulos compatíveis com novos manuscritos. Os resultados demonstram que o periódico original de publicação figurou entre as dez principais sugestões em 56,8% dos casos, evidenciando a eficácia da captura de relações temáticas. A abordagem visa reduzir o risco de rejeições por incompatibilidade de escopo (desk rejection), oferecendo uma ferramenta escalável para auxiliar pesquisadores no processo de comunicação científica.

Palavras-Chave: Comunicação Científica. Sistemas de Recomendação. Similaridade Semântica. Informação Científica.

EIXO TEMÁTICO:
Técnicas, Ferramentas e Infraestruturas

MODALIDADE:
Resumo expandido

DOI: 10.22477/x.ebbc.1034

COMO CITAR ESTE ARTIGO:

TSUNODA, Denise Fukumi; CONSTÂNCIO, Alex Sebastião; GOUVEIA, Fábio Castro; FERNANDES, Henrique Denes Hilgenberg. Recomendação automática de periódicos científicos por similaridade semântica em resumos de artigos. *In*: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1034. Disponível em: https://doi.org/10.22477/x.ebbc.1034.

1 INTRODUÇÃO

A escolha do periódico adequado para submissão de um artigo científico é uma etapa importante no processo de comunicação científica. Essa decisão influencia não apenas a visibilidade da pesquisa, mas também sua adequação temática, o alcance do público leitor e a potencial repercussão do trabalho. No entanto, identificar periódicos compatíveis com o conteúdo de um manuscrito nem sempre é uma tarefa simples, especialmente diante do grande número de periódicos disponíveis e da crescente expansão da produção científica.

Nos últimos anos, o avanço das técnicas de processamento de linguagem natural e de representação semântica de textos tem ampliado as possibilidades de análise automatizada de documentos científicos. Entre essas abordagens, destacam-se métodos baseados em *embeddings*, que permitem representar palavras, frases e textos inteiros como vetores e comparar conteúdos com base em sua similaridade semântica. Esse tipo de técnica tem sido utilizado em diferentes aplicações relacionadas à recuperação da informação, incluindo sistemas de recomendação.

Nesse contexto, este trabalho avalia uma aplicação que faz uso de similaridade semântica entre resumos de artigos científicos para recomendação automática de periódicos. A proposta baseia-se na comparação entre o resumo de um artigo submetido e resumos de artigos previamente publicados, permitindo identificar periódicos potencialmente adequados para submissão. O estudo analisa uma abordagem baseada em *embeddings* semânticos aplicada a uma base de dados da área de Ciência da Informação.

2 FUNDAMENTAÇÃO TEÓRICA

Com o crescimento da produção científica e a ampliação do número de periódi-

cos disponíveis, o processo de escolha de um periódico para submissão de um manuscrito visando dar maior visibilidade à pesquisa e o melhor posicionamento para acessar o potencial público leitor dentro de uma determinada área de conhecimento tornou-se mais complexo, estimulando o desenvolvimento de ferramentas capazes de apoiar a identificação de periódicos adequados. Reforça-se ainda que um estudo de 2015 (Trumbore; Carr; Mikaloff Fletcher, 2015) apresenta que de 25 a 30% das rejeições que ocorrem antes da revisão por pares (chamada de *desk rejection*) é motivada por escopo incompatível com os interesses do periódico.

Sistemas de recomendação têm sido amplamente investigados em diferentes domínios, podendo ser estruturados como problemas de ranqueamento de itens com base em sua relevância para o usuário (Ricci; Rokach; Shapira, 2018). Entre os desafios clássicos destacam-se questões como a inicialização a frio (cold-start), relacionadas à ausência de dados prévios sobre usuários ou itens (Roy; Dutta, 2022; Zangerle; Bauer, 2022). No caso da comunicação científica, abordagens baseadas em conteúdo têm sido frequentemente utilizadas, uma vez que permitem comparar diretamente características textuais de artigos científicos, como títulos, resumos e palavras-chave.

Avanços em processamento de linguagem natural, especialmente com modelos baseados em transformadores, têm ampliado a análise automatizada de textos científicos. Nesse contexto, *embeddings* permitem representar documentos como vetores semânticos, favorecendo tarefas como recuperação de informação e similaridade textual (Zhao *et al*., 2024) e o uso de similaridade semântica aplicada a dados bibliográficos amplia as possibilidades de organização e recuperação da informação científica.

Nos últimos anos, diferentes abordagens têm sido propostas para apoiar a escolha de periódicos científicos, incluindo ferramentas comerciais e acadêmicas baseadas na análise de metadados e similaridade textual. Sistemas como Elsevier Journal Finder e Springer Journal Suggester utilizam informações como título, resumo e palavras-chave para sugerir periódicos potencialmente adequados, evidenciando a relevância prática desse tipo de aplicação. No âmbito acadêmico, estudos têm explorado técnicas de recomendação baseadas em conteúdo, nas quais a similaridade entre textos científicos é utilizada como principal critério para inferir compatibilidade temática entre manuscritos e periódicos.

Mais recentemente, o avanço dos modelos de linguagem baseados em transformadores ampliou significativamente a capacidade de representação semântica de textos. Modelos como o BERT (Zhao *et al*., 2024) e suas variações, como o Sentence-BERT (Reimers; Gurevych, 2019), possibilitam a geração de *embeddings* densos que capturam relações semânticas mais profundas entre documentos. Essas abordagens têm sido amplamente utilizadas em tarefas de recuperação de informação e busca semântica (Karpukhin *et al*., 2020), superando limitações de métodos tradicionais baseados exclusivamente em frequência de termos, como o TF-IDF (Danish *et al*., 2024). Nesse contexto, a utilização de *embeddings* para recomendação de periódicos representa uma evolução natural das abordagens baseadas em conteúdo.

3 ENCAMINHAMENTOS METODOLÓGICOS

A pesquisa possui caráter exploratório e aplicado, com abordagem computacional voltada à implementação e avaliação de um sistema de recomendação automática de periódicos científicos baseado em similaridade semântica entre textos. As etapas da pesquisa estão detalhadas na sequência.

Como base de dados para o experimento foi utilizada a BRAPCI - Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (Gabriel Junior, 2026), que reúne metadados de artigos publicados em periódicos da área de Ciência da Informação. A base utilizada contém aproximadamente 47 mil resumos de artigos distribuídos em 105 periódicos científicos, incluindo informações como título, resumo, palavras-chave e periódico de publicação.

Inicialmente foi realizada a preparação e padronização dos dados textuais, com o objetivo de permitir o processamento automatizado dos resumos científicos. Os textos foram organizados e segmentados em sentenças utilizando o segmentador de frases da biblioteca *Pyhton SpaCy* (https://spacy.io/), resultando em um conjunto de aproximadamente 224 mil sentenças derivadas dos resumos originais.

Em seguida foram geradas representações vetoriais das sentenças por meio de *embeddings* semânticos, utilizando o modelo de linguagem gratuito *paraphrase-multilingual-mpnet-base-v2*, disponível na comunidade de desenvolvedores de Inteligência Artificial *Hugging Face* (https://huggingface.co/), processável pela biblioteca *SentenceTransformers*, da mesma comunidade. Os vetores gerados foram indexados utilizando a biblioteca FAISS (*Facebook AI Similarity Search*), que possibilita realizar consultas eficientes de similaridade vetorial em grandes volumes de dados. A recuperação de documentos similares foi realizada por meio da similaridade de cosseno, utilizando o método de vizinhos mais próximos (*k-nearest neighbors*).

O mecanismo de recomendação foi estruturado de forma que, ao submeter o resumo de um artigo, o sistema identifica sentenças semanticamente similares na base indexada e, a partir dessas correspondências, recupera os resumos associados e os respectivos periódicos de publicação. Os periódicos candidatos são

então ranqueados com base na frequência e no grau de similaridade dos textos recuperados.

Para avaliação do desempenho do sistema foi adotada a métrica *Top-k accuracy*, considerando diferentes cenários de recomendação (Top-1, Top-3, Top-5 e Top-10). A base de dados foi dividida em duas partições: 90% dos resumos foram utilizados para indexação e 10% para simulação de consultas, permitindo comparar os periódicos recomendados pelo sistema com os originais de publicação dos artigos analisados.

O ambiente computacional utilizado para a implementação do sistema foi baseado na linguagem Python, com uso das bibliotecas *Pandas*, *SentenceTransformers*, *SpaCy*, *FAISS* e *Matplotlib* para processamento de dados, geração de *embeddings*, recuperação de similaridade e análise dos resultados.

4 RESULTADOS

O experimento foi conduzido com base em um conjunto de aproximadamente 47 mil resumos de artigos científicos provenientes da base BRAPCI, distribuídos em 105 periódicos da área de Ciência da Informação. Após o processo de segmentação textual, os resumos resultaram em cerca de 224 mil sentenças, que foram utilizadas para geração das representações vetoriais dos textos por meio de *embeddings* semânticos gerados pelo modelo de linguagem e indexados pelo FAISS.

Os *embeddings* gerados pelo modelo paraphrase-multilingual-mpnet-base-v2 possuem dimensão vetorial de 768 e foram utilizados com normalização L2, de forma a viabilizar o uso consistente da similaridade de cosseno como medida de proximidade semântica. A escolha desse modelo se justifica por sua capacidade multilíngue, fator relevante considerando a predominância de textos em língua

portuguesa na base BRAPCI. Além disso, a utilização de *embeddings* densos permite capturar relações semânticas além da simples correspondência lexical entre termos.

Como a base de dados foi dividida em duas partes, foram utilizados os próprios resumos de artigos como entrada do sistema de recomendação. Para cada consulta, o sistema recuperou os textos semanticamente mais próximos na base indexada e identificou os periódicos associados a esses documentos.

Ao utilizar os próprios resumos de artigos como entrada para o sistema, segue-se uma abordagem de avaliação baseada em recuperação, na qual se verifica a capacidade do modelo em recuperar o periódico original de publicação a partir de seu conteúdo textual. Esse procedimento é análogo a métodos de validação amplamente utilizados em sistemas de recuperação de informação, permitindo avaliar a eficácia da representação semântica para compatibilidade temática entre documentos.

O desempenho avaliado por *Top-k accuracy* indica se o periódico original de publicação do artigo aparece entre os k primeiros periódicos recomendados. Os resultados obtidos são apresentados na Tabela 1.

Tabela 1 - Acurácias de recuperação em quatro cenários

ACURÁCIA	VOLUME (SENTENÇAS E RESUMOS)	TAXA DE ACERTO (TA)
Top-1	1.068	0,2557
Top-3	1.636	0,3919
Top-5	1.924	0,4609
Top-10	2.372	0,5683

Fonte: Dados da pesquisa (2025).

Os resultados obtidos devem ser interpretados considerando algumas limitações

do estudo. Em primeiro lugar, a cobertura da base BRAPCI está restrita à área de Ciência da Informação, o que pode limitar a generalização dos resultados para outras áreas do conhecimento. Além disso, a predominância de textos em língua portuguesa pode influenciar o desempenho dos modelos de *embeddings*, mesmo quando estes possuem capacidade multilíngue. Outro fator relevante refere-se à granularidade da análise, uma vez que a segmentação em sentenças pode introduzir variações na representação semântica dos documentos. Por fim, a escolha de um periódico para submissão não depende exclusivamente da similaridade temática, sendo também influenciada por fatores como escopo editorial, estratégias dos autores e critérios específicos de avaliação de cada periódico.

Em termos de desempenho, os resultados indicam que o modelo apresenta comportamento compatível com sistemas de recuperação semântica aplicados a textos científicos. A taxa de acerto de aproximadamente 25% na primeira posição (Top-1) e de até 56% entre as dez primeiras recomendações (Top-10) sugere que o modelo é capaz de identificar, com razoável consistência, periódicos tematicamente alinhados ao conteúdo dos manuscritos. Considerando a complexidade do processo de publicação científica, esses valores podem ser interpretados como indicativos de boa capacidade de ranqueamento em contextos de recomendação baseada em conteúdo.

Além do desempenho quantitativo, os testes realizados indicaram que a abordagem apresenta tempo de resposta reduzido e boa escalabilidade, mesmo com bases de dados relativamente extensas. Isso ocorre devido ao uso de indexação vetorial e busca aproximada por vizinhos mais próximos, que permite realizar consultas de similaridade de forma eficiente (Schäffer *et al.*, 2025).

De modo geral, os resultados demonstram que abordagens baseadas em repre-

sentações semânticas de textos podem contribuir para o desenvolvimento de ferramentas de apoio à comunicação científica, auxiliando pesquisadores na identificação de periódicos compatíveis com o conteúdo de seus manuscritos.

Embora abordagens tradicionais baseadas em frequência de termos, como o TF-IDF, possam ser utilizadas como *baseline* para comparação, tais métodos tendem a capturar apenas similaridade lexical, apresentando limitações na identificação de relações semânticas mais complexas entre textos. Nesse sentido, os resultados obtidos com *embeddings* sugerem uma vantagem potencial dessa abordagem na captura de proximidade temática. Ainda assim, a inclusão de comparações experimentais com métodos clássicos constitui uma oportunidade para trabalhos futuros.

5 CONSIDERAÇÕES FINAIS

Os resultados obtidos neste estudo indicam que a utilização de representações semânticas de textos permite identificar relações temáticas relevantes entre resumos científicos, possibilitando sugerir periódicos potencialmente adequados para submissão de manuscritos. A avaliação do sistema demonstrou que o periódico original do artigo foi recuperado em aproximadamente 25% dos casos na primeira posição e em até 56% quando consideradas as dez primeiras recomendações.

Além de contribuir para o desenvolvimento de ferramentas de apoio à comunicação científica, a abordagem apresentada evidencia o potencial das técnicas de processamento de linguagem natural para exploração e análise de bases de dados científicas. Métodos baseados em similaridade semântica podem ampliar as possibilidades de organização, recuperação e curadoria da informação científica.

Como perspectivas futuras, destacam-se a ampliação da base de dados utiliza-

da, a incorporação de outras fontes de informação científica e a investigação de abordagens híbridas que combinem análise semântica com indicadores bibliométricos. Essas iniciativas podem contribuir para o aprimoramento de sistemas de recomendação voltados à comunicação científica e à gestão da informação em ambientes acadêmicos.

Apesar das limitações identificadas, os resultados indicam que a abordagem proposta possui potencial como ferramenta de apoio à decisão no processo de submissão científica, podendo contribuir para a redução de rejeições por incompatibilidade de escopo e para o aprimoramento da comunicação científica.

AGRADECIMENTOS

Os autores agradecem a Rene Faustino Gabriel Júnior pelo apoio técnico na extração e organização dos dados da base BRAPCI, que viabilizou a condução deste estudo. Também agradecem ao apoio financeiro do CNPq, processos 407463/2023-2 e 315689/2023-4, e Faperj E-26/204.368/2024 (299940).

REFERÊNCIAS

ZHAO, Zihuai; FAN, Wenqi; LI, Jiatong; LIU, Yunqing; MEI, Xiaowei; WANG, Yiqi; WEN, Zhen; WANG, Fei; ZHAO, Xiangyu; TANG, Jiliang; LI, Qing. Recommender systems in the era of large language models (llms). **IEEE Transactions on Knowledge and Data Engineering,** v. 36, n. 11, p. 6889-6907, 2024.

KARPUKHIN, Vladimir; OĞUZ, Barlas; MIN, Sewon; LEWIS, Patrick; WU, Ledell; EDUNOV, Sergey; CHEN, Danqi; YIH, Wen-tau. **Dense passage retrieval for open-domain question answering**. *In*: PROCEEDINGS OF THE 2020 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP). 2020. p. 6769–6781. Disponível em: https://arxiv.org/abs/2004.04906. Acesso em: 27 abr. 2026.

GABRIEL JÚNIOR, René Faustino. **BRAPCI** – Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação. [S. l.: s. n.], 2026. Disponível em: https://brapci.inf.br/home. Acesso em: 05 mar. 2026.

REIMERS, Nils; GUREVYCH, Iryna. **Sentence-BERT: Sentence embeddings using Siamese BERT-networks**. *In*: PROCEEDINGS OF THE 2019 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. Hong Kong: ACL, 2019. p. 3982–3992. Disponível em: https://arxiv.org/abs/1908.10084. Acesso em: 27 abr. 2026.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha (ORGS.). **Recommender Systems Handbook**. New York, NY: Springer US, 2022.

DANISH, Syed Muhammad Hasan; HASNAIN, Syed; ASHRAF, Hamza; RUKAIYA, Rukaiya. **Comparative Analysis of BERT and TF-IDF for Textual Semantic Similarity Assessment**. *In*: 2024 26th International Multi-Topic Conference (INMIC). IEEE, 2024. p. 1-6.

SCHÄFER, Patrick et al. **Fast and Exact Similarity Search in Less than a Blink of an Eye**. *In*: Hong Kong, Hong Kong: IEEE, 19 maio 2025. Disponível em: https://ieeexplore.ieee.org/document/11112946/. Acesso em: 5 mar. 2026.

ROY, Deepjyoti; DUTTA, Mala. A systematic review and research perspective on recommender systems. **Journal of Big Data**, v. 9, n. 1, p. 59, 2022. Disponível em: https://dl.acm.org/doi/10.1145/564376.564421. Acesso em: 27 abr. 2026.

TRUMBORE, Susan; CARR, Mary-Elena; MIKALOFF-FLETCHER, Sara. Criteria for rejection of papers without review. **Global Biogeochemical Cycles**, v. 29, n. 8, p. 1123–1123, ago. 2015. Disponível em: https://doi.org/10.1002/2015GB005234. Acesso em: 01 jun. 2026.

ZANGERLE, Eva; BAUER, Christine. E**valuating recommender systems:** survey and framework. ACM computing surveys, v. 55, n. 8, p. 1-38, 2022. Disponível em: https://dl.acm.org/doi/pdf/10.1145/3556536 . Acesso em: 27 abr. 2026.