eib01065910603

Sistemas de recomendação e processamento de linguagem natural: uma revisão estruturada e tendências emergentes com o suporte de ferramentas de inteligência artificial

Denise Fukumi Tsunoda*
Patrick Fernandes Rezende Ribeiro**
Juliane de Lima Pires*
Kamilly Voitkiv Hubner*
Matheus Henrique Assumpção dos Reis***
Patrick Alves Bastos***
Roberto Rigo***

Artículo recibido:
3 de noviembre de 2025
Artículo aceptado:
16 de febrero de 2026

Resumo

Este artigo apresenta uma revisão estruturada da literatura sobre sistemas de recomendação que utilizam processamento de linguagem natural (PLN), abrangendo publicações entre 2020 e 2025. O corpus final compreendeu 240 artigos analisados integralmente após o processo de triagem e deduplicação (214 de 2020-2024 e 26 de 2025). A curadoria e a organização dos dados foram apoiadas por ferramentas digitais como Zotero, Rayyan, SciSpace, NotebookLM e Biblioshiny. Os resultados evidenciam a

* Departamento de Ciência e Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, Brasil
dtsunoda@ufpr.br julianepires@ufpr.br kamillyhubner@ufpr.br

** Programa de Pós-graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, Brasil
patrick.ribeiro@ufpr.br

*** Curso de Tecnologia em Análise e Desenvolvimento de Sistemas, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, Brasil
assumpcao@ufpr.br patrickalves@ufpr.br robertorigo@ufpr.br

predominância de técnicas de aprendizado profundo, com destaque para modelos como BERT, Word2Vec e GPT, além do uso crescente de modelos de linguagem de grande escala e grafos de conhecimento. No entanto, não foram encontrados registros na *Revista Brasileira de Informática na Educação (RBIE*) ou na biblioteca digital da Sociedade Brasileira de Computação (SBC-OpenLib), indicando que, em comparação a domínios globais, a produção científica em nichos específicos nacionais ainda é incipiente. Essa lacuna evidencia oportunidades relevantes para a transposição de técnicas avançadas de processamento de linguagem natural para domínios específicos ainda pouco explorados nacionalmente, como o contexto educacional, além de fomentar a formação de pesquisadores no uso de metodologias de revisão sistemática assistidas por inteligência artificial.

Palavras-chave: Sistemas de recomendação; Revisão estruturada; Inteligência artificial; Processamento de linguagem natural

Sistemas de recomendación y procesamiento de lenguaje natural: una revisión estructurada y tendencias emergentes con el soporte de herramientas de inteligencia artificial

Denise Fukumi Tsunoda, Patrick Fernandes Rezende Ribeiro, Juliane de Lima Pires, Kamilly Voitkiv Hubner, Matheus Henrique Assumpção dos Reis, Patrick Alves Bastos y *Roberto Rigo*

Resumen

Este artículo presenta una revisión estructurada de la literatura sobre sistemas de recomendación que recurren al procesamiento de lenguaje natural (PLN), que abarca publicaciones entre 2020 y 2025. El corpus final estuvo compuesto por 240 artículos analizados íntegramente tras la selección y deduplicación (214 correspondientes a 2020-2024 y 26 publicados en 2025). La curaduría y organización de los datos se apoyó en herramientas digitales como Zotero, Rayyan, SciSpace, NotebookLM y Biblioshiny. Los resultados evidencian el predominio de técnicas de aprendizaje profundo, con énfasis en modelos como BERT, Word2Vec y GPT, además del uso creciente de grandes modelos de lenguaje y grafos de conocimiento. Sin embargo, no se encontraron registros en la *Revista Brasileira de Informática na Educação (RBIE)* o en la biblioteca digital de la Sociedade Brasileira de Computação (SBC-OpenLib), lo que indica que, en comparación con los dominios globales, la producción científica en nichos nacionales específicos es

aún incipiente. Esta laguna evidencia oportunidades relevantes para la transposición de técnicas avanzadas de procesamiento de lenguaje natural a dominios específicos aún poco explorados a nivel nacional, como el contexto educativo, además de fomentar la formación de investigadores en el uso de metodologías de revisión sistemática asistidas por inteligencia artificial.

Palabras clave: Sistemas de recomendación; Revisión estructurada; Inteligencia artificial; Procesamiento de lenguaje natural

Recommender Systems and Natural Language Processing: A Structured Review and Emerging Trends Supported by Artificial Intelligence Tools

Denise Fukumi Tsunoda, Patrick Fernandes Rezende Ribeiro, Juliane de Lima Pires, Kamilly Voitkiv Hubner, Matheus Henrique Assumpção dos Reis, Patrick Alves Bastos and *Roberto Rigo*

Abstract

This article presents a structured literature review on recommender systems that use natural language processing (NLP), covering publications between 2020 and 2025. The final corpus included 240 fully analyzed articles after screening and deduplication (214 from 2020-2024 and 26 from 2025). Digital tools such as Zotero, Rayyan, SciSpace, NotebookLM, and Biblioshiny supported data curation and organization. The results highlight the predominance of deep learning techniques, with emphasis on models such as BERT, Word2Vec, and GPT, as well as the growing use of large language models (LLMs) and knowledge graphs. However, no records were found in the *Revista Brasileira de Informática na Educação* (RBIE) or the digital library of the Sociedade Brasileira de Computação (SBC-OpenLib), showing that, compared to global domains, scientific production in specific national niches is still incipient. This gap highlights relevant opportunities for transposing advanced natural language processing techniques to specific domains that are still underexplored at a national level, such as the educational context, in addition to fostering the training of researchers in the use of artificial intelligence-assisted systematic review methodologies.

Keywords: Recommender Systems; Structured Review; Artificial Intelligence; Natural Language Processing

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

Introdução

Os sistemas de recomendação estão sendo utilizados para proporcionar sugestões personalizadas aos usuários com base em suas necessidades, comportamentos, preferências e interações anteriores no entorno digital. Esses sistemas têm ganhado relevância em várias áreas, como o comércio eletrônico, os serviços de streaming, as redes sociais e também no campo educacional, onde podem apoiar a personalização de ambientes virtuais de aprendizagem. Ao auxiliarem os usuários na navegação em grandes volumes de dados, esses sistemas contribuem para decisões mais informadas e alinhadas aos interesses individuais.

Nesse cenário, a investigação sobre o uso do processamento de linguagem natural (PLN) em sistemas de recomendação revela-se particularmente pertinente, ao oferecer subsídios para o desenvolvimento de sistemas de recomendação mais precisos, interpretáveis e contextualizados, alinhados às demandas contemporâneas de personalização, criticidade e uso ético das tecnologias digitais na educação.

Diferentes abordagens para a construção desses sistemas podem ser adotadas. Conforme Ricci, Rokach e Shapira (2015: 3), as mais tradicionais são a filtragem colaborativa, a filtragem baseada em conteúdo e os sistemas híbridos, que combinam as duas primeiras. Essas abordagens, cada uma com seus pontos fortes e limitações, têm sido combinadas de modo a oferecer opções mais precisas e relevantes, aprimorando a experiência do usuário por meio da personalização dos serviços.

A integração de técnicas de PLN aos sistemas de recomendação representa um avanço significativo, ao permitir que esses sistemas compreendam e manipulem textos de forma mais eficiente para gerar recomendações acuradas, refinadas e relevantes. O PLN, enquanto subárea da inteligência artificial, dedica-se ao estudo e ao desenvolvimento de métodos que possibilitam a interpretação e a geração de informações em linguagem natural pelas máquinas. Essa integração tem se mostrado particularmente útil quando se lida com dados textuais, como descrições de produtos, resenhas de usuários, postagens em redes sociais ou textos acadêmicos.

O uso de PLN em sistemas de recomendação permite não apenas aprimorar a acurácia das recomendações baseadas no comportamento passado do usuário, mas também considerar o contexto semântico do conteúdo analisado, oferecendo sugestões mais refinadas e pertinentes. Além disso, com o crescente emprego de técnicas de aprendizado profundo, esses sistemas têm alcançado maior eficácia na análise de grandes volumes de dados não estruturados, utilizando arquiteturas como redes neurais convolucionais para imagens e *transformers* para texto.

A combinação entre PLN e sistemas de recomendação encontra-se, portanto, na vanguarda das pesquisas em inteligência artificial, com impactos relevantes em múltiplos domínios, incluindo, ainda que de forma menos explorada, o educacional (Pereira, Gomes e Primo, 2022). Ao considerar os achados deste

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

mapeamento, nota-se que a presença de estudos sobre PLN e recomendação em periódicos da área de informática na educação é incipiente, o que reforça a relevância de futuras investigações que articulem diretamente esses campos. Este estudo realiza uma revisão estruturada da literatura sobre o uso de processamento de linguagem natural em sistemas de recomendação entre os anos de 2020 a 2024 e 2025 e busca responder à seguinte questão de pesquisa: Quais são as principais abordagens e tendências no uso de processamento de linguagem natural (PLN) em sistemas de recomendação?

A filtragem colaborativa baseia-se no comportamento coletivo dos usuários para gerar recomendações. Nesse modelo, se dois usuários compartilham preferências similares, é provável que eles também apreciem os mesmos itens no futuro. Por exemplo, em plataformas de filmes, um usuário que gostou de determinados filmes pode receber sugestões de filmes que outros usuários com gostos semelhantes também apreciaram. Uma das grandes vantagens dessa abordagem é que ela não exige um entendimento profundo sobre o conteúdo dos itens, mas unicamente sobre o histórico de interações dos usuários. No entanto, ela enfrenta desafios, como o problema do *cold start*, que ocorre quando um novo item ou usuário não tem dados suficientes para gerar recomendações precisas.

A filtragem baseada em conteúdo, por outro lado, recomenda itens com base nas características dos próprios itens, como descrições, categorias ou etiquetas associadas. Como caso, um sistema de recomendação de livros pode sugerir novos títulos com base no gênero ou no autor de livros previamente lidos pelo usuário. Essa perspectiva não depende das interações de outros usuários, o que pode ser uma vantagem em contextos nos quais a quantidade de dados de usuário é limitada. Contudo, um dos problemas dessa linha metodológica é a tendência a recomendar tão-só itens muito similares aos já conhecidos, limitando a descoberta de novas preferências.

Já os sistemas híbridos buscam combinar as vantagens da filtragem colaborativa e da filtragem baseada em conteúdo. Ao integrar ambas as abordagens, esses sistemas conseguem superar as limitações de cada uma delas ao usar, por exemplo, a filtragem colaborativa para sugerir itens com base no comportamento de outros usuários, enquanto a filtragem baseada em conteúdo pode enriquecer as recomendações ao considerar as características dos próprios itens. Essa combinação reduz os problemas de cold start e melhora a personalização, embora sua implementação seja mais complexa devido à integração de diferentes fontes de dados.

A aplicação de PLN permite que os sistemas extraiam semântica e contexto de textos não estruturados, proporcionando recomendações mais personalizadas e contextualizadas. Uma das principais técnicas do PLN são os *embeddings*, que consistem em representar palavras ou frases como vetores em um espaço contínuo de

alta dimensionalidade. Modelos como Word2Vec (Mikolov *et al*., 2013a; Mikolov *et al*., 2013b) e GloVe (Pennington, Socher e Manning, 2014) foram pioneiros na criação desses embeddings, permitindo que palavras com significados semelhantes fossem representadas por vetores próximos no espaço. Estes modelos, além de capturarem o significado das palavras de forma mais precisa do que métodos tradicionais, como a representação *one-hot*, como também mantêm as relações semânticas e sintáticas entre elas.

Nos últimos anos, o uso de modelos de transformers como o BERT (*bidirectional encoder representations from transformers*; Devlin *et al*., 2019: 8) e o GPT (*generative pretrained transformer*; Brown *et al*., 2020) tem revolucionado o campo do PLN. O BERT, como prova, é um modelo pré-treinado de transformer bidirecional, que aprende a partir de grandes volumes de texto e é capaz de capturar o contexto completo das palavras em uma frase, considerando o contexto à esquerda e à direita de uma palavra. Isso permite que o BERT se destaque em tarefas como classificação de texto, análise de sentimentos e, claro, sistemas de recomendação, onde o contexto é fundamental para entender as intenções do usuário.

Por outro lado, os modelos generativos como o GPT (em suas diversas versões), Gemini, Claude e outros têm se mostrado altamente eficazes em tarefas de geração de texto e recomendação personalizada. Esses modelos utilizam treinamento baseado em grandes quantidades de dados com potencial de geração de respostas contextuais realistas, o que os torna ideais para sistemas que precisam oferecer sugestões dinâmicas e criativas, como em *chatbots* ou sistemas de recomendação de conteúdo.

A aplicação dessas técnicas no desenvolvimento de sistemas de recomendação tem se expandido, com destaque para a personalização das sugestões, a melhoria da compreensão de conteúdo textual e o aumento da precisão das previsões. Para melhor compreensão das análises realizadas neste estudo, alguns conceitos, a exemplo de algumas técnicas de PLN, são relevantes, conforme apresentado resumidamente na *tabela 1*.

Técnica de PLN	Descrição sintética	Aplicações típicas	Fonte
TF-IDF (*term frequency-inverse document frequency*)	Mede a relevância de palavras em documentos, considerando frequência local e raridade global.	Representação vetorial de textos e cálculo de similaridade.	Salton e Buckley (1988) Manning, Raghavan e Schütze (2008)
LSTM (*long short-term memory*)	Rede neural recorrente que modela dependências de longo prazo em sequências.	Histórico de interações e sequência de cliques.	Hochreiter e Schmidhuber (1997)

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

GRU (*gated recurrent unit*)	Variante simplificada da LSTM, com desempenho comparável e menos parâmetros.	Modelagem sequencial de preferências com menor custo computacional.	Cho *et al.* (2014)
CNN (*convolutional neural network*)	Rede neural que extrai padrões locais de dados, incluindo textos.	Classificação de comentários, análise de tópicos.	Kim (2014)
VAE (*variational autoencoders*)	Técnicas de codificação e reconstrução de dados, gerando representações compactas e latentes.	Compressão de descrições e inferência de embeddings de usuários.	Kingma e Welling (2014)

Tabela 1. Técnicas de PLN com respectivas descrições, aplicações e fonte
Fonte: elaboração de Denise Fukumi Tsunoda com base nos dados da pesquisa realizada pelos autores, 2025

Os modelos de linguagem (*language models*) são algoritmos treinados para compreender, gerar ou prever sequências de palavras em linguagem natural, com base em padrões estatísticos e contextuais extraídos de grandes volumes de texto. Em essência, eles atribuem uma probabilidade a sequências linguísticas, permitindo que o sistema 'entenda' o contexto e gere respostas coerentes.

Esses modelos podem variar de estruturas simples, como modelos n-gramas, até arquiteturas avançadas baseadas em redes neurais profundas, como o BERT e o GPT. Eles são amplamente utilizados em tarefas como tradução automática, resposta a perguntas, resumos automáticos, classificação de texto e, mais recentemente, em sistemas de recomendação com análise semântica. A *tabela 2* apresenta um resumo dos principais modelos de linguagens utilizados em sistemas de recomendação com PLN.

Modelo de linguagem	Descrição sintética	Aplicações em sistemas de recomendação	Fonte
Word2Vec (*word to vector*)	Vetorização semântica de palavras com base em janelas de contexto.	Similaridade entre termos e itens, geração de embeddings de conteúdo.	Mikolov *et al.* (2013a) Mikolov *et al.* (2013b)
Doc2Vec (*document to vector*)	Representação vetorial de sentenças ou documentos inteiros.	Perfil de usuários, resenhas, recomendações de texto longo.	Le e Mikolov (2014)
FastText	Extensão do Word2Vec que considera subpalavras, melhorando vocabulário e morfologia.	Recomendação multilíngue ou com textos informais ou incompletos.	Bojanowski *et al.* (2017)

BERT	Modelo pré-treinado bidirecional baseado em transformers, sensível ao contexto.	Recomendação baseada em significado contextual de sentenças.	Devlin *et al.* (2019)
GPT	Modelo autorregressivo para geração e compreensão de linguagem natural.	Geração de resumos, descrições e interações conversacionais.	Brown *et al.* (2020)
RoBERTa (*robustly optimized BERT approach*)	Versão otimizada do BERT com treinamento mais robusto e maior conjunto de dados.	Embeddings mais refinados para tarefas de *ranking* e *matching*.	Liu *et al.* (2019)
DistilBERT (*distilled BERT*)	Versão compacta e mais rápida do BERT, com desempenho próximo ao original.	Casos em que o tempo de resposta é crítico (*apps, mobile*).	Sanh *et al.* (2019)

Tabela 2. Técnicas de representação vetorial (embeddings) e modelos de linguagem com respectivas descrições, aplicações e fonte
Fonte: elaboração de Denise Fukumi Tsunoda com base nos dados da pesquisa realizada pelos autores, 2025

No contexto educacional, os sistemas de recomendação têm sido empregados para apoiar a personalização de ambientes virtuais de aprendizagem, sugerindo materiais e atividades alinhados ao perfil dos estudantes e às demandas do curso. Essa integração aproxima a tecnologia de práticas pedagógicas centradas no aluno e dialoga com investigações da comunidade brasileira de informática na educação (Baker, Isotani e Carvalho, 2011: 5).

Apresentados os principais conceitos com vistas à melhor compreensão do estudo, a próxima seção detalha os encaminhamentos metodológicos.

Metodologia

Este estudo adota uma abordagem sistemática para revisar a literatura sobre o uso de processamento de linguagem natural (PLN) em sistemas de recomendação, com foco nas publicações entre 2020 e 2025. A metodologia é composta por várias etapas que incluem a definição de critérios de pesquisa, a seleção de fontes de dados, a triagem dos artigos e a análise detalhada dos documentos recuperados, com o objetivo de identificar as tendências emergentes no uso de PLN para sistemas de recomendação. A revisão seguiu as diretrizes do protocolo PRISMA (Page *et al.*, 2021), contemplando as etapas de identificação, triagem, elegibilidade e inclusão. Foram estabelecidos critérios de inclusão e exclusão explícitos, aplicados com o apoio das ferramentas Zotero e Rayyan, que garantiram maior rastreabilidade e

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

auditabilidade do processo. A análise eliminou duplicidades, considerou apenas artigos revisados por pares e priorizou publicações de relevância comprovada entre 2020 e 2025. Embora não tenha sido realizada avaliação de risco de viés em virtude do caráter exploratório do estudo, todas as etapas do processo foram documentadas e conduzidas de modo a assegurar rigor metodológico e transparência.

Materiais

Para a análise dos artigos e a extração das informações bibliométricas, foram utilizadas ferramentas digitais baseadas em inteligência artificial, selecionadas por sua contribuição à curadoria, triagem e interpretação dos dados. Todas as decisões foram validadas por revisores humanos, garantindo rigor metodológico e rastreabilidade.

As ferramentas de inteligência artificial foram utilizadas entre o 1.º e o 31 de maio de 2025, em suas versões vigentes no período, conforme descrito a seguir:

- ChatGPT de OpenAI, versão GPT-4: ferramenta utilizada na versão paga para apoio na análise de conteúdo, identificação de tendências e sistematização dos resultados extraídos. Para assegurar a reprodutibilidade metodológica, a interação com o modelo seguiu um protocolo de dois eixos: 1. Para classificação técnica: "Atue como especialista em ciência de dados. Analise o resumo e metodologia. Extraia: a) a técnica principal de processamento de linguagem natural, b) o tipo de filtragem, c) o domínio de aplicação"; 2. Para tendências (2025): "Identifique no texto completo: a) limitações reportadas, b) se propõe nova arquitetura ou aplicação, c) tendências futuras de integração com LLMs". Reitera-se que todas as saídas foram validadas por revisão humana;
- Zotero de Corporation for Digital Scholarship, versão 6.0.30: ferramenta utilizada na versão gratuita para organização das referências, remoção de duplicatas e exportação dos artigos no formato BibTeX, compatível com o estilo Scopus para análise no Biblioshiny;
- Biblioshiny da Università degli Studi di Napoli Federico II, versão 4.3: interface gráfica do pacote bibliometrix em R, empregada para a realização das análises bibliométricas exploratórias;
- OpenAlex de OurResearch, API versão 1.6, usada entre o 16 e 23 de abril de 2025: utilizado para a recuperação de metadados dos artigos, como ano de publicação, palavras-chave, número de citações e informações sobre os periódicos;
- Rayyan de Qatar Computing Research Institute, versão web 2.0: ferramenta utilizada na versão gratuita em uma segunda etapa de deduplicação

(não detectada pelo Zotero), além do processo de *screening* e exclusão de artigos com arquivos PDF incompletos (por exemplo, contendo apenas o resumo);

- NotebookLM de Google Research, versão beta: ferramenta utilizada na versão gratuita para leitura e análise dos PDF, em lotes de até 50 documentos por vez, otimizando a identificação de trechos relevantes;
- SciSpace de Typeset Inc., versão pro: ferramenta utilizada na versão paga para análises automatizadas do texto completo dos artigos em PDF, possibilitando a extração de informações sobre técnicas, métodos e aplicações;
- Litmaps de Litmaps Ltd., versão web 1.4: ferramenta utilizada para visualização de redes de citações e identificação de artigos-nó com maior centralidade temática.

Adicionalmente ao OpenAlex foram utilizadas outras quatro bases de dados de periódicos, das quais duas são a Scopus e a Web of Science, ambas utilizadas em 16 de abril de 2025, conforme *figura 1*. Adicionalmente, foram incluídas bases de nicho, como a da *Revista Brasileira de Informática na Educação* (*RBIE*, s. d.) e da biblioteca digital da Sociedade Brasileira de Computação (SBC, s. d.) para verificar a capilaridade das técnicas de PLN em domínios específicos e compará-las com as bases generalistas. As consultas nas bases adicionais foram realizadas em 20 de agosto de 2025, utilizando os mesmos descritores aplicados nas demais, em português e inglês ("sistema de recomendação" AND "processamento de linguagem natural"; "recommender system" AND "natural language processing"), para o período de 2020 a 2024 e 2025. Nenhum artigo adicional foi identificado, de modo que o corpus permaneceu inalterado em relação às buscas realizadas em OpenAlex, Scopus e Web of Science.

Cabe destacar que a seleção das ferramentas utilizadas nesta pesquisa foi orientada não apenas por critérios de desempenho técnico, mas, sobretudo, por sua contribuição ao processo de construção e organização do conhecimento. Recursos como ChatGPT foram empregados com a finalidade de apoiar a sistematização e a interpretação dos dados, auxiliando na identificação de padrões conceituais e na estruturação temática dos resultados. Além disso, ferramentas como SciSpace e NotebookLM, ao permitirem a leitura e análise de textos completos em larga escala, atuaram como suporte à extração contextualizada de informações metodológicas e aplicadas. Zotero e Rayyan, por sua vez, foram fundamentais na etapa de curadoria bibliográfica, assegurando consistência na triagem e confiabilidade na gestão das fontes.

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

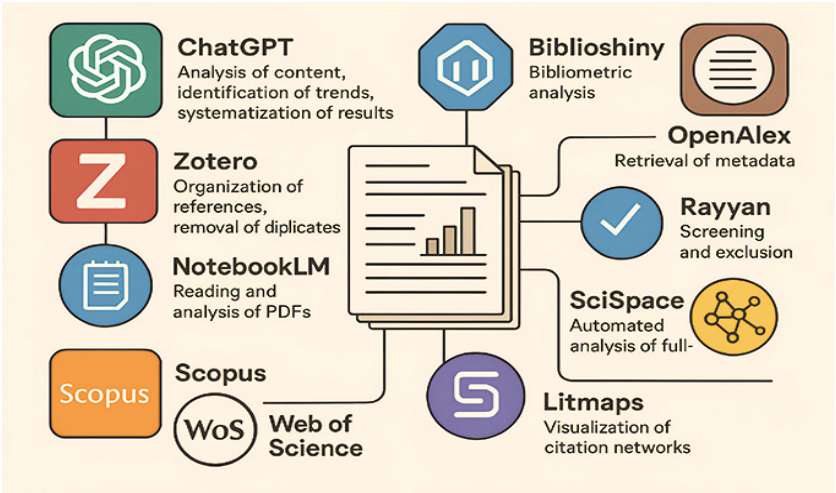

Figura 1. Resumo de todas as ferramentas utilizadas na pesquisa
Fonte: elaboração de Denise Fukumi Tsunoda com base nos dados da pesquisa realizada pelos autores, 2025

Dessa forma, cada tecnologia adotada desempenhou um papel complementar no aprofundamento analítico da revisão, contribuindo tanto para a robustez metodológica quanto para a coerência interpretativa dos achados. Cabe destacar ainda que as ferramentas baseadas em inteligência artificial (ChatGPT, SciSpace, NotebookLM e outras) foram utilizadas como apoio à triagem e organização de informações e todas as decisões de inclusão/exclusão e a extração final de dados foram validadas por revisores humanos, não havendo delegação de julgamento científico a modelos automatizados. As próximas seções detalham o uso dessas ferramentas nas diversas etapas da pesquisa.

Estratégia de pesquisa

A estratégia de busca foi formulada com o objetivo de identificar artigos que tratam da aplicação de processamento de linguagem natural (PLN) em sistemas de recomendação, utilizando a seguinte expressão de pesquisa: ((“recommendation system” OR “recommender system”) AND (“natural language processing” OR “NLP”)) OR (“sistema de recomendação” AND (“processamento de linguagem natural” OR “PLN”)). Os descritores foram escolhidos de forma a abranger tanto os sistemas de recomendação quanto as principais técnicas de PLN empregadas nesse contexto, garantindo a recuperação de estudos relevantes e alinhados à questão de pesquisa. Assim, para a análise 2020-2024, foi utilizado o OpenAlex e, para fins de filtragem, foram considerados os artigos de acesso aberto publicados entre 2020 e 2024, em inglês ou português.

A base de periódicos OpenAlex foi escolhida considerando os seguintes critérios: acesso gratuito e aberto, integração de múltiplas fontes científicas, exportação estruturada e compatível com ferramentas de curadoria e bibliometria (a exemplo do Biblioshiny), além de atualizações constantes. Seu diferencial mais relevante, contudo, reside no uso de inteligência artificial para desambiguação de autores e instituições, bem como na classificação semântica dos artigos por meio de embeddings, o que garante maior precisão na organização e análise de dados científicos. Tais recursos tornam o OpenAlex uma base mais eficiente e adequada às exigências metodológicas do presente estudo no período de 5 anos (de 2020 a 2024). A pesquisa por artigos publicados entre 2020 e 2024 foi realizada no OpenAlex, em 16 de abril de 2025, e retornou 538 registros.

Para a análise de 2025, com a mesma estratégia de busca do período 2020-2024, foram consultadas as seguintes bases de dados indexadas, além do próprio OpenAlex:

a) Scopus: uma das maiores bases de dados acadêmicas, que inclui artigos revisados por pares, conferências e periódicos de diversas áreas do conhecimento;
b) Web of Science (WOS): base de dados altamente reconhecida, abrangendo uma ampla gama de periódicos, conferências e relatórios.

A recuperação dos documentos da base 2025 foi conduzida em 23 de abril de 2025, com o objetivo de garantir uma amostra representativa de publicações nos dois idiomas. Assim, a base de dados referente ao ano de 2025 é composta por artigos recuperados da base OpenAlex, Scopus e Web of Science e totalizou 63 registros. De forma resumida, nas buscas 2020-2024 em OpenAlex, aplicaram-se filtros por ano (2020-2024), tipo (article/review) e idioma (pt/en) nos campos título, resumo e palavras-chave. Em 2025 no OpenAlex, Scopus e WOS, repetiram-se os descritores e filtros para janeiro-abril de 2025.

Todas as fontes foram escolhidas por sua relevância, abrangência e qualidade dos artigos indexados, cobrindo diferentes áreas do conhecimento e proporcionando uma ampla visão sobre os avanços nos sistemas de recomendação baseados em PLN.

Critérios de inclusão e exclusão

Os critérios de inclusão e exclusão foram definidos para garantir que os artigos selecionados atendam aos objetivos da pesquisa.

Os critérios de inclusão foram:

a) Artigos publicados entre 2020-2024 e 2025 (duas pesquisas distintas);
b) Estudos que utilizam PLN como componente central do sistema de recomendação;

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

c) Artigos que sejam estudos empíricos ou revisões de literatura relevantes;
d) Trabalhos disponíveis com acesso ao texto completo.

Os critérios de exclusão foram:

a) Trabalhos que não utilizam PLN como componente essencial do sistema de recomendação;
b) Estudos em idiomas diferentes de inglês ou português;
c) Documentos incompletos, como resumos de conferências sem acesso ao artigo completo.

Foram incluídos apenas artigos revisados por pares, publicados em acesso aberto, com foco explícito no uso de técnicas de PLN aplicadas a sistemas de recomendação. Foram excluídos trabalhos não revisados por pares, duplicados, capítulos de livros, anais de eventos sem avaliação formal e artigos que apenas mencionavam PLN ou recomendação sem estabelecer relação metodológica entre ambos. Os artigos que passaram pela triagem automatizada foram revisados manualmente para garantir o atendimento aos critérios de inclusão e exclusão. A avaliação final (pela inclusão ou exclusão do trabalho) foi realizada por dois avaliadores independentes e divergências foram resolvidas por um terceiro parecerista.

Fluxo de seleção e análise

O processo seguiu as recomendações do PRISMA (Page *et al.*, 2021) para revisões estruturadas, com adaptações, conforme segue. A triagem inicial da base 2020-2024 utilizou o Zotero com a carga do CSV gerado pelo OpenAlex. Foi realizada a remoção de duplicatas e registros inacessíveis. Essa fase visou filtrar rapidamente artigos irrelevantes ou duplicados.

Após deduplicação no Zotero sobraram 491 artigos. Estes 491 foram carregados no Rayyan que detectou outras 7 duplicações. Após minuciosa análise, verificou-se que apenas 3 estavam duplicadas e estas foram excluídas. As demais eram similares, mas não duplicações. Assim, restaram 488 artigos únicos.

Ainda no Rayyan, na etapa de triagem, foram lidos títulos, abstracts e palavras-chave de cada um dos 488 artigos, conforme segue. Cada artigo foi analisado pelo ChatGPT e por um pesquisador do grupo. As dúvidas foram resolvidas por outro pesquisador em revisão cega. Desta etapa, restaram 214 artigos. Esse processo garantiu rigor metodológico ao combinar ferramentas automatizadas com validação humana, minimizando tanto vieses de inclusão quanto a manutenção de duplicatas. Os 214 arquivos em formato PDF foram inseridos no Zotero para exportação do BibTeX que alimentou as análises métricas no Biblioshiny, uma vez que o arquivo gerado pelo Rayyan não é compatível com os padrões do Biblioshiny. Os mesmos PDF dos 214 artigos foram analisados no SciSpace e no NotebookLM3 e um arquivo XSLX foi gerado para análise no ChatGPT.

Para a composição da base de dados referente ao ano de 2025, foram recuperados 34 documentos da base OpenAlex, 23 da Scopus e 6 da Web of Science, totalizando 63 registros. Após o processo de deduplicação realizado no Rayyan, restaram 52 artigos únicos para a etapa de triagem. Cada um desses 52 artigos foi analisado colaborativamente pelo ChatGPT e por um pesquisador da equipe, resultando na seleção de 26 artigos para a próxima fase. Todos os PDF dos 26 artigos selecionados foram devidamente baixados para análise textual no SciSpace, e um arquivo no formato XLSX foi produzido para subsidiar a etapa de análise temática e técnica no ChatGPT. A *tabela 3* sumariza as etapas do fluxo da revisão, bem como explicita os estudos incluídos e excluídos. Todas as bases de dados utilizadas na pesquisa estão compartilhadas em <https://doi.org/10.5281/zenodo.15466656> de forma pública.

Etapa	2020-2024	2025	Total
Registros identificados (OpenAlex, Scopus e WOS)	960 (OpenAlex)	34 (OpenAlex), 6 (WOS), 23 (Scopus)	1 023
Registros após remoção de duplicado (Zotero e Rayyan)	488	52	540
Registros submetidos à triagem (título, resumo e palavras-chave)	488	52	540
Registros excluídos na triagem	274	26	300
Estudos incluídos após triagem	214	26	240
Estudos incluídos na análise final (Os PDF analisados no Zotero, SciSpace, NotebookLM e ChatGPT)	214	26	240

Tabela 3. Resumo da aplicação da metodologia PRISMA
Fonte: elaboração de Denise Fukumi Tsunoda e Patrick Fernandes Rezende Ribeiro com base nos dados produzidos na pesquisa realizada pelos autores, 2025

Após a seleção final dos artigos, foi realizada uma análise qualitativa e quantitativa dos dados. A análise qualitativa envolveu a leitura e interpretação do conteúdo dos artigos para identificar as técnicas de PLN aplicadas nos sistemas de recomendação, bem como as áreas de aplicação mais exploradas, como comércio eletrônico, educação personalizada e análise de sentimentos (Liu, 2012).

Já a análise quantitativa envolveu a extração de dados bibliométricos para entender a distribuição temporal das publicações e a frequência de uso de determinados modelos de PLN, como BERT, GPT e embeddings, bem como a identificação das palavras-chave mais recorrentes.

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

Apresentação e análise dos resultados

Em uma análise global, o crescimento em número de artigos nos últimos cinco anos é expressivo, conforme apresentado na *figura 2*. A distribuição temporal das publicações evidencia crescimento contínuo entre 2020 e 2024 e os anos de 2023 e 2024, somados respondem por mais de 65 % da produção científica do intervalo considerado.

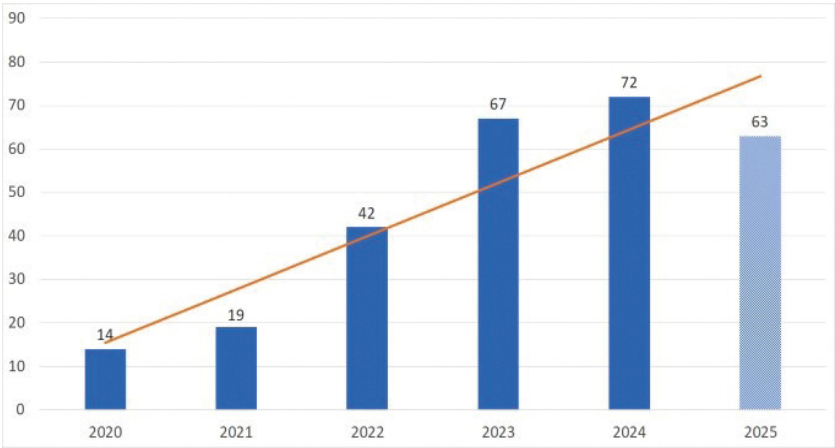

Figura 2. Publicações sobre processamento de linguagem natural e sistemas de recomendação nos últimos 5 anos
Fonte: elaboração de Denise Fukumi Tsunoda e Patrick Fernandes Rezende Ribeiro com base nos dados produzidos na pesquisa realizada pelos autores, 2025
Nota: a coluna de 2025 refere-se a dados parciais (janeiro-abril), indicando tendência de alta

Observa-se que a produção científica sobre PLN em sistemas de recomendação mantém uma trajetória de ascensão acelerada. Enquanto o período de 2020 a 2024 registrou um crescimento constante, quadruplicando o volume anual (de 14 para 72 publicações), os dados de 2025 revelam uma mudança de patamar. Apenas no primeiro quadrimestre do ano, foram identificados 63 documentos relevantes em OpenAlex, Scopus e WOS. Embora o valor absoluto ainda seja inferior ao total acumulado de 2024, a média mensal de publicações em 2025 supera os anos anteriores, sugerindo que o pico de interesse acadêmico está ocorrendo no presente momento, impulsionado pela popularização dos modelos de linguagem de grande escala.

O crescimento observado a partir de 2022 pode estar associado à consolidação de modelos pré-treinados, como BERT e GPT; à democratização de bibliotecas, como a Hugging Face; e à popularização de aplicações personalizadas em domínios como educação, mercado de trabalho e saúde.

O Biblioshiny apresenta 1 417 autores distintos, uma média de 4.55 autores por documento, apenas 14 artigos com autores individuais, 6 357 referências, 149 fontes e uma média de 24.21 citações por documento.

O Litmaps apresentou os artigos ordenados por ordem de citação, conforme *figura 3*. Um Litmap (abreviação de Literature Map) é uma representação visual interativa da literatura científica, na qual cada artigo é um nó e as conexões representam relações conceituais, temáticas ou de citação entre eles. A inspeção das afiliações institucionais correspondentes aos dez artigos de maior centralidade visualizados na *figura 3* permitiu identificar a origem geográfica dessa produção de alto impacto. Verificou-se que nove destes trabalhos são provenientes de instituições da China, enquanto um (Gugnani e Misra, 2020) origina-se da Índia. Esse dado qualitativo, extraído dos nós de maior destaque na rede, corrobora a hegemonia asiática na fronteira de pesquisa sobre modelos de linguagem de grande escala e recomendação. No cenário nacional, destaca-se qualitativamente o artigo de Bazzan *et al*. (2023), que embora não figure no topo das citações globais, exemplifica a aplicação prática no contexto brasileiro.

De pesquisadores brasileiros, destaque para o artigo "An Information Management Model for Addressing Residents' Complaints Through Artificial Intelligence Techniques", de Bazzan *et al*. (2023), que propõe um modelo de gestão da informação para tratamento de reclamações de moradores em empreendimentos residenciais utilizando técnicas de inteligência artificial. A partir de um estudo de caso em uma construtora brasileira, o modelo foi desenvolvido com base na metodologia *design science research* e contempla um sistema hierárquico de classificação de defeitos; um menu estruturado de palavras para o registro de reclamações e um sistema de recomendação baseado em algoritmos de aprendizado de máquina. As tecnologias de PLN e aprendizado supervisionado foram aplicadas para automatizar a categorização das queixas, melhorar a coleta de dados e reduzir o tempo das inspeções técnicas, contribuindo para a eficiência dos serviços de assistência técnica e para a geração de conhecimento para a melhoria da qualidade do produto.

A *figura 3* apresenta a estrutura de influência da rede analisada, ordenando os nós por volume de citações. O trabalho de maior centralidade no período recente é o de Yang *et al.* (2023), identificado automaticamente na visualização do software como "Chen, 2023" devido à extração de metadados do segundo autor. Com 91 citações contabilizadas na base de dados utilizada, este estudo se destaca como um hub de referência para a aplicação de modelos de linguagem de grande escala em personalização.

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

Tags	Author	Year	Title	References	Citations
	Chen	2023	PALR: Personalization Aware LLMs for Recommendation	23	91
	Li	2023	GPT4Rec: A Generative Framework for Personalized Recommendation and User Interests Int...	30	84
	Aljunid	2020	An Efficient Deep Learning Approach for Collaborative Filtering Recommender System	24	74
	Wu	2024	Exploring Large Language Model for Graph Data Understanding in Online Job Recommendati...	31	72
	Gugnani	2020	Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation	18	62
	Geng	2023	VIP5: Towards Multimodal Foundation Models for Recommendation	85	60
	Ding	2021	An Overview of Machine Learning-Based Energy-Efficient Routing Algorithms in Wireless Sen...	85	54
	Ji	2023	GenRec: Large Language Model for Generative Recommendation	19	46
	Cho	2020	McDRAM v2: In-Dynamic Random Access Memory Systolic Array Accelerator to Address the ...	55	45
	Wang	2020	Using Natural Language Processing Techniques to Provide Personalized Educational Material...	55	44

Figura 3. Artigos ordenados por ordem de citação no Litmaps
Fonte: elaboração de Denise Fukumi Tsunoda e Patrick Fernandes Rezende Ribeiro com base nos dados produzidos na pesquisa realizada pelos autores, 2025

O artigo "PALR: Personalization-Aware LLMs for Recommendation" investiga o uso de modelos de linguagem de grande escala no contexto de sistemas de recomendação personalizados. A proposta parte da observação de que, embora os modelos apresentem avanços notáveis em tarefas de compreensão e geração de linguagem, sua eficácia na personalização de recomendações ainda é limitada por abordagens genéricas.

A predominância deste estudo na rede de citações (conforme visualizado na *figura 3*) não é acidental, mas reflete uma mudança de paradigma na área. Enquanto os primeiros trabalhos com modelos de linguagem de grande escala focavam em geração de texto genérico, o modelo PALR (Yang *et al.*, 2023) preencheu uma lacuna crítica ao introduzir a 'consciência de personalização' (*personalization-awareness*) na arquitetura. Essa inovação técnica explica sua alta taxa de citação e centralidade: ele ofereceu à comunidade evidências empíricas de que é possível superar as métricas de *baselines* tradicionais ao ajustar o modelo de linguagem de grande escala com o histórico de comportamento do usuário, validando a tendência de hibridização apontada anteriormente na análise quantitativa deste estudo.

A análise da rede de conexões, detalhada na *figura 4*, revela que a relevância de Yang *et al.* (2023) não é isolada. O gráfico evidencia uma forte aresta de conexão (co-citação) com o *survey* de Lin *et al.* (2024). Essa ligação visual, representada pela espessura e proximidade dos nós, sugere estatisticamente que o campo está se estruturando em torno de dois eixos complementares: a fundamentação teórica fornecida pela revisão de Lin *et al.* (2024) e a validação empírica de arquiteturas propostas por Yang *et al.* (2023). Diferente de uma dispersão aleatória, a topologia da rede indica uma rápida consolidação de frameworks de recomendação baseados em *prompts* e ajuste fino (*fine-tuning*) entre 2023 e 2024.

Para compreender a estrutura de colaboração e influência no cluster mais denso da rede, gerou-se uma visualização de vizinhança (*figura 4*) utilizando o artigo de Yang *et al*. (2023) como nó semente (*seed paper*). Essa abordagem metodológica foi adotada para garantir a legibilidade das conexões, uma vez que a projeção estática da rede completa (240 nós) resultaria em sobreposição excessiva de arestas, o que dificultaria a interpretação analítica dos dados.

A topologia apresentada na *figura 4* revela uma forte conexão de co-citação entre o trabalho empírico de Yang *et al*. (2023) e o *survey* de Lin *et al.* (2024). A análise desta aresta específica é crítica: ela demonstra que o campo de modelos de linguagem de grande escala em recomendação não cresce de forma dispersa, mas articulada. O *survey* de Lin *et al*. (2024) atua como um organizador teórico que sistematiza *prompts* e *fine-tuning*, enquanto o modelo PALR de Yang *et al*. (2023) valida essas teorias na prática. A densidade de conexões entre esses dois nós e os demais autores periféricos, como Wu e Geng, sugere a formação de uma 'comunidade de prática' consolidada, focada em resolver os problemas de alucinação e explicabilidade identificados na revisão sistemática.

No que tange ao ecossistema tecnológico de implementação, a análise textual via SciSpace permitiu mapear as ferramentas de desenvolvimento predominantes no corpus de 2020-2024. Identificou-se a hegemonia da linguagem Python, com 27 menções explícitas, frequentemente associada a bibliotecas de aprendizado profundo, como PyTorch, com 4 artigos; TensorFlow; e Scikit-Learn. A linguagem R aparece em segundo plano com 6 ocorrências, seguida por menções isoladas a Java (Yang, 2022) e abordagens poliglotas (Shaikh *et al*., 2023; Velpula *et al*., 2024).

É relevante notar, contudo, que a maioria dos estudos (178 artigos) não explicita a linguagem de programação ou framework utilizado. Essa ausência de detalhamento técnico reflete a natureza de parte significativa da literatura, que prioriza a discussão de modelos matemáticos, arquiteturas conceituais e avaliações algorítmicas em detrimento da implementação de software. Um exemplo representativo desse grupo é o trabalho de Pires, Rizzi e Almeida (2024) que, embora não detalhe o código, contribui com diretrizes críticas para a avaliação intrínseca de embeddings em filtragem colaborativa, demonstrando que o rigor teórico independe da explicitação da *stack* tecnológica.

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

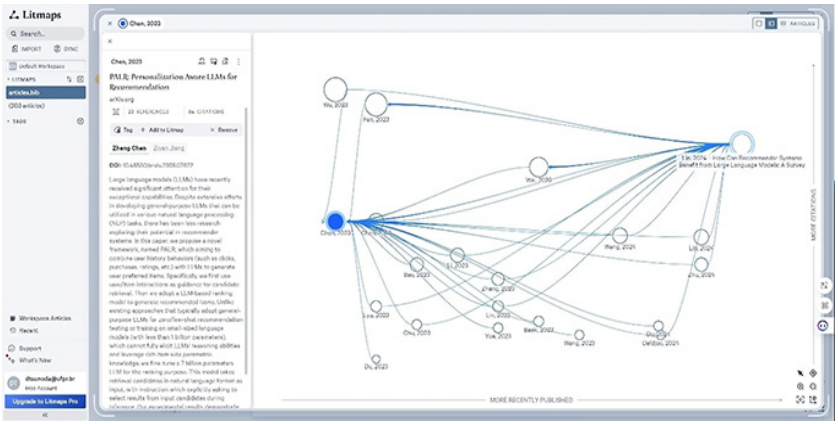

Figura 4. Litmap do artigo "PALR: Personalization Aware LLMs for Recommendation" de Yang *et al*. (2023)
Fonte: elaboração de Denise Fukumi Tsunoda e Patrick Fernandes Rezende Ribeiro com base nos dados produzidos na pesquisa realizada pelos autores, 2025

Quanto à aplicação do sistema de recomendação que utiliza PLN, a *tabela 4* apresenta as análises e os resultados apontam uma diversidade de aplicações, lideradas por entretenimento focado em filmes e comércio eletrônico, com a educação (18 %) aparecendo como um domínio emergente relevante.

Tipo de sistema de recomendação	Núm. de artigos	Percentual	Destaques
Outro/não especificado	57	27 %	Diversos artigos (57) abordam técnicas genéricas de PLN e recomendação sem delimitar claramente o domínio de aplicação. Inclui temas como frameworks, arquitetura de sistemas e comparações metodológicas.
Filmes	44	21 %	Forte presença histórica, refletindo o uso de benchmarks como MovieLens e IMDB. Frequentemente utilizados para testar modelos com avaliações explícitas dos usuários.
Educação	39	18 %	Surgimento de sistemas de recomendação para cursos, conteúdos educacionais, tutoriais e personalização do ensino. Indica forte ligação com estratégias de aprendizagem adaptativa.
Produtos	30	14 %	Foco em recomendação baseada em análises de sentimentos, resenhas e descrições de itens com destaque para *marketplaces* e comércio eletrônico.

Vagas/currículos	14	7 %	Tendência crescente de uso de PLN em análise de currículos, compatibilidade de perfil e análise semântica de descrições de vagas. Frequentemente combinados com embeddings e classificação textual.
Livros	10	5 %	Sistemas que recomendam leituras com base em resenhas, estilo de escrita ou similaridade textual entre obras e perfis de leitura.
Saúde	9	4 %	Uso de PLN para interpretar prescrições, resumos clínicos, interações medicamentosas e sintomas. Forte presença de análises semânticas e sentimentais.
Restaurantes	5	2 %	Recomendação baseada em comentários e avaliações de clientes. Processamento de linguagem natural aplicado a análise de sentimento e extração de preferências.
Música	5	2 %	Extração de preferências musicais a partir de letras de músicas, resenhas e etiquetas semânticas.
Moda	1	0 %	Aplicação que utiliza descrições de estilo, tendências e preferências textuais para sugestão de roupas.
Total	**214**	**100 %**	

Tabela 4. Principais aplicações identificadas para uso de PLN em sistemas de recomendações
Fonte: elaboração de Denise Fukumi Tsunoda e Patrick Fernandes Rezende Ribeiro com base nos dados produzidos na pesquisa realizada pelos autores, 2025

Ainda com a base 2020-2024 e após análise dos conteúdos dos 214 documentos, verifica-se nas abordagens dispostas na *tabela 5* que o BERT e o TF-IDF lideram com ampla margem, revelando uma coexistência entre técnicas modernas, baseadas em transformers, e clássicas, baseadas em estatísticas. As técnicas de redes neurais recorrentes como LSTM, GRU e RNN continuam populares, especialmente para modelagem de sequência, enquanto o Word2Vec mantém sua utilidade como base de vetorização semântica leve.

Quanto à evolução dessas abordagens, a análise da distribuição temporal corrobora a hipótese de transição tecnológica. Verifica-se que abordagens estatísticas como TF-IDF mantêm uma presença constante ao longo de todo o período (2020-2024), atuando como baseline robusto. Em contrapartida, as arquiteturas baseadas em transformers demonstram uma curva de adoção clara: enquanto o BERT se consolida como padrão dominante a partir de 2021, os modelos generativos (GPT) e menções genéricas a transformers concentram-se majoritariamente nos anos mais recentes da amostra (2023-2024). Essa concentração tardia, oposta à distribuição uniforme das técnicas clássicas, valida estatisticamente a afirmação de que tais modelos representam a tendência futura imediata da área.

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

Técnica ou modelo	Núm. de artigos	Participação	Observação
BERT	23	19 %	Forte presença, indicando uso intensivo de embeddings contextuais pré-treinados.
TF-IDF	22	19 %	Ainda muito utilizado, principalmente em modelos baseados em conteúdo.
LSTM	18	15 %	Popular em tarefas de sequência de texto, como modelagem de preferências.
Word2Vec	14	12 %	Aplicado na vetorização semântica de textos.
CNN	6	5 %	Usado em análises estruturais e extração de características textuais.
Autoencoder	5	4 %	Aplicado na redução de dimensionalidade e modelagem de preferências.
GPT	5	4 %	Indica uma tendência emergente de uso de modelos generativos em recomendações.
Doc2Vec	5	4 %	Variante do Word2Vec focada em documentos.
GRU	4	3 %	Variante mais eficiente que o LSTM, utilizada em alguns modelos sequenciais.
RNN (*recurrent neural network)*	4	3 %	Base tradicional para sequências, sendo superada por LSTM e GRU.
Transformer	4	3 %	Uso crescente, muitas vezes relacionado ao BERT e GPT.
VAE	3	3 %	Empregado em modelos probabilísticos de recomendação.
LDA (*latent Dirichlet allocation*)	2	2 %	Técnica clássica de modelagem de tópicos, com uso decrescente.
NLP (como termo genérico)	2	2 %	Termo amplo, mas usado de forma pouco específica nos artigos.
Embedding (genérico)	1	1 %	Representa o uso de representações vetoriais, mas foi citado genericamente.
Total	**118**	**100 %**	

Tabela 5. Principais técnicas encontradas no corpus 2020-2024
Fonte: elaboração de Denise Fukumi Tsunoda e Patrick Fernandes Rezende Ribeiro com base nos dados produzidos na pesquisa realizada pelos autores, 2025

A análise das técnicas, detalhada na *tabela 5*, evidencia a hegemonia de arquiteturas baseadas em transformers. Cabe ressaltar uma decisão metodológica na contabilização desses dados: embora modelos como BERT e GPT sejam intrinsecamente transformers, optou-se por discriminá-los individualmente para evidenciar a alta frequência de adoção desses modelos pré-treinados específicos. A categoria 'transformer', por sua vez, agrupa os estudos que referenciam a arquitetura de forma genérica ou utilizam variantes com menor representatividade estatística no corpus, tais como RoBERTa (Liu *et al*., 2019), DistilBERT (Sanh *et al*., 2019) ou BART (*bidirectional and auto-regressive transformers*) (Qiu *et al*., 2024).

Assim, considerando os 214 artigos analisados, as tendências observadas são:

1. Domínio dos transformers e modelos pré-treinados, como BERT e GPT, apontando para uma migração de abordagens clássicas, como TF-IDF e Word2Vec, para modelos mais contextuais e precisos;
2. Combinação com técnicas clássicas: muitos artigos ainda integram TF-IDF e Word2Vec como parte de modelos híbridos ou comparações de desempenho;
3. Crescimento do uso de autoencoders variacionais (VAE), sugerindo um interesse em modelagem profunda e representações compactas;
4. Popularidade de modelos sequenciais, como LSTM e GRU, para capturar dinâmicas temporais e contextuais do usuário;
5. Início de aplicação de modelos generativos, como GPT, embora ainda em menor escala, sinalizando uma nova fronteira de personalização.

Com base na análise dos 26 artigos do corpus de 2025, observa-se um panorama atual e emergente sobre o uso de PLN em sistemas de recomendação. As abordagens predominantes refletem a consolidação de técnicas baseadas em aprendizado profundo e a adoção crescente de modelos de grande escala, em conjunto com abordagens clássicas utilizadas como baseline ou em estruturas híbridas.

Entre os métodos mais empregados, destaca-se o BERT, identificado em nove artigos distintos, ou seja, 34.6% dos 26 artigos de 2025. Sua aplicação está centrada na representação contextualizada de textos, permitindo maior acurácia em tarefas como análise de sentimentos, perfis semânticos de usuários e classificação de conteúdos recomendados. Apesar da complexidade computacional, o BERT tem sido preferido por sua capacidade de capturar nuances linguísticas em grande escala. Os títulos dos artigos que mencionam BERT incluem aplicações variadas, como recomendação de empregos, análise de sentimentos, sistemas educacionais e resumos de notícias.

Além disso, a utilização de TF-IDF, em 7 artigos, e Word2Vec, em 2 artigos, demonstra que técnicas clássicas de vetorização ainda têm papel relevante, especialmente em arquiteturas híbridas ou como comparativos experimentais. O uso de modelos tradicionais de aprendizado de máquina, como Random Forest,

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

XGBoost e Naive Bayes, ainda persiste, particularmente em cenários onde interpretabilidade ou baixo custo computacional são prioritários.

Discussões e tendências futuras

A análise qualitativa do corpus de 2025 permitiu identificar uma tendência emergente distinta: a ascensão dos modelos de linguagem de grande escala. Dentre o conjunto de trabalhos recuperados neste período (26), identificou-se que apenas quatro artigos abordam explicitamente essa tecnologia ou arquiteturas generativas conversacionais, razão pela qual foram destacados para esta análise de fronteira. Esses modelos vêm sendo explorados não apenas para geração de texto, mas como mecanismos de recomendação baseados em diálogo, compreensão semântica profunda e adaptação em tempo real ao perfil do usuário. Paralelamente, nos demais trabalhos, observa-se a continuidade do uso de transformers, GRU e embeddings semânticos, indicando um movimento em direção à personalização contextualizada e à robustez de modelos híbridos.

Entre os desafios mais destacados nos artigos estão a explicabilidade dos modelos complexos; a necessidade de integração de dados heterogêneos, sejam textuais, estruturados e interacionais; a escalabilidade dos sistemas baseados em modelos de linguagem de grande escala; e a mitigação de fraudes e manipulações, como ataques de avaliações falsas.

O NotebookLM aponta dois artigos que referenciam o uso de grafos de conhecimento. Um deles, "Knowledge Graphs and Pretrained Language Models Enhanced Representation Learning for Conversational Recommender Systems", de Qiu *et al.* (2024), propõe o KERL (*knowledge-enhanced entity representation learning*), um framework inovador para sistemas de recomendação conversacional (*conversational recommender systems* [CRS]) que integra grafos de conhecimento e modelos de linguagem pré-treinados (*pre-trained language models* [PLMs]), como o BERT e o BART. O objetivo, conforme apresenta o autor, é superar limitações comuns desses sistemas, como incapacidade de lidar com mudanças nos interesses dos usuários e dificuldade em gerar respostas informativas. Para isso, o KERL emprega uma combinação de codificação textual de entidades com redes neurais em grafos, como as *relational graph convolution networks* (RGCNs), além de técnicas como codificação posicional e aprendizado contrastivo, visando capturar a ordem e o contexto semântico dos elementos em uma conversa.

O sistema é avaliado em dois conjuntos de dados, ReDial e INSPIRED, nos quais demonstra desempenho superior em comparação com modelos de referência, como KGSF, C2-CRS e UniCRS. Os resultados evidenciam ganhos expressivos tanto nas tarefas de recomendação (Recall@K) quanto na geração de respostas

(Dist-n, Item Ratio e avaliação humana), evidenciando que a fusão de conhecimento estrutural com representações linguísticas proporciona respostas mais fluentes, contextualizadas e informativas. O trabalho ainda introduz o WikiMKG, um grafo de conhecimento construído a partir da Wikipedia com descrições textuais, e destaca o papel crítico dessas descrições na melhoria das representações semânticas e na personalização das recomendações.

O artigo destaca que, embora o modelo KERL apresente desempenho robusto, a interpretabilidade das recomendações ainda é limitada. Isso dificulta a explicação clara de por que certas recomendações foram feitas, o que é um obstáculo importante para a confiança e adoção prática em ambientes sensíveis ou críticos. Também há menção à necessidade de melhorar a transparência e o controle sobre o raciocínio por trás das recomendações, especialmente ao integrar múltiplas fontes de dados complexas como grafos e texto.

Finalmente, os autores apontam algumas limitações, como a questão da explicabilidade das recomendações geradas, indicando a necessidade de criar mecanismos que possam justificar as decisões do sistema com base nos dados utilizados. Isso contribuiria para mitigar vieses e aumentar a confiabilidade, além de potencializar a exploração de perfis de usuários pré-treinados, visando representar cenários mais próximos da realidade, onde os sistemas de recomendação têm à disposição históricos de interações ou preferências manifestas pelos usuários.

A baixa explicabilidade dos modelos também é citada como lacuna de pesquisa no artigo "An Intelligent Job Recommendation System based on Semantic Embeddings and Machine Learning", de Singla e Verma (2025). O artigo afirma que a abordagem proposta, que combina embeddings semânticos, aprendizado de máquina e várias medidas de similaridade, demonstra o potencial para fornecer recomendações de vagas de emprego "confiáveis, explicáveis e ideais" (520). Os autores mencionam a falta de explicabilidade como uma característica de desvantagem de abordagens como modelos de linguagem de grande escala e sugerem a adição de explicações como um aprimoramento futuro para o sistema proposto, reconhecendo sua importância.

Em síntese, a análise do corpus de 2025 revela que o campo de PLN em sistemas de recomendação segue uma trajetória de evolução híbrida. Observa-se a consolidação massiva de modelos baseados em aprendizado profundo, como BERT e embeddings contextuais, enquanto a incorporação de modelos de linguagem de grande escala se apresenta como uma tendência emergente, identificada qualitativamente nos estudos mais recentes, ainda que numericamente incipiente (quatro trabalhos) se comparada às técnicas tradicionais que proporcionam estabilidade ao ecossistema (22 estudos). O cenário atual aponta, portanto, para uma transição gradual rumo a soluções cada vez mais personalizadas, interpretáveis e adaptativas, com alta integração entre componentes semânticos, computacionais e cognitivos.

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

Os resultados evidenciam a centralidade de técnicas de aprendizado profundo, em especial modelos como BERT, embeddings contextuais e a ascensão de modelos de linguagem de grande escala. Esses achados têm implicações diretas para múltiplos domínios de aplicação identificados neste estudo, incluindo a informática na educação, que aparece como um campo emergente relevante (conforme *tabela 4*). No entanto, ao contrastar o panorama internacional com as bases nacionais consultadas, observa-se uma disparidade: a ausência de registros específicos na RBIE e na SBC-OpenLib sugere que, embora o tema esteja aquecido globalmente, a produção científica brasileira indexada nesses repositórios ainda não incorporou tais técnicas. Essa lacuna local não indica inexistência de interesse, mas sim uma oportunidade estratégica para que pesquisadores nacionais alinhem suas investigações às tendências globais de personalização e mediação tecnológica.

A revisão evidenciou que as técnicas de PLN são transversais a múltiplos domínios. A análise específica das bases nacionais revelou uma oportunidade latente para aplicação dessas técnicas no contexto educacional brasileiro, ainda pouco explorado se comparado ao cenário global. Essa lacuna abre caminho para estudos futuros que articulem os achados aqui discutidos com a personalização de trajetórias formativas e o apoio à tomada de decisão em sistemas educacionais digitais. Os achados têm implicações para ambientes virtuais de aprendizagem, especialmente na curadoria de recursos educacionais, na personalização de trilhas e no apoio à tutoria inteligente. A integração de embeddings contextuais e modelos de linguagem de grande escala pode aprimorar a recomendação de conteúdos e atividades, desde que acompanhada de mecanismos de explicabilidade e controle de vieses.

Ao mapear de forma sistemática os avanços internacionais e contrastá-los com a ausência de contribuições nacionais, este estudo não apenas identifica tendências emergentes, mas também ressalta a oportunidade estratégica para que a comunidade científica brasileira ocupe esse espaço.

Considerações finais

Este estudo buscou responder à seguinte questão de pesquisa: Quais são as principais abordagens e tendências no uso de processamento de linguagem natural (PLN) em sistemas de recomendação? A partir da análise integrada de 214 artigos publicados entre 2020 e 2024, com apoio de ferramentas como ChatGPT, SciSpace, Rayyan, OpenAlex, Zotero, Python e Biblioshiny, foi possível identificar padrões técnicos, temáticos e metodológicos que caracterizam o atual estado da arte na área.

Aponta-se que o período 2020-2025 indica a consolidação de embeddings contextuais (BERT) e a emergência de modelos de linguagem de grande escala em

cenários de recomendação, coexistindo com técnicas clássicas, como TF-IDF e Word2Vec. Pesquisas futuras devem priorizar explicabilidade, cenários educacionais reais e avaliações comparáveis, incluindo custos e latência de modelos.

Observa-se que as abordagens baseadas em conteúdo permanecem amplamente utilizadas, sobretudo em domínios como educação, produtos e entretenimento. No entanto, cresce significativamente a adoção de estratégias híbridas que integram dados comportamentais com representações linguísticas, especialmente aquelas baseadas em modelos de linguagem pré-treinados. Essa tendência é evidenciada pelo uso frequente de modelos como BERT, Word2Vec, Doc2Vec, GPT e autoencoders, que ampliam as possibilidades de personalização e entendimento semântico nos sistemas recomendadores.

As análises apontam também para uma ampla utilização de ferramentas e bibliotecas como TensorFlow, PyTorch, Scikit-Learn, Gensim, Spacy e NLTK, quase sempre associadas à linguagem de programação Python, consolidando-a como padrão técnico dominante. Do ponto de vista empírico, conjuntos de dados amplamente utilizados como MovieLens, Amazon, Yelp e IMDB continuam sendo preferidos para testes e validações, embora isso também indique uma limitação de escopo quanto à diversidade de contextos de aplicação, como moda, saúde ou setor bancário, ainda pouco representados na literatura recente.

Outra descoberta importante refere-se ao papel do PLN nos sistemas de recomendação. Em muitos estudos, o PLN é elemento central da arquitetura proposta, com funções que vão desde a representação semântica de perfis até a geração automática de recomendações. Contudo, em um número considerável de trabalhos, o PLN aparece de forma periférica, sendo utilizado apenas em etapas preliminares, como tokenização ou vetorização básica de textos.

Entre os desafios técnicos mapeados na literatura analisada, destaca-se a questão da explicabilidade (*interpretability*) dos modelos avançados. Conforme evidenciado nos estudos qualitativos de 2025, como os de Qiu *et al*. (2024) e Singla e Verma (2025), a transição para arquiteturas de 'caixa-preta', como aprendizado profundo e modelos de linguagem de grande escala, impõe desafios para a justificativa das recomendações, demandando abordagens de transparência inovadoras. Outro desafio recorrente identificado é a complexidade computacional associada ao treinamento e inferência desses modelos, que impacta a latência em sistemas de recomendação em tempo real. No âmbito metodológico desta revisão, aponta-se como oportunidade futura a expansão do ecossistema de ferramentas de inteligência artificial, ao explorar soluções de orquestração, como LangChain ou LlamaIndex, para ampliar a automação e a profundidade da análise semântica em revisões sistemáticas.

Finalmente, os resultados obtidos neste estudo evidenciam a crescente transversalidade da aplicação de PLN em sistemas de recomendação, apontando

tendências e desafios que impactam múltiplos domínios, desde setores consolidados, como entretenimento e comércio eletrônico, até contextos emergentes como a educação. Ao sistematizar o estado da arte da área, este trabalho contribui para subsidiar o desenvolvimento tecnológico e a tomada de decisão em sistemas inteligentes, independentemente do domínio de aplicação.

Agradecimentos

Os autores expressam seu reconhecimento à Universidade Federal do Paraná (UFPR), ao Programa de Posgrado em Ciência de Dados (PPGCD/UFPR) e ao Programa de Posgrado em Gestão da Informação (PPGGI/UFPR) pelo apoio institucional brindado durante o desenvolvimento desta investigação. Asimismo, reconhecemos a colaboração dos colegas e estudantes vinculados ao Laboratório de Excelência em Inteligência Artificial (LexIA), que contribuem com as valiosas discussões acadêmicas e apoiam a organização dos dados. Finalmente, agradecemos ao patrocínio e apoio ao financiamento do Instituto de Ciência, Tecnologia e Inovação (ICTi) / Itaú, que tornou possível a realização deste trabalho.

Referências

Baker, Ryan Shaun Joazeiro de, Seiji Isotani e Adriana Maria Joazeiro Baker de Carvalho. 2011. "Mineração de dados educacionais: oportunidades para o Brasil". *Revista Brasileira de Informática na Educação* 19 (2): 3-13. https://doi.org/10.5753/RBIE.2011.19.02.03

Bazzan, Jordana, Márcia Elisa Echeveste, Carlos Torres Formoso, Bernardo Altenbernd e Márcia Helena Barbian. 2023. "An Information Management Model for Addressing Residents' Complaints Through Artificial Intelligence Techniques". *Buildings* 13 (3), e737. https://doi.org/10.3390/buildings13030737

Bojanowski, Piotr, Edouard Grave, Armand Joulin e Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information". *Transactions of the Association for Computational Linguistics* 5: 135-46. https://doi.org/10.1162/tacl_a_00051

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan *et al*. 2020. "Language Models are Few-Shot Learners". Pré-publicação Arxiv. https://doi.org/10.48550/arXiv.2005.14165

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk e Yoshua Bengio. 2014. "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation". Em *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, organizado por Alessandro Moschitti, Bo Pang e Walter Daelemans, 1724-34. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

Devlin, Jacob, Ming-Wei Chang, Kenton Lee e Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Pré-publicação Arxiv.
https://doi.org/10.48550/arXiv.1810.04805

Gugnani, Akshay, e Hemant Misra. 2020. "Implicit Skills Extraction Using Document Embedding and its Use in Job Recommendation". *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (8): 13286-93.
https://doi.org/10.1609/aaai.v34i08.7038

Hochreiter, Sepp, e Jürgen Schmidhuber. 1997. "Long Short-Term Memory". *Neural Computation* 9 (8): 1735-80.
https://doi.org/10.1162/neco.1997.9.8.1735

Kim, Yoon. 2014. "Convolutional Neural Networks for Sentence Classification". Em *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, organizado por Alessandro Moschitti, Bo Pang, Walter Daelemans, 1746-51. Association for Computational Linguistics.
https://doi.org/10.3115/v1/D14-1181

Kingma, Diederik P., e Max Welling. 2014. "Auto-encoding Variational Bayes". Pré-publicação Arxiv.
https://doi.org/10.48550/arXiv.1312.6114

Le, Quoc V., e Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents". Pré-publicação Arxiv.
https://doi.org/10.48550/arXiv.1405.4053

Lin, Jianghao, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu *et al.* 2024. "How Can Recommender Systems Benefit from Large Language Models: A Survey". Pré-publicação Arxiv.
https://doi.org/10.48550/arXiv.2306.05817

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Springer.
https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy *et al.* 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". Pré-publicação Arxiv.
https://arxiv.org/abs/1907.11692

Manning, Christopher D., Prabhakar Raghavan e Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
https://doi.org/10.1017/CBO9780511809071

Mikolov, Tomas, Kai Chen, Greg Corrado e Jeffrey Dean. 2013a. "Efficient Estimation of Word Representations in Vector Space". Pré-publicação Arxiv.
https://doi.org/10.48550/arXiv.1301.3781

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado e Jeffrey Dean. 2013b. "Distributed Representations of Words and Phrases and their Compositionality". Pré-publicação Arxiv.
https://doi.org/10.48550/arXiv.1310.4546

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D Mulrow, Larissa Shamseer *et al.* 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews". *BMJ* 372, 71.
https://doi.org/10.1136/bmj.n71

Pennington, Jeffrey, Richard Socher e Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation". Em *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, organizado por Alessandro Moschitti, Bo Pang e Walter Daelemans, 1532-43. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Pereira, Aluisio José, Alex Sandro Gomes e Tiago Thompsen Primo. 2022. "Design de sistema e recomendação educacional: abordagens com Mágico de Oz". Em *Anais do XXXIII Simpósio Brasileiro de Informática na Educação (SBIE 2022)*, 1184-95. Sociedade Brasileira de Computação. https://doi.org/10.5753/sbie.2022.225760

Pires, Pedro R., Bruna B. Rizzi e Thiago A. Almeida. 2024. "Why Ignore Content? A Guideline for Intrinsic Evaluation of Item Embeddings for Collaborative Filtering". Em *Brazilian Symposium on Multimedia and the Web (Webmedia)*, 345-354. Sociedade Brasileira de Computação. https://doi.org/10.5753/webmedia.2024.243199

Qiu, Zhangchi, Ye Tao, Shirui Pan e Alan Wee-Chung Liew. 2024. "Knowledge Graphs and Pretrained Language Models Enhanced Representation Learning for Conversational Recommender Systems". *IEEE Transactions on Neural Networks and Learning Systems* 36 (4): 6107-21. https://doi.org/10.1109/tnnls.2024.3395334

RBIE (*Revista Brasileira de Informática na Educação*). s. d. About the Journal. Acessado em 19 de março, 2026. https://journals-sol.sbc.org.br/index.php/rbie

Ricci, Francesco, Lior Rokach e Bracha Shapira, orgs. 2015. *Recommender Systems Handbook*, 2.º ed. Springer. https://doi.org/10.1007/978-1-4899-7637-6

Salton, Gerard, e Christopher Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval". *Information Processing & Management* 24 (5): 513-23. https://doi.org/10.1016/0306-4573(88)90021-0

Sanh, Victor, Lysandre Debut, Julien Chaumond e Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter". Pré-publicação Arxiv. https://doi.org/10.48550/arXiv.1910.01108

SBC (Sociedade Brasileira de Computação). s. d. SBC OpenLib (SOL). Acessado em 19 de março, 2026. https://sol.sbc.org.br

Shaikh, Aryaan, Nikita Newalkar, Sakshi Gaikwad, Namrata Kadav e Chaitali Shewale. 2023. "Autocomplete Recommendation Plugin and Summarizing Text Using Natural Language Processing". *Journal of Innovation Information Technology and Application (JINITA)* 5 (2): 114-23. https://doi.org/10.35970/jinita.v5i2.1912

Singla, Priyanka, e Vishal Verma. 2025. "An Intelligent Job Recommendation System Based on Semantic Embeddings and Machine Learning". *Journal of Information Systems Engineering and Management* 10 (5s): 520-42. https://doi.org/10.52783/jisem.v10i5s.681

Velpula, Koteswara Rao, Hema Pavuluri, Poojitha Neeluri, Anushka Pappala e Mounika Narra. 2024. "Recommendation System for Code Validation and Optimal Refactoring". *International Journal of Advanced Research in Computer and Communication Engineering* 13 (3): 80-87. https://doi.org/10.17148/IJARCCE.2024.13313

DOI: https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106

Yang, Yixiao. 2022. "Improving the Robustness to Data Inconsistency Between Training and Testing for Code Completion by Hierarchical Language Model". Pré-publicação Arxiv.
https://arxiv.org/abs/2003.08080v2

Yang, Fan, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang e Yanbin Lu. 2023. "PALR: Personalization Aware LLMs for Recommendation". Pré-publicação Arxiv.
https://arxiv.org/abs/2305.07622

Para citar este texto:

Tsunoda, Denise Fukumi, Patrick Fernandes Rezende Ribeiro, Juliane de Lima Pires, Kamilly Voitkiv Hubner, Matheus Henrique Assumpção dos Reis, Patrick Alves Bastos e Roberto Rigo. 2026. "Sistemas de recomendação e processamento de linguagem natural: uma revisão estruturada e tendências emergentes com o suporte de ferramentas de inteligência artificial". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 40 (106): 79-108.
https://dx.doi.org/10.22201/iibi.24488321xe.2026.106.59106