A Comprehensive Review on Multilingual News Recommender Systems and their Challenges

Siddhant
M.Tech Scholar, CSE Department, UIET, MDU, ROHTAK

How to Cite this Article:

Siddhant, (2026). A Comprehensive Review on Multilingual News Recommender Systems and their Challenges. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.429

https://doi.org/10.55041/ijcope.v2i5.429

ABSTRACT: Multilingual News Recommender Systems (MNRS) aim to deliver personalized news across different languages, addressing the limitations of traditional English-centric recommendation approaches. With the rapid growth of regional digital news consumption, especially in linguistically diverse countries like India, multilingual recommendation has become essential for user engagement. . Recent advances in natural language processing, including multilingual transformers such as mBERT, XLM-RoBERTa, and LASER embeddings, have enabled improved cross-lingual semantic understanding. However, challenges such as low-resource languages, lack of multilingual datasets, inconsistent translation quality, semantic drift, script variation, and code-mixed text continue to restrict the effectiveness of MNRS. This review paper provides a comprehensive analysis of existing techniques, datasets, and evaluation methods used in multilingual news recommendation, highlighting their strengths and limitations. The study also identifies core research gaps and outlines future directions for building more accurate, inclusive, and real-world-ready multilingual news recommender system.

KEYWORDS: Multilingual News Recommender Systems (MNRS), Cross-Lingual Transfer, Low-Resource Languages, Multilingual NLP, Monolingual Bias

1.INTRODUCTION

The rapid expansion of digital journalism has led to an overwhelming increase in multilingual news content available online. Users increasingly consume news in their preferred regional languages, especially in linguistically diverse countries such as India, where hundreds of languages and dialects coexist. As a result, personalized news delivery has become a critical requirement for improving user experience and content relevance. News Recommender Systems (NRS) address this challenge by predicting and suggesting news articles that match user interests. While traditional NRS rely heavily on English datasets and monolingual embeddings, modern multilingual recommendation requires deeper semantic understanding across languages, scripts, and writing styles.

Recent advancements in multilingual natural language processing (NLP) — particularly transformer-based models such as mBERT, XLM-RoBERTa, and LASER — have improved cross-lingual text representation, enabling more accurate multilingual news recommendation. These models help capture semantic similarities across languages without relying on machine translation alone. Despite such progress, creating effective multilingual news recommendation remains challenging due to scarcity of multilingual datasets, regional language diversity, code-mixing, inconsistent grammar structures, and low-resource languages.

This review paper analyzes existing multilingual news recommendation techniques, examines their strengths and drawbacks, compares key multilingual NLP models, and highlights major gaps that still limit large-scale deployment. The review aims to provide a clear understanding of current advancements and offer future research directions for building more robust, inclusive, and language-agnostic news recommender systems

2. BACKGROUND

News Recommender Systems (NRS) have become an integral part of modern digital news platforms, enabling personalized content delivery to users based on their reading behaviors, preferences, and contextual information. Unlike traditional recommendation domains such as movies or e-commerce—where item lifecycles are long and user preferences are relatively stable—the news domain is highly dynamic. News articles rapidly become outdated, user interests shift frequently, and new events emerge continuously. These characteristics make news recommendation a challenging yet essential task in the era of information overload.

A traditional News Recommender System typically relies on one of two foundational approaches. Content-Based Filtering (CBF) recommends articles similar to those a user has previously interacted with, based on textual features such as keywords, topics, TF-IDF vectors, or more recently, word embeddings and transformer-based representations. Collaborative Filtering (CF), on the other hand, focuses on user–user or item–item similarity by leveraging historical interaction patterns, such as clicks, likes, or reading time. While CBF excels in handling new articles and capturing textual relevance, it suffers from over-specialization. CF, conversely, can uncover latent user preferences but struggles with sparse data and new users.

To overcome these limitations, hybrid approaches combining content and collaborative signals emerged, integrating semantic article features with implicit user feedback. The rise of deep learning further transformed news recommendation through Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, attention mechanisms, and transformer-based models such as BERT and its multilingual variants. These methods significantly enhanced the system's ability to model long-range dependencies, extract nuanced semantic representations, and capture evolving user interests.

However, the news recommendation landscape becomes far more complex in multilingual environments. In linguistically diverse countries like India, users consume news in multiple regional languages including Hindi, Tamil, Telugu, Bengali, Marathi, Malayalam, Punjabi, and others. Each language differs in script, grammar, morphology, vocabulary, and writing style, making cross-lingual text modeling inherently challenging. Moreover, many regional languages are low-resource, meaning they lack large annotated datasets required to train robust machine learning models.

To address multilingual text understanding, researchers increasingly rely on multilingual natural language processing (NLP) models, most notably mBERT (Multilingual BERT), XLM-RoBERTa, LASER embeddings, IndicBERT, and MuRIL. These models can represent multiple languages within a shared

semantic space, enabling cross-language similarity and classification without explicit translation. Such models are particularly useful for multilingual news recommendation, where the system must recognize semantic equivalence across languages—for example, a political news article in Hindi and a similar one in English.

Despite these advancements, several challenges persist. Multilingual datasets for news recommendation are scarce, and existing datasets like MIND, Adressa, and Plista are predominantly English or European-language oriented. Indian languages also present unique issues such as code-mixing, where users blend English with regional languages (e.g., Hinglish: "aaj match kaun jeetega?"), significantly complicating preprocessing and semantic modeling. Script diversity across Indian languages further amplifies tokenization and embedding challenges. Additionally, multilingual recommendation introduces concerns of cultural bias, fairness, and regional relevance, as models trained on English-dominant corpora may inadequately capture local sentiment, context, and social meaning.

In summary, the background of multilingual news recommendation encompasses a complex interplay of recommender system techniques, cross-lingual NLP advancements, linguistic diversity, and dynamic user behavior. Understanding these foundations is essential for analyzing existing research, identifying current limitations, and exploring the need for more inclusive and linguistically aware news recommender systems suitable for diverse cultural and linguistic populations

3. LITERATURE REVIEW

The evolution of news recommender systems has progressed through several distinct phases, beginning with early traditional recommendation techniques, advancing into deep-learning-based architectures, and finally transitioning into the modern era of multilingual and cross-lingual news recommendation. A chronological understanding of these developments is essential to contextualize current multilingual approaches.

Early research on recommender systems relied heavily on Content-Based Filtering (CBF) and Collaborative Filtering (CF) methods, which used keywords, TF-IDF vectors, and user-item interaction matrices to produce recommendations. Foundational studies in the late 1990s and early 2000s established these paradigms, laying the groundwork for subsequent news recommendation research. CBF methods were effective for analyzing news articles due to their high textual density, while CF methods captured user behavior patterns. However, both approaches suffered from limitations such as sparse data, overspecialization, and inability to handle dynamic news content (Adomavicius & Tuzhilin, 2005; Ricci et al., 2011).

The shift toward deep neural architectures began in the mid-2010s, with models designed to capture the contextual richness and temporal sensitivity of news articles. One of the earliest significant contributions in document modeling was the Hierarchical Attention Network (HAN) by *Yang et al. (2016)*, which introduced hierarchical attention for learning document semantics. This approach laid a strong foundation for subsequent news text encoders. The introduction of DKN (Deep Knowledge-Aware Network) by *Wang et al. (2018)* marked a step forward by integrating knowledge graphs with CNN-based representations, enhancing semantic richness but remaining monolingual.

The field saw major advancements in 2019, when several influential English-language news recommenders were introduced. NRMS (Neural News Recommendation with Multi-Head Self-Attention) by *Wu et al.* utilized self-attention to capture semantic dependencies within articles, while NPA (Neural Personalized Attention) by the same authors introduced personalized attention mechanisms to model user-specific reading patterns. Also in 2019, LASER (Language-Agnostic Sentence Representations) by *Artetxe et al.* provided multilingual sentence embeddings covering over 90 languages, although not specifically optimized for news recommendation.

With the increasing availability of multilingual data, multilingual transformer models began to shape the field. In 2020, *Conneau et al.* introduced XLM-RoBERTa, trained on large multilingual corpora across 100+ languages, enabling cross-lingual text understanding. In parallel, *Kakwani et al. (2020)* developed IndicBERT, an efficient transformer optimized for 12 major Indian languages. These models enabled researchers to explore cross-lingual transfer for news recommendation, though neither was originally trained on news-domain data.

Significant improvements continued in 2021, with *Khanuja et al.* introducing MuRIL, a multilingual model developed specifically for 17 Indian languages, improving representation for regional content. *Chen et*

al. proposed cross-lingual behavior transfer techniques to model user interactions across English, Spanish, and German. *Huang et al.* advanced semantic modeling via graph neural networks, suggesting entity-level relationships could enhance news recommendation — although multilingual extensions remained limited.

By 2022, multilingual and cross-lingual news recommendation had become an active research area. *Das et al.* proposed a simple English–Bengali hybrid recommendation model using TF-IDF, embeddings, and translation. *Park et al.* examined ideological bias and fairness issues, while *Singh et al.* adapted BERT for English–Hindi cross-lingual clustering. *Qi et al.* introduced DeepFusion, a multi-view fusion architecture combining textual, contextual, and metadata features.

A major breakthrough occurred in 2023 when *Qi et al.* evaluated zero-shot cross-lingual recommendation using mBERT, showing significant performance drops without target-language data. Most importantly, xMIND (MIND Your Language) by *Wu et al. (2023)* introduced the first large-scale multilingual news recommendation dataset, covering 14 languages through machine translation. This dataset enabled systematic evaluation of multilingual recommenders under zero-shot, few-shot, and multilingual conditions, and revealed substantial challenges in cross-lingual semantic alignment.

The field matured rapidly in 2024. *Xia et al.* introduced NaSE (News-Adaptive Sentence Encoder), a domain-adapted multilingual embedding model trained on multilingual news corpora. NaSE significantly outperformed general-purpose multilingual encoders like XLM-R for cross-lingual news similarity. Meanwhile, *Li et al.* published a comprehensive survey on Large Language Models (LLMs) for news recommendation, highlighting their potential for multilingual reasoning, contextual comprehension, and user preference modeling.

Recent work in 2025 has focused on addressing the low-resource language challenge. PPT (Preference Pattern Transfer) by *Li et al. (2025)* proposed a two-tower cross-lingual model capable of transferring reading behavior from high-resource to minor languages, demonstrating strong improvements for low-resource news domains. Additional studies extended multilanguage personalization through cross-lingual alignment, demonstrating progress but also emphasizing the persistent scarcity of regional-language datasets.

Overall, the literature demonstrates a transition from traditional filtering techniques to advanced neural architectures, and more recently to multilingual and cross-lingual models capable of supporting linguistically diverse populations. Despite progress, challenges such as low-resource language support, code-mixing, cultural nuance, data scarcity, and real-world multilingual behavior modeling continue to hinder robust multilingual news recommendation.

Comparison table of Major Approaches and Developments in News Recommender Systems (Year-wise)

Year	Authors	Title	Approach / Method	Key Contribution
2016	*Yang et al.*	*Hierarchical Attention Networks*	*Attention-based hierarchical text encoding*	*Introduced hierarchical attention for document modelling ; foundation for news representation*
2018	*Wang et al.*	*DKN: Deep Knowledge-Aware Network*	*CNN + Knowledge Graph*	*Integrated entity-level knowledge with news text; improved semantic relevance in monolingual news recommenders*
2019	*Wu et al.*	*NRMS: Neural News Recommendation with Multi-Head*	*Multi-head self-attention*	*Enhanced semantic extraction from news content; strong English-language baseline*

		Self-Attention		
2020	*Conneau et al.*	*XLM-RoBERTa*	*Multilingual Transformer*	*Enabled cross-lingual semantic representation across 100+ languages; foundation for multilingual news recommendation*
2021	*Khanuja et al.*	*MuRIL: Multilingual Representations for Indian Languages*	*Indian-language transformer model*	*Improved embeddings for 17 Indian languages; beneficial for multilingual/regional news contexts*
2023	*Wu et al.*	*MIND Your Language (xMIND)*	*Multilingual dataset (machine-translated)*	*First multilingual benchmark dataset for news recommendation covering 14 languages*
2024	*Xia et al.*	*News Without Borders (NaSE)*	*Domain-adapted multilingual news encoder*	*Achieved superior cross-lingual performance using news-specific multilingual embeddings*

4. RESEARCH GAP

Although significant progress has been made in news recommendation—from early neural architectures to advanced multilingual and cross-lingual systems—several critical research gaps remain evident in the existing literature. A major limitation observed across studies is the lack of real multilingual and regionally diverse datasets, as most available corpora, including xMIND (Wu et al., 2023), are generated through machine translation rather than native multilingual content, resulting in loss of cultural context, semantic drift, and inconsistent terminology. Furthermore, while multilingual models such as XLM-R (Conneau et al., 2020), IndicBERT (Kakwani et al., 2020), and MuRIL (Khanuja et al., 2021) provide strong cross-lingual representations, they suffer from domain mismatch, as they are not trained specifically on news data; only NaSE (Xia et al., 2024) attempts domain adaptation, yet its performance remains limited for low-resource languages. Despite the existence of large multilingual embedding models, low-resource and regional languages—particularly Indian languages—remain under-represented, with only a few recent approaches like PPT (Li et al., 2025) attempting to address minor-language cold-start issues. Another critical gap is the absence of robust solutions for code-mixing and mixed-script text, such as Hinglish or Tamil-English blends, which dominate real-world digital news consumption but are not addressed by any of the reviewed systems. Moreover, cross-lingual user modeling remains weak; most models assume monolingual users, while real users often switch languages fluidly, and existing works like Chen et al. (2021) and PPT (2025) provide only partial solutions. The field further suffers from the lack of standardized evaluation protocols for multilingual recommendations, as current benchmarks rely on artificial translations and fail to reflect multilingual user behavior or mixed-language contexts. Finally, concerns related to fairness, cultural bias, ideological bias, and real-time performance remain largely unexplored, with only minimal research addressing political bias (Park et al., 2022), and no studies evaluating fairness or temporal latency in multilingual environments. Collectively, these gaps highlight the need for more comprehensive, culturally grounded, and linguistically inclusive research to enable reliable real-world multilingual news recommendation systems

5. CONCLUSION

The evolution of news recommender systems over the past decade demonstrates a clear transition from traditional filtering methods to sophisticated neural architectures and, more recently, to multilingual and cross-lingual recommendation approaches. Early models focused primarily on monolingual environments and relied heavily on content-based or collaborative filtering techniques. With the rise of deep learning, attention-based models and neural user–news encoders significantly improved semantic understanding and personalization. However, as multilingual digital news consumption increased—especially in linguistically diverse regions like India—the limitations of monolingual systems became apparent. Recent advances such as multilingual transformers (XLM-R, IndicBERT, MuRIL), multilingual datasets (xMIND), domain-adapted encoders (NaSE), and low-resource solutions (PPT) mark substantial progress toward bridging linguistic gaps, yet the literature shows persistent shortcomings. These include a lack of real multilingual datasets, weak support for low-resource regional languages, poor handling of code-mixed content, inadequate cross-lingual user modeling, limited standardized evaluation protocols, and underexplored fairness and real-time performance challenges. Overall, while research in multilingual news recommendation has expanded rapidly in recent years, significant opportunities remain for developing more inclusive, culturally grounded, and robust systems capable of supporting diverse multilingual user populations in real-world environments.

6.REFERENCES

[1]G. Adomavicius & A. Tuzhilin. Toward the Next Generation Of Recommender Systems: A Survey of the state-of-the-art and Possible Extensions. https://ieeexplore.ieee.org/document/1423975.

[2] Artetxe, M., Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7, 597–610.

[3] Chen, Y., Zhang, H., & Zhou, X. (2021). Cross-lingual user behavior modeling for news recommendation. Proceedings of the 30th ACM International Conference on Information and Knowledge Management.

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

[5] Das, S., Ghosh, S., & Banerjee, S. (2022). A hybrid multilingual recommendation model for English-Bengali news. International Journal of Information Management, 63, 102455.

[6] Huang, C., Feng, F., & Zhao, D. (2021). Graph-based news recommendation with contextualized entity representations. Proceedings of the Web Conference (WWW).

[7] Kakwani, D., Kunchukuttan, A., et al. (2020). IndicBERT: A multilingual ALBERT model for Indian languages. arXiv preprint arXiv:2009.08701.

[8] Khanuja, S., Bapna, R., et al. (2021). MuRIL: Multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730.

[9] Li, R., Zhang, Y., & Chen, Q. (2024). A survey on LLM-based news recommender systems. arXiv preprint arXiv:2402.09797.

[10] Li, T., Wang, X., & Liu, Y. (2025). PPT: Preference pattern transfer for minor-language news recommendation. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.

[11] Park, J., Lee, S., & Kim, H. (2022). Fairness and bias in news recommendation systems: An empirical analysis. ACM Transactions on Recommender Systems.

[12] Qi, T., Zhang, Y., & Wu, Q. (2022). DeepFusion: A multi-view deep learning framework for news recommendation. Proceedings of the 45th International ACM SIGIR Conference.

[13] Qi, Z., Li, Y., & Chen, Z. (2023). Zero-shot cross-lingual news recommendation using multilingual BERT. IEEE Access, 11, 45387–45399.

[14] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. Springer.

[15] Singh, A., Gupta, M., & Sharma, V. (2022). Cross-lingual BERT-based clustering for English-Hindi news. Journal of Intelligent Information Systems.

[16] Wang, H., Zhang, F., & Wang, J. (2018). DKN: Deep knowledge-aware network for news

recommendation. Proceedings of the 27th International Conference on World Wide Web (WWW).

[17] Wu, C., Wu, F., & Xie, X. (2019). Neural news recommendation with multi-head self-attention. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

[18] Wu, C., Wu, F., & Xie, X. (2019). NPA: Neural personalized attention model for news recommendation. Proceedings of the 25th ACM SIGKDD Conference.

[19] Wu, Y., Qi, T., & Wu, Q. (2023). MIND Your Language: A multilingual dataset for cross-lingual news recommendation. Proceedings of the 46th ACM SIGIR Conference.

[20] Xia, J., Li, X., & Feng, F. (2024). News Without Borders: Domain-adaptive multilingual news sentence encoder. Transactions of the ACL.

[21] Yang, Z., Yang, D., & Zhou, B. (2016). Hierarchical attention networks for document classification. Proceedings of NAACL-HLT