Evaluation Metrics for Deep Learning–Based Semantic Search: A Critical Review with a User Satisfaction Perspective

ADESANYA Adetola Joel, AYOADE Akintayo Michael, Folarin Israel Bolaji

Department of Computer Science, Lead City University, Ibadan, Oyo State, Nigeria.

DOI: https://dx.doi.org/10.51244/IJRSI.2026.1305000044

Received: 04 May 2026; Accepted: 16 May 2026; Published: 26 May 2026

ABSTRACT

Background: Semantic search, driven by deep learning models like BERT and Sentence-BERT (SBERT), has greatly improved information retrieval. It has shifted from matching keywords to capturing the context of search and user intent. However, to evaluate how effective these systems are, traditional system-focused metrics such as precision, recall, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG) are still used. These metrics do not adequately reflect user experience. They often overlook important behavioral and contextual factors such as user engagement, search satisfaction, relevance perception, and interaction quality in real-world environments. This review examines existing evaluation metrics for deep learning-based semantic search. It identifies their strengths and limitations, as well as how well they capture real-world user satisfaction. It also explores helpful ways to incorporate user-centered approaches into the evaluation of these systems.

Method: A critical review approach was used, synthesizing literature from 2020 to 2025 across databases like IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. Studies on semantic search evaluation, deep learning-based retrieval, and user-centered metrics were thematically analyzed for information. The reviewed studies were selected using predefined inclusion and exclusion criteria, and the analysis categorized evaluation methods into traditional and user-centered approaches.

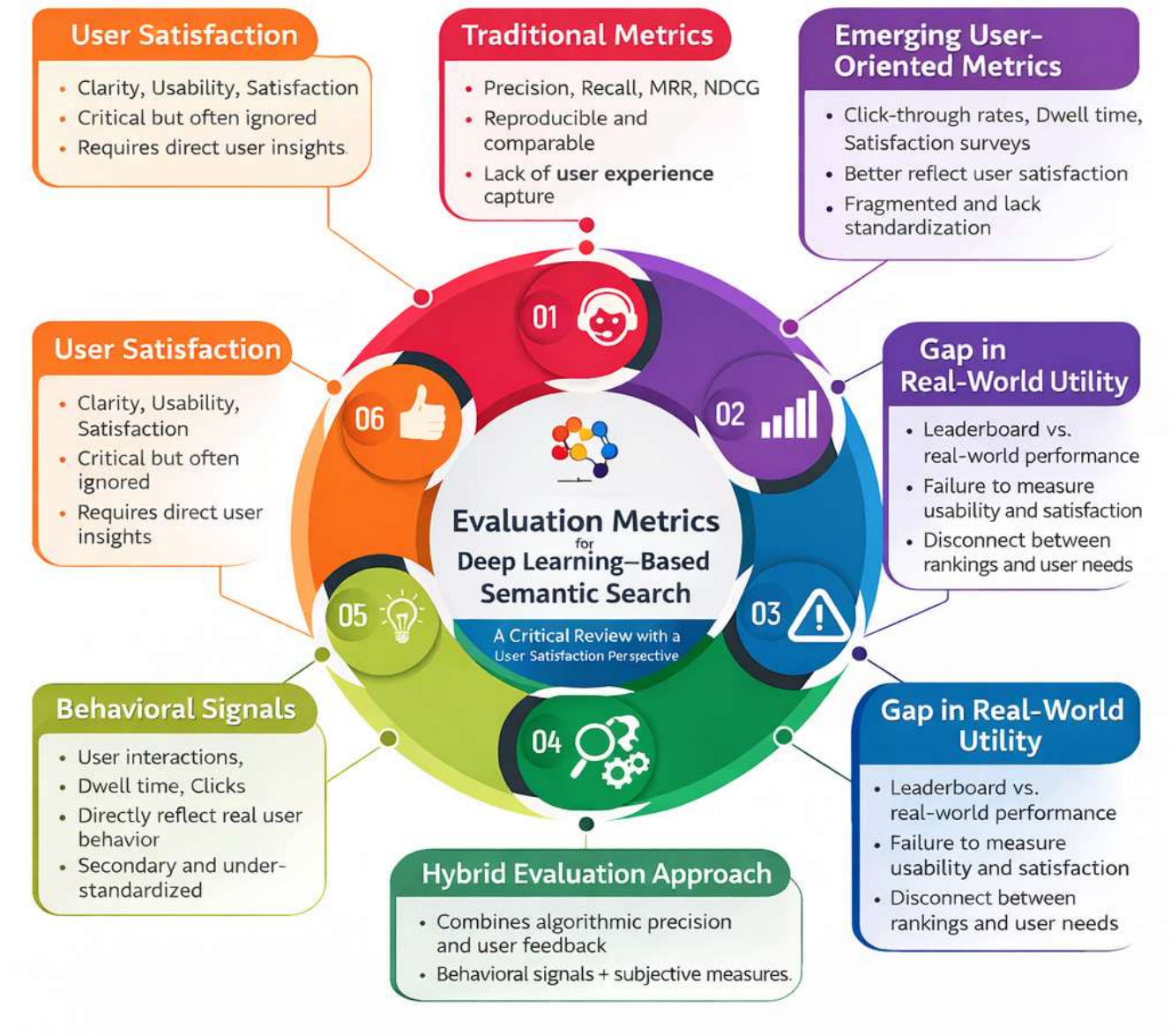
Findings: The review finds that while traditional metrics provide reproducibility and comparability, they fail to capture important aspects of user experience such as clarity, usability, and satisfaction. Emerging user-oriented alternatives like click-through rates, dwell time, and satisfaction surveys offer valuable insights, but they remain secondary, fragmented, and lack standardization. The review highlights an ongoing gap between the leaderboard performance of search systems and their real-world utility. The review further reveals that many high-performing semantic retrieval systems achieve strong benchmark scores while still failing to fully satisfy users in practical search scenarios.

Conclusion: Semantic search evaluation must change from traditional, system-focused measures to hybrid metrics that integrate algorithmic precision with user-centered awareness. By combining these traditional metrics with behavioral signals and subjective feedback, future evaluation methods can ensure that semantic search systems are not only technically sound but also practical, usable, and satisfying for end-users. The study therefore recommends the development of standardized hybrid evaluation frameworks capable of balancing retrieval accuracy with measurable user experience indicators.

Keywords: Semantic search, information retrieval, evaluation metrics, deep learning, user satisfaction.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Infograph



INTRODUCTION

The field of Information Retrieval (IR) has changed significantly over the past few decades. It has moved from keyword-based retrieval systems to more advanced semantic search models that take advantage of improvements in natural language processing (NLP) and deep learning [1]. Early IR models, like the Boolean and vector space models, focused on the overlap of words between queries and documents [2]. While these models performed effectively in limited domains, they often had trouble understanding user intent, dealing with synonyms and word meanings, and addressing the subtle differences in natural language. The rise of semantic search helped to overcome some of these challenges by shifting focus from simple keyword matching to understanding meanings and context [3]. Deep learning methods, especially transformer-based models like BERT and Sentence-BERT (SBERT), have transformed semantic search even more. They allow systems to handle complex language structures and provide more relevant, context-aware results on a large scale [4].

As semantic search technologies continue to improve, evaluation has become a key issue. Evaluation metrics play a central role in driving progress within information retrieval, providing the standards against which new models are tested and compared. Traditional metrics such as precision, recall, mean average precision (MAP), normalized discounted cumulative gain (NDCG), and mean reciprocal rank (MRR) remain the primary standards [3,5]. These measures offer consistent, system-focused evaluations and have been vital in encouraging

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

algorithmic innovation. However, they were originally designed for static, labeled datasets and focus on document relevance from a computational viewpoint rather than from the perspective of actual users [6].

This reliance on system-focused benchmarks has created an important gap. While semantic search systems show strong performance in controlled evaluations, their effectiveness in real-world situations often remains uncertain. Users evaluate a search system not only on technical accuracy but also on how useful, understandable, and contextually relevant the retrieved results are [7]. A document considered "relevant" by traditional metrics may still fail to satisfy a user if it is too technical, poorly structured, or misaligned with their underlying intent. As a result, existing evaluation methods risk ignoring factors like usability, satisfaction, and task success, which are increasingly seen as essential to meaningful information retrieval [8].

This paper aims to provide a critical review of evaluation metrics in the context of deep learning-based semantic search. It highlights the strengths and weaknesses of current algorithmic approaches, identifies gaps in capturing user experience, and emphasizes the need for user-focused evaluation frameworks. By integrating insights from both computational metrics and human-centered measures, this review contributes to the ongoing discussion on how to better assess the effectiveness of semantic search in practice [7]. The research process adopted in this study is illustrated in Figure 1, which presents the study flowchart and outlines the stages of the review methodology.

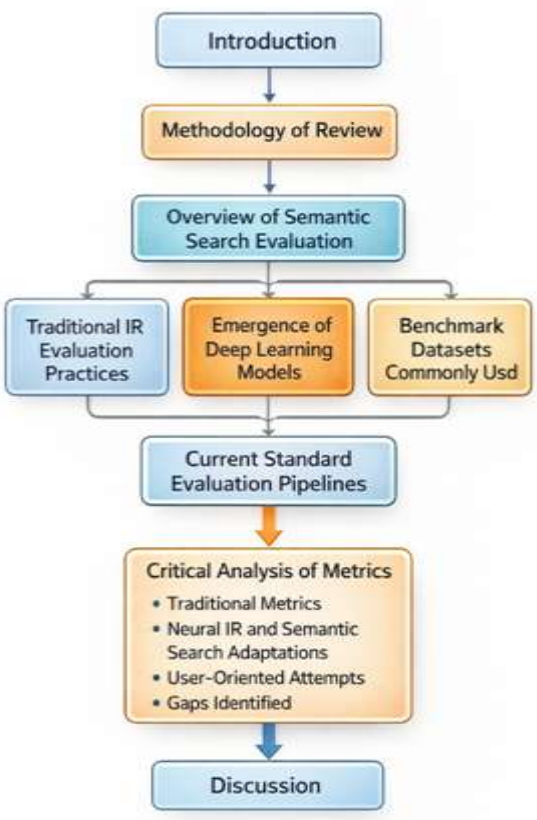


Figure 1: Study Flowchart

Methodology of Review

This study uses a critical review approach to examine evaluation metrics in deep learning-based semantic search, focusing on how well they capture user experience. Unlike systematic reviews that follow strict protocols to cover all available literature, a critical review is more analytical and allows for an in-depth exploration of concepts, theories, and emerging practices related to the research issue [9].

The literature was gathered from key academic and technical sources such as IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. These databases were chosen because they include leading journals and

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

conferences in computer science, artificial intelligence, and information retrieval, ensuring access to both foundational studies and recent developments [10].

A mix of keywords and Boolean operators helped retrieve relevant publications. Main search terms included "semantic search," "evaluation metrics," "information retrieval evaluation," "deep learning IR," and "user satisfaction in search systems." These terms were used both alone and in various combinations to enhance coverage while keeping relevance to the research topic [11].

To ensure quality and relevance, the review focused on peer-reviewed articles, conference papers, and significant technical reports published in the last 5 years (2020-2025). Studies were included if they discussed evaluation methods for semantic search, deep learning-based retrieval, or user-centered approaches to information retrieval assessment. Preference was given to studies with empirical findings, comparative evaluations, or applied implementations related to semantic retrieval systems. Publications that were purely theoretical, lacked empirical support, or did not directly discuss evaluation practices were excluded. Similarly, opinion pieces, non-peer-reviewed preprints, and articles centered only on traditional keyword-based information retrieval with no relevance to semantic or deep learning contexts were also left out. After the screening and filtering process, a total of 33 studies were selected for detailed review and analysis [12].

The chosen literature was analyzed thematically and critically, considering both technical and user-centered perspectives. The thematic analysis involved grouping the selected studies into major categories such as traditional retrieval metrics, behavioral evaluation metrics, user satisfaction indicators, and hybrid evaluation approaches. The review did not seek to cover everything; instead, it prioritized works that provided significant insights into the development of evaluation metrics, pointed out gaps in existing methods, or suggested user-oriented frameworks. The findings were synthesized to identify common strengths and weaknesses in current evaluation practices and to outline ways to incorporate user experience into assessing semantic search systems [13]. Figure 2 presents the methodology flowchart for this review process.

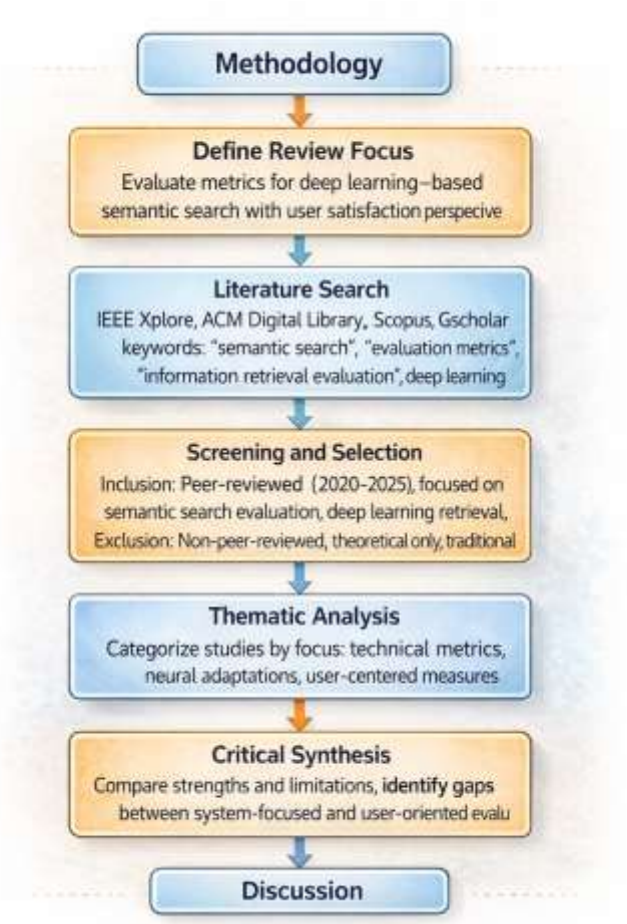


Figure 2: Methodology Flowchart

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Overview of Semantic Search Evaluation

Traditional IR Evaluation Practices

Evaluation has always been important for the development of information retrieval (IR). In traditional IR, performance was mainly measured using system-focused, relevance-based metrics that compared retrieved documents to set ground truth labels [1,14]. Some of the most common metrics include precision, which is the proportion of relevant retrieved documents [7], recall, the proportion of relevant documents that were successfully retrieved, and the F1-score, which is their harmonic mean. For ranked retrieval tasks, more complex measures like Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), and Mean Reciprocal Rank (MRR) became standard [2]. These metrics offered consistent, quantitative benchmarks for comparing algorithms and have been crucial in advancing IR research. However, they were created for static datasets with binary or graded relevance judgments and tend to prioritize algorithm efficiency over user experience [15]. Figure 3 summarizes the traditional information retrieval evaluation metrics discussed in this section.

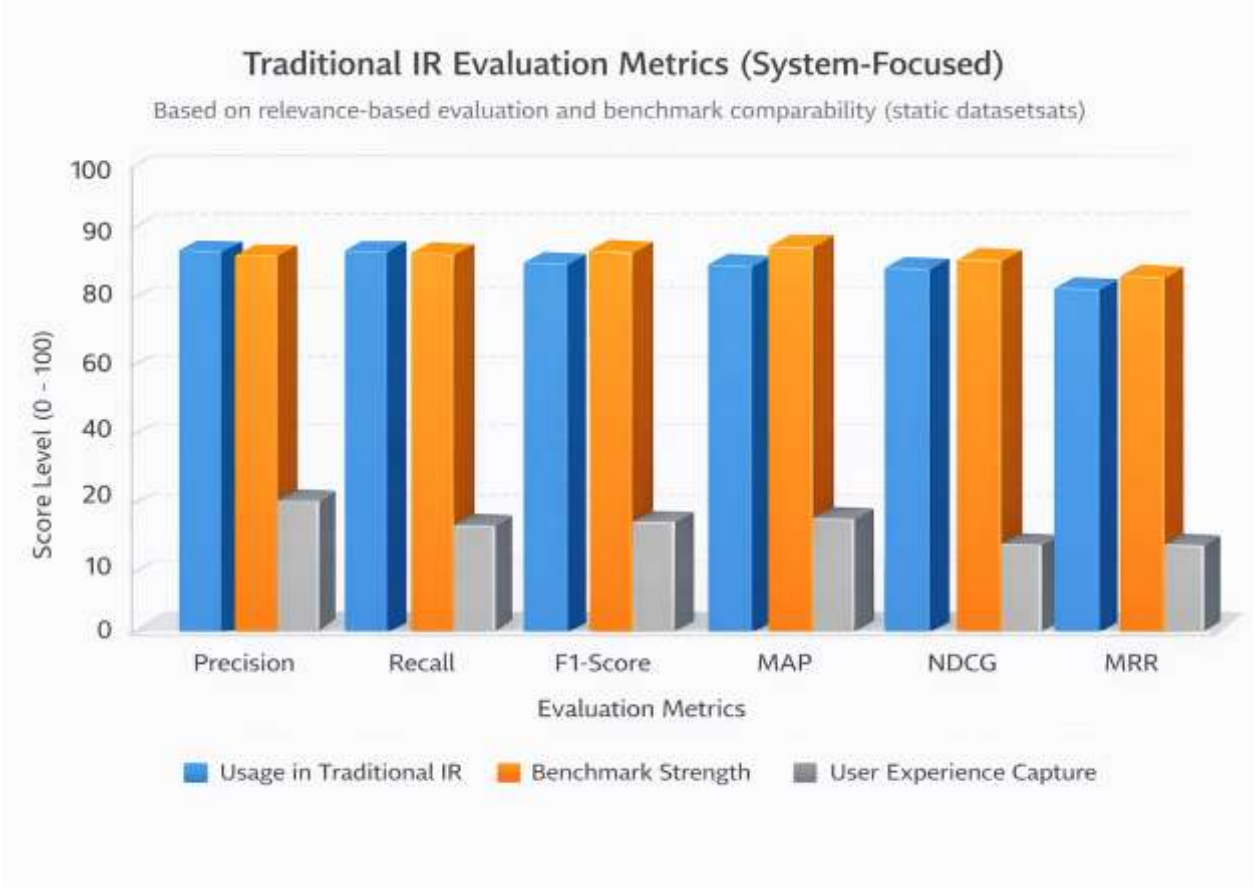


Figure 3: Traditional Information Retrieval Metrics

Emergence of Deep Learning Models

The limitations of traditional information retrieval, especially its focus on keyword matching, led to the development of semantic search. This new approach highlights the importance of context and understanding user intent. The use of deep learning methods has greatly improved this model [3,4,15]. Transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) and its variations, such as Sentence-BERT (SBERT), offer contextual embeddings that capture subtle meanings that go beyond simple keyword matching. Models like ColBERT (Contextualized Late Interaction over BERT) introduced effective late-interaction methods to find a balance between accuracy and scalability. Meanwhile, Dense Passage Retrieval (DPR) used dual-encoder setups to enhance passage-level retrieval [14]. These models have consistently outperformed traditional methods like BM25 on standard datasets, proving that deep learning is now the leading

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

approach for semantic search [12]. Figure 4 illustrates the SBERT semantic retrieval pipeline used in modern semantic search systems.

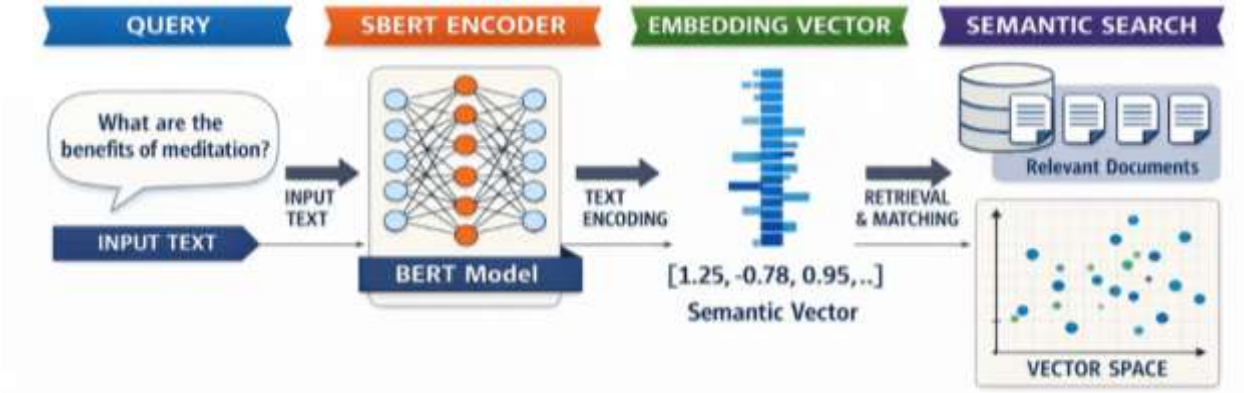


Figure 4: SBERT Semantic Retrieval Pipeline

Benchmark Datasets Commonly Used

The development and evaluation of semantic search models depend on benchmark datasets that create standardized test environments. One of the most influential datasets is MS MARCO (Microsoft Machine Reading Comprehension). This dataset includes large-scale passage ranking and question-answering tasks that are widely used to train and test neural retrieval systems [13,16]. Natural Questions (NQ) comes from Google search queries and offers a more realistic view of real-world information needs. Likewise, the TREC (Text Retrieval Conference) collections, which include TREC CAR (Complex Answer Retrieval) and TREC-COVID, have been essential evaluation resources [17]. These datasets contain relevance judgments that allow the use of MAP, NDCG, and similar metrics. They are fundamental to evaluation pipelines in both academic and industrial research [18]. Figure 5 presents an overview of the benchmark datasets discussed in this section.

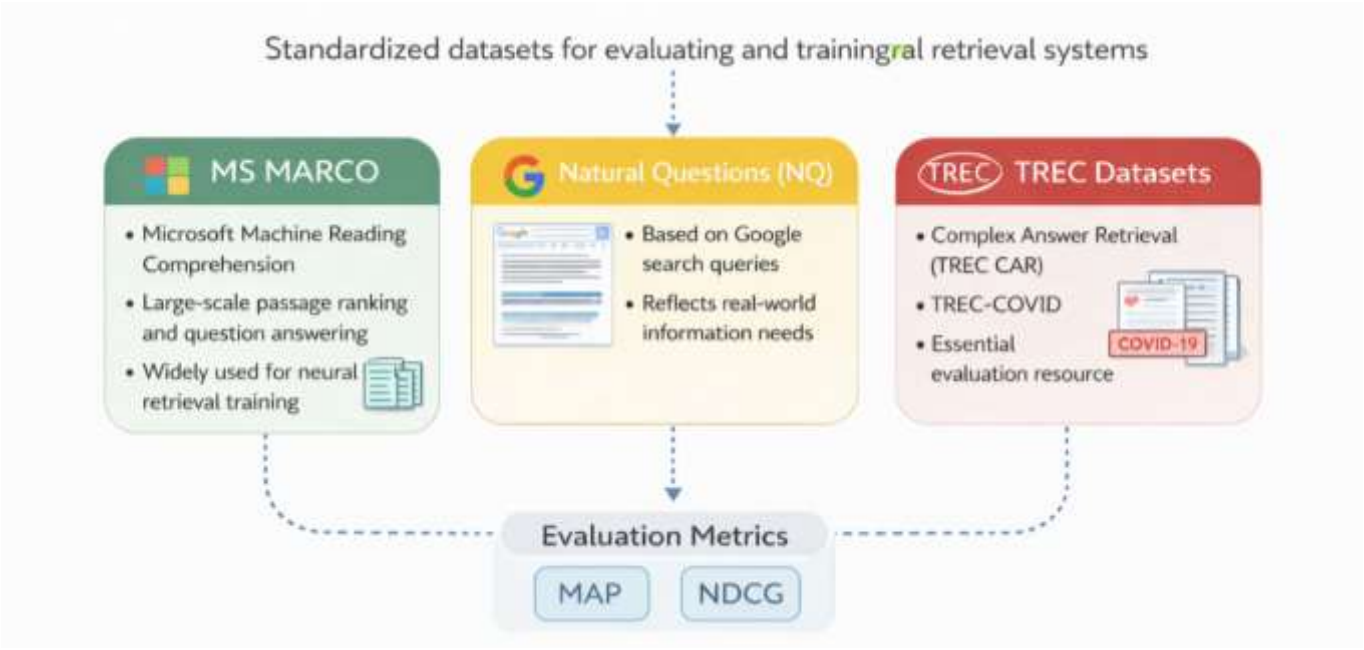


Figure 5: Dataset Overview

Current Standard Evaluation Pipelines

In practice, evaluating semantic search systems still follows a standard experimental process. Models are usually trained on large annotated datasets, fine-tuned for retrieval tasks, and tested against benchmark datasets using system-focused metrics like NDCG@10, MAP, and MRR [19]. Leaderboards for datasets such as MS MARCO

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

have become central to showing top performance, creating a competitive culture around small metric improvements. While this process ensures comparability and reproducibility, it also reinforces the dominance of algorithmic benchmarks that may not fully reflect real-world user satisfaction [14,16]. Consequently, there is growing concern that the current evaluation culture may lead to optimizing models for leaderboard results instead of meeting practical user needs [15].

The Normalized Discounted Cumulative Gain (NDCG). NDCG evaluates ranked retrieval by assigning higher scores to relevant documents appearing near the top of the results list. It tries to account for graded relevance that is, not all relevant documents are equally useful. NDCG at position k (NDCG@k) is defined as: $NDCG@k = \frac{\text{DCG@k}}{\text{IDCG@k}}$, where DCG@k is the Discounted Cumulative Gain up to rank k, and IDCG@k is the ideal (maximum possible) DCG [19].

Mean Average Precision (MAP) practically is an aggregate measure of precision that considers the ranking of relevant documents across multiple queries. For each query, the average precision (AP) is computed by averaging the precision values at each rank where a relevant document appears. MAP is the mean of these AP values across all test queries. It is calculated as; $MAP = \frac{1}{Q}\sum_{i=1}^{Q} A\,P_i$ where Q is the number of queries [12,17].

Mean Reciprocal Rank (MRR) as a different metric measures the effectiveness of systems in retrieving the first relevant result. It is particularly useful for applications such as question answering and chatbot response ranking. It is defined as; $MRR = \frac{1}{Q}\sum_{i=1}^{Q} \frac{1}{rank_i}$, where $rank_i$ is the rank position of the first relevant document for the *the i-th query* [15,19]. Figure 6 summarizes these common system-focused evaluation metrics used in semantic search.

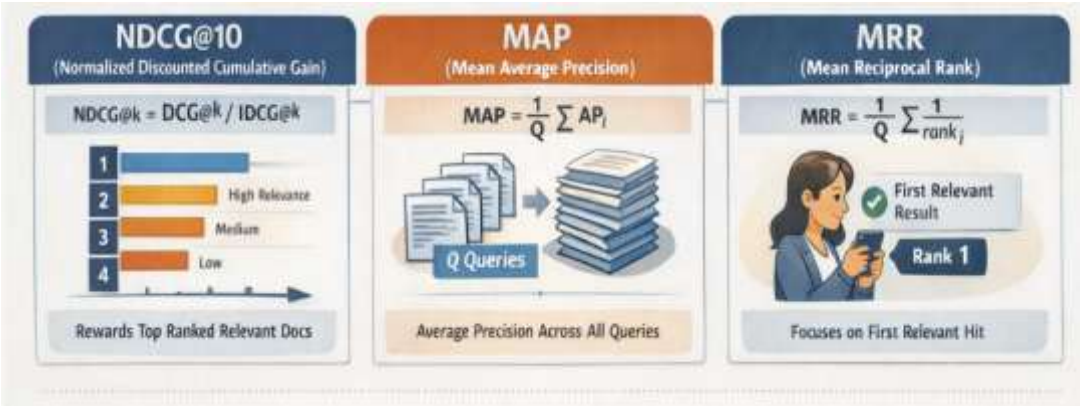


Figure 6: Common System-Focused Evaluation Metrics for Semantic Search

Critical Analysis of Metrics

Traditional Metrics

Traditional evaluation metrics are the foundation of information retrieval (IR) research [20]. Their strengths come from being simple, comparable, and reproducible. Metrics like precision, recall, and MAP are easy to calculate, widely understood, and allow for consistent comparisons across models and datasets. They promote comparability by offering a common standard within the IR community. This ensures that improvements can be tracked over time and among different research groups [21].

However, their weaknesses are becoming more apparent in the age of semantic search. Traditional metrics focus mainly on the system and are designed for static relevance judgments instead of dynamic user interactions. They do not capture behavioral signals, such as how users navigate, change queries, or engage with the results they retrieve [22]. Additionally, these metrics often treat relevance as binary or graded but overlook elements like clarity, usability, and contextual fit, which greatly affect how users view results as satisfactory [23].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Neural IR and Semantic Search Adaptations

The rise of deep learning-based retrieval has led to changes in existing metrics. Measures like Mean Reciprocal Rank (MRR), Recall@K, and improvements in NDCG have become popular for evaluating semantic search [21]. These metrics are especially helpful for tasks such as passage retrieval, question answering, and ranking relevance at scale. For example, Recall@K checks if relevant results show up in the top k retrieved documents. This illustrates the importance of ranking in real-world systems [24].

However, most of these evaluations occur offline using benchmark datasets, where performance is assessed against pre-labeled relevance sets. This ensures reproducibility but may lead to overfitting models to specific dataset features [20]. In contrast, online evaluations, like A/B testing in live settings, provide real-world feedback. Yet, they are less common in academic research due to resource and ethical challenges. This conflict between offline accuracy and online reality underscores a significant issue in current evaluation practices [25].

User-Oriented Attempts

Recognizing the limitations of purely algorithmic measures, researchers have tested user-oriented proxies. Metrics like click-through rates (CTR), dwell time, and query reformulations provide behavioral signals of satisfaction and engagement [26]. For instance, longer dwell times and fewer query reformulations often indicate successful retrieval. Mixed-methods approaches that combine user studies with system performance analysis offer richer insights. Controlled experiments, surveys, and interviews can reveal subjective views of relevance and usability, adding to quantitative system metrics [27].

Despite these improvements, user-oriented efforts are still not fully developed and are often seen as secondary to algorithmic benchmarks. They are also tougher to standardize, as behavioral signals differ across domains, user groups, and tasks [26,28].

Gaps Identified

The review shows ongoing gaps in current evaluation practices. Most importantly, there is a lack of subjective satisfaction measures. Such measures indicate whether users feel their information needs were met within a specific context [29]. Current metrics rarely include personalization or consider the variety of user backgrounds, intents, and expectations. Contextual factors, like task difficulty or specific needs, are also ignored. As a result, semantic search models may score well on leaderboards but still fail to provide satisfactory experiences for end-users [20,29]. Figure 7 highlights the key gaps identified in current semantic search evaluation frameworks.

Figure 7: Gaps in Semantic Search Evaluation Frameworks

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Table 1: Comparison of Traditional and User-Centered Evaluation Metrics in Semantic Search

Metric	Type	Advantages	Limitations	Practical Applicability
Precision	Traditional	Simple and interpretable; measures retrieval accuracy	Ignores ranking quality and user context	Useful for exact-match retrieval tasks
Recall	Traditional	Measures of completeness of retrieval	May retrieve many irrelevant results	Suitable for high-recall domains such as healthcare
MAP	Traditional	Evaluates ranking quality across queries	Does not reflect subjective user experience	Effective for benchmark comparisons
NDCG	Traditional	Accounts for ranking position and graded relevance	Limited in capturing usability and satisfaction	Widely used in semantic search benchmarking
MRR	Traditional	Measures effectiveness of first relevant result retrieval	Focus only on the first relevant item	Useful in question answering and chatbot systems
CTR (Click-Through Rate)	User-Centered	Reflects user interaction and engagement	Can be influenced by interface design and ranking bias	Useful in live search and recommendation systems
Dwell Time	User-Centered	Indicates depth of user engagement with results	Difficult to standardize across users and tasks	Valuable for measuring perceived relevance
Satisfaction Surveys	User-Centered	Captures subjective user perception directly	Time-consuming and difficult to scale	Useful in usability and UX evaluation
Task Completion Rate	User-Centered	Measures practical success of retrieval systems	Context-dependent and task-specific	Effective in applied and domain-specific evaluations
Hybrid Evaluation Frameworks	Combined Approach	Balances technical accuracy with user experience	More complex to implement and standardize	Recommended for future semantic search evaluation

DISCUSSION

This review highlights several key insights into evaluating deep learning-based semantic search systems. First, traditional metrics like precision, recall, MAP, and NDCG have been crucial for advancing information retrieval research [24,30]. They provide standardized and repeatable benchmarks. Their simplicity and comparability have facilitated rapid development within the field and created clear performance hierarchies. However, these advantages also hide their biggest flaw: they are disconnected from practical user experience [31].

A central finding is the gap between high metric scores and low user experience quality. Semantic search models often perform excellently on benchmark datasets, yet these results do not always lead to positive practical experiences [32]. For instance, a top-ranked document may be technically relevant but use overly complex language, lack clarity, or not fit with the user's specific context. In these situations, a system optimized for NDCG or MRR can still fail at its main goal, which is to help users efficiently find the information they need. This disconnect shows that current evaluation processes often aim for leaderboard success instead of real user benefit [33].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Several practical semantic search applications further demonstrate this limitation. Retrieval models based on BERT and SBERT frequently achieve strong benchmark performance on datasets such as MS MARCO and BEIR, yet users may still encounter contextually weak or less meaningful search results in practical environments. In many cases, systems optimized primarily for ranking metrics fail to adequately capture user intent, contextual relevance, and perceived usefulness. This highlights the limitation of relying solely on leaderboard-focused evaluation and reinforces the need for hybrid evaluation frameworks that combine technical performance with user-centered indicators such as click behavior, dwell time, and satisfaction feedback [41].

The practical implications are significant. In applied environments, from healthcare decision support to e-commerce product searches, user experience is key to system effectiveness. Users evaluate retrieval systems not just on relevance but also on usefulness, clarity, timeliness, and how well they fit the context [34]. Poorly designed evaluations can lead to systems that seem superior in research but frustrate or mislead users in practice. For organizations, this can decrease trust, engagement, and adoption of retrieval technologies, undermining the value of even technically advanced systems [35].

These insights also provide lessons for designing future metrics. One approach is to develop hybrid frameworks that combine offline algorithmic evaluation with user-focused indicators. For example, NDCG@10 could be paired with satisfaction surveys, dwell-time analysis, or task completion rates [36]. Such combined methods would maintain reproducibility while capturing the subjective and contextual aspects of performance. Another lesson is the need for context-aware metrics that reflect specific domain requirements. In medical retrieval, for instance, clarity and reliability may be more important than lexical precision, while in e-commerce, speed and personalization might be prioritized [37].

Overall, the discussion highlights that the evolution of semantic search evaluation must align with the advancement of semantic search itself. Just as retrieval models have progressed from basic keyword matching to deeper contextual understanding, evaluation metrics must shift from static relevance judgments to measures that capture human experience [38,39]. Only then can semantic search systems be fully optimized for their intended purpose, which is to help users retrieve meaningful and usable information.

CONCLUSION AND FUTURE DIRECTIONS

This review has critically examined how we evaluate deep learning-based semantic search systems, pointing out both the strengths and weaknesses of current metrics. Traditional measures like precision, recall, MAP, and NDCG are still useful because they are simple, comparable, and reproducible [35,37]. They have built the foundation for measuring progress in information retrieval and continue to be important for system development. However, their flaws are also clear: these metrics focus on the system, are made for static datasets, and do not fully capture the details of user experience, task success, and contextual relevance [40].

The findings highlight the urgent need for evaluation frameworks that include user feedback, behavioral signals, and individual user perception along with algorithm performance. New user-focused methods like click-through analysis, dwell time, satisfaction surveys, and mixed-methods evaluations present good opportunities to connect technical success with practical applications [32]. However, these methods are still in early development and lack the consistency needed for broad acceptance [33].

Future progress will need a more collaborative approach. Evaluating semantic search should not rely only on information retrieval metrics; it should also include insights from human-computer interaction (HCI), user experience (UX) research, and cognitive science [31,34]. This interdisciplinary approach can create richer, context-aware evaluation models that show not just if a result is relevant, but also if it is usable, understandable, and satisfying [23,33].

In conclusion, the future assessment of semantic search should move away from static, algorithm-centered measures toward hybrid, user-centered frameworks. By combining algorithm rigor with human experience, researchers and practitioners can create search systems that are not only powerful in computation but also meaningful and effective in real-world settings [40].

REFERENCES

1. L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," *arXiv preprint*, arXiv:2402.05672, 2024.
2. X. Li and J. Li, "Angle-optimized text embeddings," *arXiv preprint*, arXiv:2309.12871, 2023.
3. A. Esteva, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev, and R. Socher, "COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization," NPJ Digit. Med., vol. 4, no. 1, p. 68, 2021.
4. M.-K. Ghali, A. Farrag, D. Won, and Y. Jin, "Enhancing knowledge retrieval with in-context learning and semantic search through generative AI," Knowl.-Based Syst., 2025, Art. no. 113047.
5. T. Hellert, J. Montenegro, and A. Pollastro, "PhysBERT: A text embedding model for physics scientific literature," *APL Mach. Learn.*, vol. 2, no. 4, 2024.
6. L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Improving text embeddings with large language models," *arXiv preprint*, arXiv:2401.00368, 2023.
7. A. Moffat, "Batch evaluation metrics in information retrieval: Measures, scales, and meaning," *IEEE Access*, vol. 10, pp. 105564–105577, 2022.
8. L. L. de Oliveira, D. S. Vargas, A. M. A. Alexandre, F. C. Cordeiro, D. D. S. M. Gomes, M. D. C. Rodrigues, ... and V. P. Moreira, "Evaluating and mitigating the impact of OCR errors on information retrieval," *Int. J. Digit. Libr.*, vol. 24, no. 1, pp. 45–62, 2023.
9. X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, and Z. Dou, "From matching to generation: A survey on generative information retrieval," *ACM Trans. Inf. Syst.*, vol. 43, no. 3, pp. 1–62, 2025.
10. A. Salemi and H. Zamani, "Evaluating retrieval quality in retrieval-augmented generation," in *Proc. 47th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, July 2024, pp. 2395–2400.
11. W. Zhang, M. Zhang, S. Wu, J. Pei, Z. Ren, M. de Rijke, ... and P. Ren, "Excluir: Exclusionary neural information retrieval," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 12, Apr. 2025, pp. 13295–13303.
12. M. Gaur, K. Gunaratna, V. Srinivasan, and H. Jin, "Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, June 2022, pp. 10672–10680.
13. N. Girdhar, M. Coustaty, and A. Doucet, "Digitizing history: transitioning historical paper documents to digital content for information retrieval and mining—a comprehensive survey," *IEEE Trans. Comput. Social Syst.*, 2024.
14. O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, 2024.
15. X. Qi, Y. Zhang, S. Cao, S. Yan, and H. Su, "Human–computer interaction based on the intelligent information retrieval method for customer satisfaction in power system service," *Int. J. Model. Simul. Sci. Comput.*, vol. 14, no. 01, p. 2341004, 2023.
16. A. M. Flores, M. C. Pavan, and I. Paraboni, "User profiling and satisfaction inference in public information access services," *J. Intell. Inf. Syst.*, vol. 58, no. 1, pp. 67–89, 2022.
17. C. Siro, M. Aliannejadi, and M. de Rijke, "Understanding user satisfaction with task-oriented dialogue systems," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 2018–2023.
18. C. Siro, M. Aliannejadi, and M. De Rijke, "Understanding and predicting user satisfaction with conversational recommender systems," *ACM Trans. Inf. Syst.*, vol. 42, no. 2, pp. 1–37, 2023.
19. C. Bauer, M. Fröbe, D. Jannach, U. Kruschwitz, P. Rosso, D. Spina, and N. Tintarev, "Overcoming methodological challenges in information retrieval and recommender systems through awareness and education," *Bauer et al.* [2023a], pp. 51–67.
20. T. E. Kim and A. Lipani, "A multi-task based neural model to simulate users in goal oriented dialogue systems," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, July 2022, pp. 2115–2119.
21. M. B. Yılmaz and K. Rızvanoğlu, "Understanding users' behavioral intention to use voice assistants on smartphones through the integrated model of user satisfaction and technology acceptance: a survey approach," *J. Eng. Des. Technol.*, vol. 20, no. 6, pp. 1738–1764, 2022.
22. C. L. Hsu and J. C. C. Lin, "Understanding the user satisfaction and loyalty of customer service chatbots," *J. Retail. Consum. Serv.*, vol. 71, p. 103211, 2023.
23. J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng, "Semantic models for the first-stage retrieval: A comprehensive review," *ACM Trans. Inf. Syst. (TOIS)*, vol. 40, no. 4, pp. 1–42, 2022.

24. X. Li, K. Dong, Y. Q. Lee, W. Xia, H. Zhang, X. Dai, ... and R. Tang, "CoIR: A comprehensive benchmark for code information retrieval models," *arXiv*, *preprint*, arXiv:2407.02883, 2024.
25. H. W. Kim, D. H. Shin, J. Kim, G. H. Lee, and J. W. Cho, "Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval," *Seizure: Eur. J. Epilepsy*, vol. 114, pp. 1–8, 2024.
26. J. Gao, C. Xiong, P. Bennett, and N. Craswell, *Neural Approaches to Conversational Information Retrieval*, vol. 44. Heidelberg: Springer, 2023.
27. N. Thakur, L. Bonifacio, M. Fröbe, A. Bondarenko, E. Kamalloo, M. Potthast, ... and J. Lin, "Systematic evaluation of neural retrieval models on the Touché 2020 argument retrieval subset of BEIR," in *Proc. 47th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, July 2024, pp. 1420–1430.
28. W. Gao, H. Wang, Q. Liu, F. Wang, X. Lin, L. Yue, ... and S. Wang, "Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, July 2023, pp. 983–992.
29. J. I. Friese and N. Fuhr, "Towards Reproducibility of Interactive Retrieval Experiments: Framework and Case Study," in *Proc*. *Eur. Conf. Inf. Retrieval (ECIR)*, Apr. 2025, pp. 146–160. Cham: Springer Nature Switzerland.
30. A. Anand, L. Lyu, M. Idahl, Y. Wang, J. Wallat, and Z. Zhang, "Explainable information retrieval: A survey," *arXiv preprint*, arXiv:2211.02405, 2022.
31. L. Heryawan, D. Novitaningrum, K. R. Nastiti, and S. N. Mahmudah, "Medical record document search with TF-IDF and vector space model (VSM)," *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 14, no. 3, pp. 847–852, 2024.
32. F. M. Hasyim and F. Fahmi, "A literature review: The importance of term normalization in vector space mode," *Literatify: Trends Libr. Develop.*, vol. 5, no. 1, pp. 18–28, 2024.
33. C. I. Nakpih, "A modified Vector Space Model for semantic information retrieval," *Nat. Lang. Process. J.*, vol. 8, p. 100081, 2024.
34. D. Paseru, R. Kuera, and S. Salmon, "Comparison of Vector Space Model and BM25F Methods in Book Search," *J. RESTIKOM: Riset Tek. Inform. Komput.*, vol. 5, no. 3, pp. 513–522, 2023.
35. M. A. A. Shiddiqi and A. Sanmarino, "Vector Space Model-based Information Retrieval Systems at South Sumatera Regional Libraries," *J. Comput. Sci. Appl. Eng. (JOSAPEN)*, vol. 1, no. 2, pp. 49–53, 2023.
36. P. B. Bahtera and D. S. Kartawijaya, "Content Classification of the Official Website of the Ministry of Foreign Affairs of the Republic of Indonesia (MoFA RI) using Vector Space Model (VSM)," *MALCOM: Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 4, pp. 1309–1319, 2024.
37. B. P. Zen, I. Susanto, K. Putriyani, and Sintiya, "Automatic document classification for Tempo news articles about COVID-19 based on term frequency, inverse document frequency (TF-IDF), and Vector Space Model (VSM)," in *AIP Conf. Proc.*, vol. 2952, no. 1, July 2024, p. 060003.
38. J. J. M. De Araujo and A. A. J. Sinlae, "The Visualization of the Vector Space Model in Searching for Immigration News in the East Nusa Tenggara Region," in *Proc. Nat. Conf. Electr. Eng., Informatics, Ind. Technol. Creative Media*, vol. 3, no. 1, pp. 924–931, 2023.
39. Y. Wang, Y. Hou, H. Wang, Z. Miao, S. Wu, Q. Chen, ... and M. Yang, "A neural corpus indexer for document retrieval," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 25600–25614, 2022.
40. S. Bruch, C. Lucchese, M. Maistro, and F. M. Nardini, "Special Section on Efficiency in Neural Information Retrieval," *ACM Trans. Inf. Syst.*, vol. 42, no. 5, pp. 1–4, 2024.
41. T. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in *Proc. 35th Conf. on Neural Inf. Process. Syst. (NeurIPS)*, 2021.