Performance Evaluation of Embedding-Based and Keyword-Based Retrieval in Text Description-Based Hotel Recommendation

Ilham Yusuf Faturochman[1*], Aprilisa Arum Sari[2], Nibras Faiq Muhammad[3]

[1,2,3] *Information Engineering, Universitas Duta Bangsa Surakarta, Bhayangkara Street No. 55, Tipes, Serengan District, Surakarta City, Central Java 57154, Indonesia*

Keywords

MAP; MiniLM; Precission@3; Recommendation System; TF-IDF

Correspondence Email:
Ilhamyusuf1004@email.com

Abstract

The massive volume of textual descriptions on hotel booking platforms makes it difficult for recommendation systems to accurately match user preferences. Traditional keyword-based retrieval methods, such as TF-IDF, often struggle to capture semantic relationships when relevant terms do not explicitly overlap. This study evaluates the performance of keyword-based (TF-IDF) and embedding-based (paraphrase-multilingual-MiniLM-L12-v2) retrieval approaches in a content-based hotel recommendation system using a small-scale dataset. The dataset consists of 30 unique Traveloka hotels in Yogyakarta collected from Kaggle, representing a resource-constrained experimental setting. Evaluation was conducted using a black-box approach with 10 dynamic synthetic queries and assessed through Precision@3 (P@3) and Mean Average Precision (MAP). The results indicate that MiniLM achieved higher retrieval effectiveness than TF-IDF, with a mean P@3 of 0.3667 and a mean MAP of 0.1378, compared with 0.3000 and 0.1333, respectively. These findings suggest that embedding-based retrieval is more effective in capturing semantic information, including synonym usage and implicit contextual relationships, within the evaluated dataset. Therefore, compact embedding models such as MiniLM may provide an alternative approach to traditional keyword-based retrieval methods for small-scale recommendation systems.

1. Introduction

The rapid growth of textual data volume across various information systems has driven the need for retrieval and recommendation methods that can effectively generate relevant information, particularly within text-based recommendation systems[1], [2]. In the field of information retrieval, keyword-based methods such as TF-IDF and BM25 remain widely used due to their simplicity and effectiveness in ranking textual documents[3]. However, these approaches suffer from a fundamental limitation in capturing semantic relationships between documents when there is no direct keyword overlap[4], [5]. Previous research on Indonesian journal metadata retrieval reported that BM25 achieved a higher Mean Average Precision (MAP) score than TF-IDF (0.74 and 0.59 respectively), indicating that retrieval effectiveness can vary substantially among keyword-based ranking methods[3].

Nevertheless, word-weighting methods are still commonly utilized as a baseline in content-based filtering and information retrieval research[6]. TF-IDF has been shown to outperform Count Vectorizer in content-based recommendation systems, yielding higher accuracy and relevance scores in document similarity taks[6]. On the other hand, recent advances in embedding-based approaches enable text to be represented as dense semantic vectors that capture contextual and semantic relationships beyond exact keyword matching[4], [7]. Studies on scientific publication recommendation have demonstrated that word embeddings and sentence encoders can improve the modeling of semantic relationships among documents[8], resulting in more relevant recommendations compared with purely lexical representations[1]. Furthermore, integrating semantic information into recommendation systems has been shown to improve recommendation quality by better representing contextual meaning and user interests[9], [2].

Numerous studies have demonstrated the potential of embedding-based approaches to improve retrieval quality, particularly for long and descriptive textual documents where semantic matching plays an important role[1], [10], [11], [12]. However, most of these studies were conducted on large-scale corpora or relatively complex processing environments[10], [13].In practice, not all recommendation system developers have access to massive datasets or adequate computational resources. Consequently, further evaluation regarding the effectiveness of embedding methods under data scarcity conditions is highly required. Given these circumstances, this study focuses on comparing keyword-based retrieval and embedding-based retrieval approaches on a small-scale dataset to determine the extent to which embedding methods can still deliver improvements in retrieval quality over traditional methods.

This study utilizes a subset of data from the Kaggle dataset titled "New Year Hotel Rooms Availability from Traveloka (2022)," limited to 30 hotels in Yogyakarta. The selection of Yogyakarta is based on its characteristics as one of the primary tourist destinations in Indonesia[14], [15],featuring a high diversity of hotels and tourism facilities. This small-scale dataset is used to represent realistic conditions often faced by system developers with limited data availability. To bypass the extensive training requirements typical of embedding approaches, this study leverages the pretrained Sentence Transformer model paraphrase-multilingual-MiniLM-L12-v2, which is based on the Sentence-BERT framework, and compares its performance against TF-IDF as the baseline[16], [17], [18].

The purpose of this research is to compare the performance of embedding-based retrieval and traditional keyword-based methods within a text description-based hotel recommendation system. The evaluation was conducted using 10 dynamic queries designed to represent various user needs in hotel searches. This study contributes an empirical evaluation of a multilingual embedding model for hotel recommendation on a small-scale dataset. The findings of this study are expected to serve as a consideration baseline for selecting appropriate retrieval methods in data-constrained recommendation systems, thereby enhancing recommendation quality without requiring a massive data corpus.

2. Research Methods

The research layout is designed as an empirical comparative experiment conducted through several structured phases, including dataset preparation, text preprocessing, feature extraction using both methods, query execution, and evaluation calculation. All computational experiments were executed on a machine equipped with an AMD A9 processor and 4 GB of RAM, running Python 3.10 as the primary programming language. The core software environment integrated several open-source libraries, specifically scikit-learn for the TF-IDF baseline implementation, sentence-transformers for processing the embedding vectors, and pandas along with numpy for data manipulation and evaluation matrix tracking.

2.1 Data Collection and Dataset Characteristics

The primary material used in this study is secondary textual data extracted from the Kaggle repository under the title "New Year Hotel Rooms Availability from Traveloka (2022)", The original dataset contains approximately 400 hotel records from multiple Indonesian cities. For experimental control, only the top 30 unique hotels located in Yogyakarta were selected. Each hotel entry contains unstructured textual attributes,

including the hotel name, explicit amenities, geographic vicinity, and descriptive overviews that portray the ambiance and services offered.

2.2 Technical Implementation of Retrieval Models

The system processes the text corpus through two distinct pipelines to perform information retrieval. The first pipeline implements Term Frequency-Inverse Document Frequency (TF-IDF) as the lexical baseline using the word-based configuration of the TfidfVectorizer provided by the scikit-learn library. Each hotel description is represented as a sparse vector weighted by term frequency and inverse document frequency across the corpus[6].During preprocessing, all text is automatically converted to lowercase to ensure tokenization consistency, while no external stopword removal is applied in order to preserve the original structure of the Indonesian textual descriptions. Similarity between query vectors and document vectors is calculated using Cosine Similarity, and the resulting scores are used to rank the retrieved hotel recommendations.

The second pipeline utilizes the pretrained Sentence Transformer model, specifically paraphrase-multilingual-MiniLM-L12-v2, which is built upon the Sentence-BERT architecture[16], [17].This model transforms each hotel description into a dense semantic embedding of 384 dimenssions[16], enabling semantically related texts to be represented by nearby vectors within the embedding space without requiring additional fine-tuning[17].Similarity between query embeddings and hotel embeddings is likewise computed using Cosine Similarity. This approach follows the dense retrieval paradigm, where semantic relevance is measured in a continuous vector space rather than through exact keyword matching[4].

To ensure a consistent comparison between the two retrieval approaches, both models employ the same ranking strategy. For each query, the system calculates similarity scores against all hotel documents, ranks the results in descending order, and returns the Top-3 recommendations with the highest similarity scores. This Top-K configuration (K = 3) is aligned with the Precision@3 evaluation metric used in this study.

2.3 Experimental Evaluation

The evaluation framework adopts an established black-box information retrieval testing methodology. The evaluation employed 10 manually designed synthetic queries representing common hotel-search scenarios, including business travel, family vacations, luxury stays, accessibility needs, and amenity-focused searches. Ground-truth labels were manually assigned based on the semantic relevance between each query and hotel description prior to testing. Before computing the evaluation metrics, the system calculates the text proximity between the query vector (A) and the hotel document vector (B) using Cosine Similarity. The mathematical formulation for this similarity coefficient is defined as follows:

$$Cosine\ Similarity(A,B) = \frac{A \cdot B}{||A||||B||} \qquad (1)$$

Retrieval performance and relevance ranking are strictly measured using two standard validation metrics, namely Precision at Top-3 (Precision@3) to evaluate immediate recommendation quality and Mean Average Precision (MAP) to determine the overall ranking accuracy across all executed queries[19], [20].The precision at a specific cutoff K (where $K=3$ in this study) is formulated as:The precision at a specific cutoff K is formulated as:

$$P@K = \frac{1}{K}\sum_{i=1}^{K} rel(i) \qquad (2)$$

Where $rel(i)$ is an indicator function that outputs 1 if the hotel recommended at rank i is present within the ground-truth database, and 0 otherwise. Furthermore, Average Precision (AP) is computed for each individual query to evaluate the specific ranking quality by tracking precision at every relevant document drop:

$$AP = \frac{1}{|R|}\sum_{i=1}^{K} P@i * rel(i) \quad (3)$$

Where $|R|$ represents the total number of actual relevant hotels defined in the ground truth for that specific query. Finally, the Mean Average Precision (*MAP*) is calculated by taking the arithmetic mean of the *AP* scores across all Q evaluated queries (where Q=10);

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP_q \quad (4)$$

3. Result and Discussion

This section outlines the systematic findings gathered from the computational experiment comparing the traditional TF-IDF pipeline and the paraphrase-multilingual-MiniLM-L12-v2 embedding-based dense retrieval mechanism. The data generated through the 10 dynamic query tests are thoroughly evaluated, followed by a rigorous theoretical reflection contrasting the results against previous information retrieval literature.

3.1 Analytical Demonstration of Experimental Results

To demonstrate transparency in metric calculation prior to displaying the complete dataset, a manual verification example is conducted using query q2 under the MiniLM pipeline. For query q2, the system processes the user request and returns three top hotel recommendations (*K=3*). According to the validated ground-truth database, two out of the three retrieved hotels are marked as relevant (occurring at rank 1 and rank 3), while the second-ranked hotel is irrelevant. Applying the formulas specified in the evaluation framework, the explicit calculations for Precision at Top-3 (*P@3*) and Average Precision (*AP*) are performed as follows:

$$P@3 = \frac{2}{3} = 0.6667 \quad (5)$$

$$AP = \frac{(1.0 * 1) + (0.0 * 0) + (0.6667 * 1)}{5} = \frac{1.6667}{5} = 0.3333 \quad (6)$$

Following this precise mathematical formulation, the automated script logged the retrieval scores across all 10 user queries. The complete empirical results are systematically structured in Table 1.

Table 1. Evaluation Performance per Query for TF-IDF and MiniLM Pipelines

Query ID	Table Column Head				
	Query Text	P@3 TF-IDF	MAP TF-IDF	P@3 MiniLM	MAP MiniLM
Q1	"Cari hotel di daerah Depok yang nyaman untuk kerja dan meeting."	0.6667	0.2333	0.3333	0.2000
Q2	"tempat menginap mewah yang mengesankan dan ada main bola sodok"	0.0000	0.0000	0.6667	0.4000
Q3	"hotel yang nyaman untuk beristirahay dan menghilangkan rasa lelah"	0.3333	0.2000	0.3333	0.0667
Q4	"Cari tempat menginap untuk liburan keluarga yang ada banyak aktifitas	0.3333	0.1000	0.6667	0.3333

Query ID	Table Column Head				
	Query Text	P@3 TF-IDF	MAP TF-IDF	P@3 MiniLM	MAP MiniLM
	buat anak-anak"				
Q5	"Cari hotel yang praktis untuk wisatawan tanpa kendaraan pribadi dan mudah bepergian ke berbagai tempat."	0.0000	0.0000	0.3333	0.0667
Q6	"Cari hotel yang ada tempat main bola sodok dan fasilitas hiburan malam."	0.0000	0.0000	0.3333	0.0667
Q7	"Butuh penginapan dengan koneksi internet cepat karena harus ikut meeting online."	0.3333	0.0667	0.0000	0.0000
Q8	"Cari hotel yang punya tempat olahraga lengkap untuk tetap menjaga kebugaran selama liburan."	0.6667	0.3333	0.3333	0.0667
Q9	"Ingin hotel yang ada tempat santai di rooftop untuk menikmati pemandangan kota Yogyakarta malam hari."	0.3333	0.0667	0.3333	0.0667
Q10	"Cari penginapan yang menyediakan antar jemput bandara supaya tidak repot cari transportasi."	0.3333	0.3333	0.3333	0.1111
Mean	Overall Evaluation Score	0.3000	0.1333	0.3667	0.1378

As explicitly referenced in Table 1, the collective mean results indicate that the embedding-based MiniLM model yields an overall higher performance compared to the traditional TF-IDF lexical baseline. The final summary highlights that MiniLM achieves a Mean P@3 of 0.3667, showing an upward trajectory over TF-IDF which scores 0.3000. Additionally, the overall ranking accuracy measured through Mean MAP reflects a subtle margin of victory for the dense model, securing a score of 0.1378 against the 0.1333 baseline achieved by TF-IDF. This corresponds to a relative improvement of approximately 22.2% in P@3 and 3.4% in MAP over the TF-IDF baseline.

3.2 Theoretical Interpretation and Comparative Discussion

A granular analysis of the individual query behavior in Table 1 yields deep, provides useful insights regarding retrieval dynamics in data-constrained scenarios. The vocabulary mismatch problem manifests starkly in queries containing colloquialisms, idioms, or implicit user desires, such as q2 ("tempat main bola sodok"), q5 ("wisatawan tanpa kendaraan pribadi"), and q6 ("fasilitas hiburan malam"). On these test beds, TF-IDF suffers from complete failure, resulting in absolute zero scores (P@3 = 0.0000, MAP = 0.0000). Because traditional word-weighting relies on string matching, it is structurally unable to bridge the gap between user phrasing and formal text descriptions. Conversely, MiniLM successfully captures the underlying semantic intent, providing a P@3 of 0.6667 for q2 and q4, and 0.3333 for q5 and q6. This observation supports the foundational theories established by previous research on semantic encoders[1], [17],confirming that projecting text into continuous dense spaces alleviates vocabulary mismatches and MiniLM consistently

outperformed TF-IDF in the evaluated dataset, although the margin of improvement was relatively modest in terms of MAP.

In contrast, an essential divergence emerges when investigating queries dominated by highly specific, unambiguous geographical terms or strict keyword identifiers. In query q1 ("di daerah Depok") and q8 ("tempat olahraga lengkap"), the TF-IDF pipeline markedly outperforms MiniLM, reaching a high P@3 score of 0.6667. This demonstrates that lexical matching excels at isolating distinct tokens like "Depok" (a specific regional sub-district name) or concrete amenities. The sentence encoder, utilizing context pooling to map sentences globally, occasionally dilutes the paramount importance of localized proper nouns within limited texts, pushing documents with thematic overlap higher than those with specific keyword presence.

The overarching reflection of this research answers the core objective: determining the capability of pretrained multilingual embedding models within data-scarce micro-environments. Dense retrieval models are commonly evaluated on large-scale corpora[4], [13].However, This study provides an initial empirical evaluation by proving that leveraging a robustly pretrained architecture like paraphrase-multilingual-MiniLM-L12-v2 enables system developers to sidestep local data limits. Even on a constraint as tight as 30 records, the embedding space holds sufficient semantic alignment to provide superior contextual recommendations. This establishes a highly critical engineering baseline for developing light, resource-friendly, yet intelligent recommendation pipelines where massive historical datasets are completely inaccessible. The superiority of MiniLM over TF-IDF observed in this study is consistent with findings reported by Guesmi et al. [1],who demonstrated that semantic embeddings improve recommendation relevance by capturing contextual relationships beyond exact lexical matching. Similarly, Huang [2]reported that semantic information enhances recommendation quality by providing richer contextual representations.

4. Conclusions

This study achieved its objective of comparing keyword-based retrieval and embedding-based retrieval approaches in a text-based hotel recommendation system using a small-scale dataset. Based on the experimental results, the paraphrase-multilingual-MiniLM-L12-v2 model outperformed the TF-IDF baseline, achieving a Mean Precision@3 score of 0.3667 compared to 0.3000 and a Mean Average Precision (MAP) score of 0.1378 compared to 0.1333. These results indicate that embedding-based retrieval is more effective in capturing semantic relationships between user queries and hotel descriptions, particularly when relevant concepts are expressed using different vocabulary.

The findings suggest that pretrained multilingual embedding models can remain effective even when applied to limited datasets, reducing the dependency on large-scale training data and extensive computational resources. From a practical perspective, this approach may assist developers of small-scale recommendation systems in improving retrieval quality without requiring complex training procedures or large historical datasets. Nevertheless, the experimental results also show that TF-IDF still performs competitively for queries containing explicit keywords and location-specific terms, indicating that lexical matching remains valuable in certain retrieval scenarios.

Future research may explore hybrid retrieval architectures that combine lexical and semantic representations in order to leverage the strengths of both approaches. In addition, experiments involving larger datasets, more diverse query sets, and additional embedding models could provide a broader understanding of retrieval performance in hotel recommendation systems and other text-based recommendation domains.

5. References

[1] M. Guesmi, M. A. Chatti, L. Kadhim, S. Joarder, and Q. U. Ain, “Semantic Interest Modeling and Content-Based Scientific Publication Recommendation Using Word Embeddings and Sentence Encoders,” *Multimodal Technol. Interact.*, vol. 7, no. 9, 2023, doi: 10.3390/mti7090091.

[2] R. Huang, “Improved content recommendation algorithm integrating semantic information,” *J. Big*

Data, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00776-7.

[3] M. Syarif, R. Sa'adah, M. R. A. Listi, and R. Manisha, "Analisis Perbandingan Algoritma BM25 dan TF-IDF untuk Temu Kembali Metadata Jurnal Indonesia pada Temujurnal . com," *J. ICT Inf. Commun. Technol.*, vol. 25, pp. 173–180, 2025.

[4] W. X. Zhao, J. Liu, R. Ren, and J. R. Wen, "Dense Text Retrieval Based on Pretrained Language Models: A Survey," *ACM Trans. Inf. Syst.*, vol. 42, no. 4, 2024, doi: 10.1145/3637870.

[5] M. L. Sidi and S. Gunal, "applied sciences A Purely Entity-Based Semantic Search Approach for Document Retrieval," *Appl. Sci. MDPI*, vol. 13, no. 18, 2023, doi: 10.3390/app131810285.

[6] M. A. Mazta, E. Saputra, and M. R. A, "'Perbandingan Kinerja TF-IDF dan Count Vectorization pada Sistem Rekomendasi Judul Skripsi Berbasis Content-Based Filtering,'" *J. Inform. Teknol. DAN SAINS*, vol. 7, no. 4, pp. 1807–1816, 2025.

[7] L. Wang *et al.*, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," *Assoc. Comput. Linguist.*, pp. 1–17, 2024, doi: 10.18653/v1/2024.acl-long.500.

[8] A. Duhan, Aryan; Singhal, Aryan; Sharma, Shourya; Neeraj; MK, "Semantic Search and Recommendation Algorithm," 2024. [Online]. Available: https://arxiv.org/abs/2412.06649

[9] V. Bellini *et al.*, "A qualitative analysis of knowledge graphs in recommendation scenarios through semantics-aware autoencoders," *J. Intell. Inf. Syst.*, pp. 787–807, 2024, doi: 10.1007/s10844-023-00830-z.

[10] N. Muennighoff, N. Tazi, N. Reimers, and H. Face, "MTEB : Massive Text Embedding Benchmark," pp. 2014–2037, 2023, doi: 10.18653/v1/2023.eacl-main.148.

[11] J. Wang, J. X. Huang, and J. Sheng, "An efficient long-text semantic retrieval approach via utilizing presentation learning on short-text," *Complex Intell. Syst.*, vol. 10, no. 1, pp. 963–979, 2024, doi: 10.1007/s40747-023-01192-3.

[12] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J. Y. Nie, *C-Pack: Packed Resources For General Chinese Embeddings*, vol. 1, no. 1. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657878.

[13] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," *Adv. Neural Inf. Process. Syst.*, 2021.

[14] Santoso, "Analisis Pertumbuhan Jumlah Kamar Hotel,Wisatawan dan Mahasiswa Perguruan Tinggi Pariwisata," vol. 12, no. 1, pp. 43–69, 2014.

[15] C. A. Melyani, A. Kesumawati, R. B. F. Hakim, and A. H. Primandari, "Hotel Recommendation System with Content-Based Filtering Approach (Case Study: Hotel in Yogyakarta on Nusatrip Website)," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 15, no. 1, pp. 152–157, 2022, doi: 10.36456/jstat.vol15.no1.a5375.

[16] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*, Advances in Neural Information Processing Systems, 2020, pp. 5776–5788.

[17] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 3982–3992, 2019, doi: 10.18653/v1/D19-1410.

[18] T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant, "SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval," in *Proceedings of ACM Conference (Conference'17)*, Washington, DC, USA: Association for Computing Machinery, 2021.

[19] S. Sivarajkumar *et al.*, *"Clinical Information Retrieval: A Literature Review,"* vol. 8, no. 2. Springer International Publishing, 2024. doi: 10.1007/s41666-024-00159-4.

[20] A. Moffat, "Batch Evaluation Metrics in Information Retrieval: Measures, Scales, and Meaning," *IEEE Access*, vol. 10, no. July, pp. 105564–105577, 2022, doi: 10.1109/ACCESS.2022.3211668.