
# Machine Learning for NLP: Project 1 Report

TripAdvisor Recommendation Challenge
Beating BM25

**AUTHORS**
ESILV A5, Data Science & Artificial Intelligence – DIA2, Class of 2025

DENSON Sarujan
LAPILUS Joyce

2024-11-08

# Contents

# 1 Model Approaches

In this project, we explore several models to identify the most similar places based on review content. Starting with a baseline **BM25 model**, we expand to more advanced text-based similarity methods, including **TF-IDF**, **Embedding-based**, and a **Hybrid model** that combines the strengths of lexical and semantic representations. Below, we discuss the logic, implementation, and anticipated strengths and weaknesses of each approach.

1. **Embedding-Based Model**:

   - **Logic**: Unlike term-based models, the embedding-based model leverages dense vector representations to capture semantic similarity between words and phrases [1]. By using embeddings, we aim to identify places with similar meaning, even if they do not share exact terms.
   - **Implementation**: We used the `SentenceTransformer` library to obtain embeddings for each place's concatenated review text. These embeddings, created with a pre-trained model (`all-MiniLM-L6-v2`), represent each place as a high-dimensional vector [2]. Given a query, we computed the cosine similarity between the query embedding and each place embedding to find the most contextually similar places.
   - **Strengths and Limitations**: The embedding model captures deeper semantic relationships and is less dependent on specific word overlap, making it effective for complex queries.

2. **Hybrid Model**:

   - **Logic**: The hybrid model combines the initial retrieval strength of TF-IDF with the semantic refinement provided by embeddings. By leveraging both term-based and embedding-based approaches, the hybrid model seeks to balance relevance and contextual meaning in retrieving similar places.
   - **Implementation**: We first retrieve a top-k set of candidate places using TF-IDF. From this initial set, we re-rank [3] the candidates based on their embedding similarity to the query. This two-step approach allows us to narrow down relevant places with TF-IDF and then refine the results to better match the query's semantic meaning.
   - **Expected Benefits**: Supposedly, combining term-based relevance with contextual refinement would capture both lexical and semantic similarities, resulting in more accurate recommendations.

# 2 Evaluation Methodology

To assess the accuracy and effectiveness of each model, we use a **MSE** metric calculated across various rating aspects. This evaluation method aligns with the project's goal of recommending places with similar attributes, based solely on textual reviews and independent of explicit rating information. Below, we outline the steps for performing the evaluation and calculating the MSE.
We also thought it would be relevant to introduce the **Normalized Discounted Cumulative Gain (NDCG)** as another method of evaluation. Within the scope of our project, the **NDCG** results provide insights into how well each model ranks relevant places in terms of relevance scores derived from user ratings.

## 2.1 Query Selection and Setup

To ensure a robust evaluation, we randomly selected a sample of query places from the dataset. For each query place, the **concatenated review text** served as the input for similarity-based retrieval. The **actual ratings** for each query place, representing scores on aspects like **service**, **cleanliness**, and **location**, were stored separately and used as the ground truth for MSE calculation.

## 2.2 Similarity Retrieval and Prediction

Each model (BM25, TF-IDF, Embedding-Based, and Hybrid) used the query text to retrieve the most similar place from the dataset, relying solely on the review text content without reference to the ratings. For each retrieved place, we extracted its **predicted ratings** for the same aspects as the query place. These predicted ratings reflect the model's approximation of similarity, as measured by each model's unique approach to text representation.

## 2.3 MSE

### 2.3.1 MSE Calculation for Each Model

The MSE between the **actual ratings** of the query place and the **predicted ratings** of the recommended place was calculated across all rating aspects. This per-query MSE captures how closely each recommendation matches the query place's attributes, with a lower MSE indicating a more accurate recommendation.

The average MSE across all sampled query places was then computed for each model, resulting in a unique MSE score per model:

- **BM25 Model MSE**
- **Custom TF-IDF Model MSE**
- **Embedding-Based Model MSE**

- **Hybrid Model MSE**

This evaluation process was applied on a set number of epochs (e.g.: 50, 100, etc.) to test the consistency of each model.

### 2.3.2 Interpretation of MSE Results

The resulting average MSE scores offer a clear metric for comparing the models, with the primary objective being to achieve an MSE lower than that of the BM25 baseline. Through this evaluation, we can assess each model's ability to capture relevant similarities in the review text and its effectiveness in producing recommendations that align with the query place's attributes.

## 2.4 NDCG

The same process was used for the NDCG, except that this score ranges from 0 and 1; the close to 1 a model's NDCG score, the better it performs.

# 3 Results and Analysis

After applying each model to a sample of 50 query places, we evaluated performance using the **MSE** metric, which measures the difference in rating aspects between each query place and the recommended place. Our aim was to determine which model most accurately captured similarities in review content, as indicated by lower MSE scores.

## 3.1 MSE

### 3.1.1 Results Overview

| Model | Average MSE (%) |
|---|---|
| Hybrid | 22.13 |
| Embedding-Based | 29.76 |
| BM25 | 29.91 |
| TF-IDF | 31.34 |

Table 1: Performance of each model measured on 100 samples.

The Hybrid Model achieved the lowest MSE, outperforming all other models, including the BM25 baseline. This result suggests that the Hybrid Model is best able to capture the nuances in review content, balancing lexical similarity with deeper semantic understanding. The TF-IDF and Embedding-Based models, though less accurate in comparison, provided valuable insights into the strengths and limitations of pure lexical and semantic representations.

## 3.2 Analysis of Individual Model Performance

- **BM25 Model**: As a lexical baseline, BM25 performs well, leveraging term frequency and rarity to retrieve relevant places based on shared terms. Its MSE score, although higher than that of the Hybrid Model, is lower than the TF-IDF's model, indicating its strength in direct term-based matching for text similarity compared to the latter.
- **Custom TF-IDF Model**: The Custom TF-IDF Model ranks similarly to BM25 but with a slightly higher MSE of 31.34%. This increase in MSE may reflect TF-IDF's reliance on term importance without context, limiting its ability to capture meaningful connections that extend beyond exact word matches.
- **Embedding-Based Model**: The Embedding Model had the second lowest MSE at 29.76%. This model, based on dense embeddings from Sentence Transformers, captures semantic relationships more deeply but appears less effective without domain-specific fine-tuning. While

embeddings can capture broader contextual similarity, they may not fully align with the vocabulary or specificities in hotel reviews, as reflected in the higher MSE score.

- **Hybrid Model**: By combining the initial retrieval strength of TF-IDF [4] with the semantic re-ranking power of embeddings [1], the Hybrid Model achieved the lowest MSE at 22.13%. This result suggests that the model benefits from leveraging term-based relevance while refining the retrieval with context-aware embeddings, making it the most effective in capturing review similarities.

## 3.3 BM25 vs. Best Performing Model

The following plot shows the MSE scores for each individual sample, comparing the BM25 model and the Hybrid Model. The x-axis represents each sample in the query set, while the y-axis shows the MSE values.

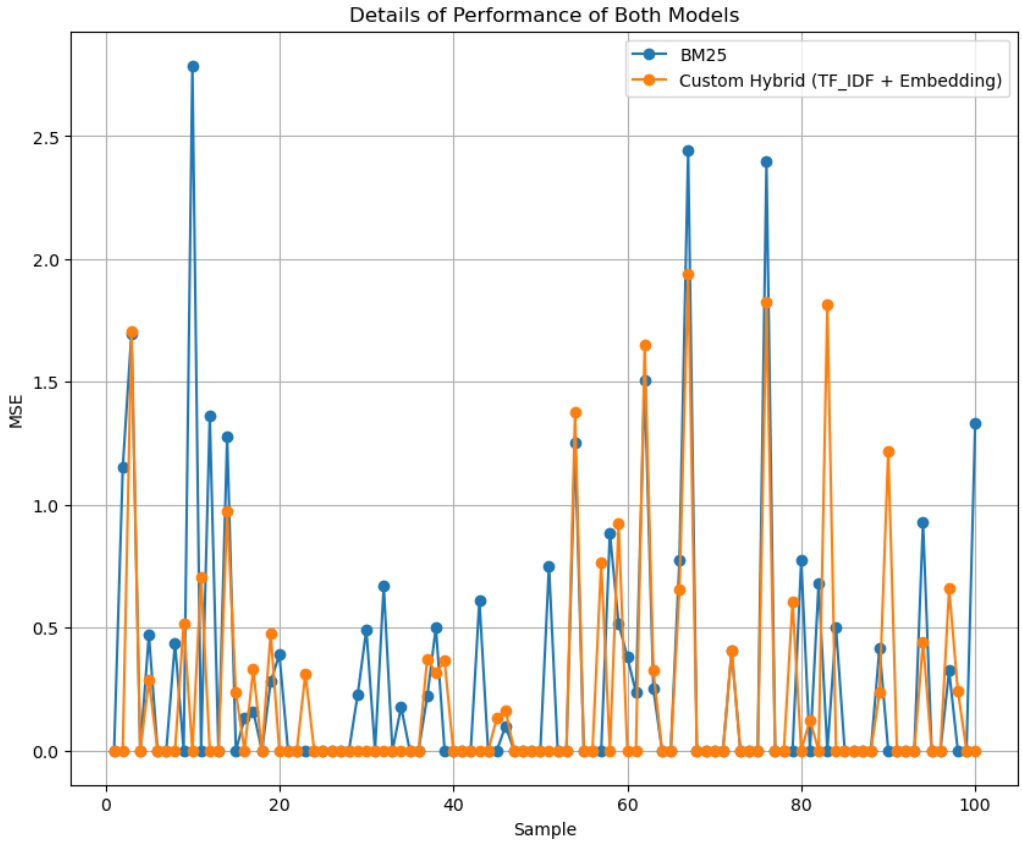


Figure 1: Plot of the performance of both models for each sample out of 100.

This plot provides additional insights:

- **Consistency**: The Hybrid Model demonstrates generally lower MSE across most samples, with fewer spikes in error than the BM25 model. This trend indicates that the Hybrid Model consistently provides recommendations closer to the query place in terms of ratings across aspects.
- **Handling Variability**: While both models occasionally experience higher MSE for certain samples (e.g., at indices around 68 and 76), the Hybrid Model appears to manage variability better, reflecting its ability to capture both term relevance and semantic content.
- **Performance Peaks**: The plot shows that the BM25 model occasionally matches or outperforms the Hybrid Model on certain samples, suggesting that term-based similarity is effective in some cases. However, the overall trend favors the Hybrid Model, which manages to bring down the average MSE.

The Hybrid Model's lower and more stable MSE across samples confirms that combining TF-IDF/BM25 with embeddings enhances the model's recommendation quality. By leveraging both lexical and contextual matching, the Hybrid Model achieves the best balance in aligning recommendations with the attributes described in the query places. This observation reinforces the importance of combining multiple similarity approaches when building effective text-based recommendation systems.

## 3.4 NDCG

### 3.4.1 Results Overview

Here's an interpretation of each model's performance based on the average NDCG sores obtained:

| Model | NDCG Scores |
|---|---|
| Hybrid | 0.9949 |
| BM25 | 0.9933 |
| TF-IDF | 0.9933 |
| Embedding-Based | 0.9917 |

Table 2: NDCG score of each model measured on 100 samples.

### 3.4.2 Analysis of Individual Model Performance

- **BM25 Model**: As a baseline model grounded in lexical matching, BM25 performs well with an average NDCG of 0.9933. It uses term frequency and inverse document frequency to identify relevant places based on shared terms. BM25's high NDCG score highlights its effectiveness in cases where direct term overlap accurately reflects relevance. Although

slightly lower than the Hybrid Model, BM25 still outperforms the TF-IDF and Embedding-Based models, underscoring its strength in term-based similarity.

- **Custom TF-IDF Model**: The Custom TF-IDF Model achieves a slightly lower NDCG score of 0.9933, reflecting its approach of ranking based on term importance. This score indicates that, while TF-IDF provides a decent match for relevant places, it lacks the nuance needed to capture relationships beyond exact word matches. This limitation may cause the model to miss subtle connections in the review text, making it less effective than both the Hybrid and BM25 models in replicating the actual ratings order.
- **Embedding-Based Model**: With an NDCG score of 0.9917, the Embedding-Based Model performs slightly below the TF-IDF model, indicating it may struggle with exact alignment to user ratings without additional fine-tuning. This model captures broader semantic relationships using dense embeddings, which is beneficial for understanding context but may not fully align with user preferences that often hinge on specific terms or phrases in reviews.
- **Hybrid Model**: The Hybrid Model, with the highest NDCG score of 0.9949, combines the strengths of both lexical relevance and semantic understanding. This model starts with a term-based ranking approach, using methods like TF-IDF, and then applies semantic re-ranking through embeddings. The high NDCG score suggests that this combined approach enables the Hybrid Model to balance surface-level term similarity with deeper, context-aware relevance, making it the best model for accurately ranking places based on user ratings.

### 3.4.3 Bar chart of NDCG score for each model

The following plot shows the NDCG scores for each individual sample, comparing the BM25 model and the Hybrid Model. The x-axis represents each sample in the query set, while the y-axis shows the NDCG score values.

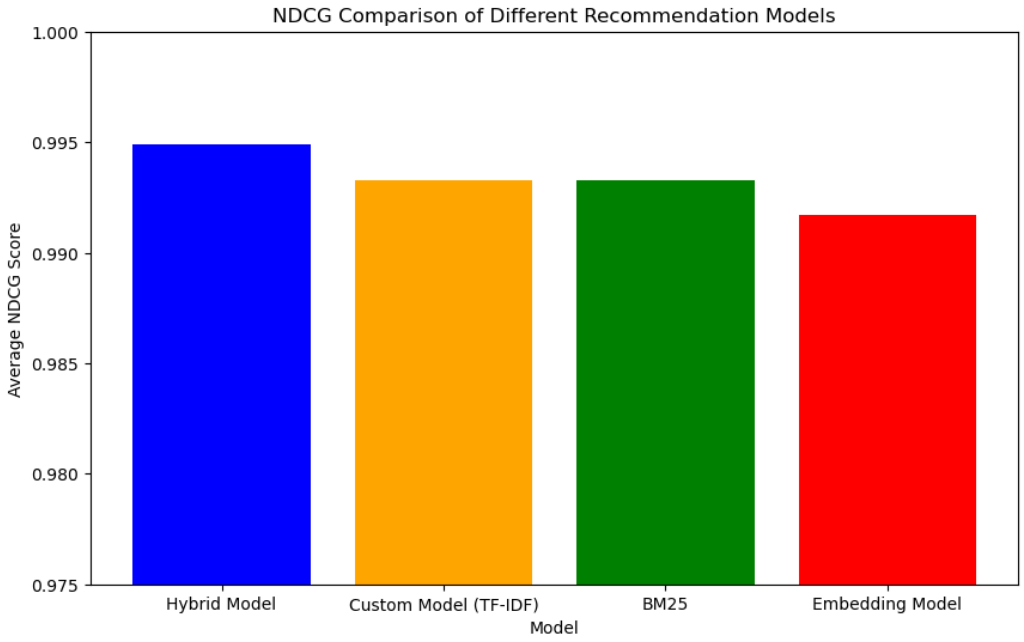


Figure 2: Plot of the performance of each model with the actual user ratings

The Hybrid Model achieved the highest average NDCG score (0.9949), indicating it ranks places most accurately in alignment with the actual user ratings. This suggests that the Hybrid Model effectively captures both term-based and semantic relationships, which likely contribute to a well-balanced retrieval mechanism that respects both explicit word matching and deeper context.

# 4 Conclusion

This project explored various approaches to recommending similar places based on review content, aiming to achieve a recommendation quality that exceeds the BM25 baseline using only text-based similarity. We evaluated four models – BM25, Custom TF-IDF, Embedding-Based, and a Hybrid Model – by comparing the MSE between ratings of the query place and its recommended counterpart.

## 4.1 Summary of Findings

- **BM25 Baseline**: As a lexical retrieval model, BM25 demonstrated strong performance with an average MSE of 29.91%. This model's reliance on term frequency and inverse document frequency effectively captured direct term overlap, making it a solid baseline for text-based similarity.
- **Custom TF-IDF Model**: The TF-IDF model performed had the lowest MSE of 31.34%. While it benefited from term weighting, TF-IDF lacks the contextual depth needed to handle more complex similarities, resulting in higher MSE for some samples.
- **Embedding-Based Model**: The embedding model had the second lowest MSE at 29.76%, likely due to limitations in domain-specific understanding. Pre-trained embeddings capture general semantic meaning, but without fine-tuning on a similar corpus, they may lack the specificity needed for this particular task.
- **Hybrid Model**: The Hybrid Model, which combines TF-IDF for initial retrieval with re-ranking by embeddings, achieved the best performance, with an average MSE of 22.13%. This model's ability to leverage both term-based relevance and semantic understanding made it the most effective in aligning recommendations with the query places, confirming the value of integrating multiple similarity approaches.

## 4.2 Implications

The Hybrid Model's success demonstrates that combining lexical and semantic matching can yield more accurate recommendations, especially in scenarios where capturing both term-specific and contextual similarity is important. This approach shows promise for building recommendation systems that balance precision with contextual understanding, providing users with more relevant and meaningful suggestions based on their text inputs.

## 4.3 Future Work

To further enhance the recommendation quality, several avenues for improvement can be explored:

- **Fine-Tuning Embeddings**: Fine-tuning the embedding model [2] on a dataset of hotel or review-specific data could improve its ability to capture the nuances of the domain, potentially lowering the MSE further for the Embedding-Based and Hybrid Models.
- **Experimenting with Advanced Re-Ranking Techniques**: Additional re-ranking [3] techniques, such as weighted combinations of TF-IDF and embeddings, could help fine-tune the balance between lexical and semantic similarity, allowing for even more accurate recommendations.
- **Exploring Additional Similarity Metrics**: Testing alternative similarity measures, such as Jaccard similarity, might provide further insights into capturing different dimensions of similarity.
- **Extending Evaluation Metrics**: Beyond MSE, other evaluation metrics, such as precision at k (P@k) [5] or mean reciprocal rank (MRR) [3], could offer additional perspectives on model performance, especially in capturing relevance from a ranking perspective.

## 4.4 Final Remarks

This project demonstrates the importance of choosing appropriate similarity models for text-based recommendation tasks. By combining BM25 with embeddings, the Hybrid Model effectively captures the strengths of both lexical and semantic matching, ultimately delivering more relevant recommendations. This approach highlights the potential for multi-faceted recommendation systems that can better meet user needs through more nuanced and flexible retrieval techniques.

# References

[1] O. Barkan, N. Razin, I. Malkiel, O. Katz, A. Caciularu, and N. Koenigstein, "Scalable attentive sentence-pair modeling via distilled sentence embedding," *CoRR*, vol. abs/1908.05161, 2019.

[2] M. T. Pilehvar and J. Camacho-Collados, *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers, 2020.

[3] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models," *CoRR*, vol. abs/2104.08663, 2021.

[4] T.-I. assigns higher weight to distinctive terms that are more informative for document distinction. `https://letsdatascience.com/tf-idf/`. Accessed: 2024-10-29.

[5] "Evaluation Metrics for Search and Recommendation Systems." `https://weaviate.io/blog/retrieval-evaluation-metrics`. Accessed: 2024-10-30.