ProfileRec: Efficient Accommodation Review Ranking using Sentence Embeddings and Nearest-Neighbor Search

Rajorshi Chaudhuri[1,*], Pranav Bhatki[* 1,†] and Yash Dubal[* 1,†]

[1]BookMyShow, Mumbai, India

Abstract

In this paper, we present our 2nd place solution of the RecTour 2024 Challenge. The competition task was to rank the reviews of accommodations for users based on the characteristics of the user and the accommodation. For the final solution, our team ”BMS Hunters”, used a combination of sentence embeddings with nearest-neighbor search. We achieved a final leaderboard score of 0.0829 for MRR@10 and 0.2679 for Precision@10. The full implementation of our model is available on GitHub. [1]

Keywords

Recommender Systems, RecTour 2024 Challenge, Tourism, Hotels, Accommodations, Reviews, SBERT, BallTree

1. Introduction

With the vast number of reviews available for popular accommodations, users often struggle to quickly discover the most relevant and helpful ones. An intuitive solution to this problem would be to display the reviews solely in chronological order or based on ”helpfulness” votes. However, this does not account for the preferences or specific needs of the individual user. For example, a family might value different aspects of a hotel compared to a solo traveler, and the ranked reviews should reflect these nuances. Furthermore, many reviews lack ”helpfulness” votes, leading to presentation bias where only a few reviews dominate visibility.

The RecTour 2024 Challenge[1], organized by Booking.com, focused on solving this problem by ranking reviews based on user and accommodation features. This aims to improve the booking experience by personalizing the review order to match the context of each user.

The rest of the paper is structured as follows. Section 2 provides an overview of the dataset. In Section 3, we discuss the baseline and various models used. Section 4 reports the experimental results of our proposed method. Finally, we conclude the paper with a summary of the proposed approach and its key findings.

2. Dataset Description

The competition training dataset consisted of three files:

1. **Users** - This file contains information regarding anonymized users and accommodation features. It has over 1.6 million unique rows and includes the following columns:
    - *user_id*: Unique identifier for the user.
    - *accommodation_id*: Unique identifier for the accommodation.
    - *guest_type*: Type of guest (e.g., solo, couple, family).

[1] https://github.com/Sirius79/rectour2024

Workshop on Recommenders in Tourism (RecTour 2024), October 18th, 2024, co-located with the 18th ACM Conference on Recommender Systems, Bari, Italy.

[*] Corresponding author.

[†] These authors contributed equally.

rajorshi.chaudhuri@bookmyshow.com (R. Chaudhuri); pranav.bhatki@bookmyshow.com (P. Bhatki[*]); yash.dubal@bookmyshow.com (Y. Dubal[*])

- *guest_country*: Country of the guest.
- *room_nights*: Number of room nights booked.
- *month*: Month of the booking.
- *accommodation_type*: Type of accommodation.
- *accommodation_country*: Country of the accommodation.
- *accommodation_score*: Overall score of the accommodation.
- *accommodation_star_rating*: Star rating of the accommodation.
- *location_is_ski*: Indicator if the location is a ski resort.
- *location_is_beach*: Indicator if the location is a beach destination.
- *location_is_city_center*: Indicator if the accommodation is located in the city center.

2. **Review** - This file contains information regarding reviews. It has over 1.6 million rows and includes the following columns:
    - *review_id*: Unique identifier for the review.
    - *accommodation_id*: Unique identifier for the accommodation being reviewed.
    - *review_title*: Title of the review.
    - *review_positive*: Positive comments in the review.
    - *review_negative*: Negative comments in the review.
    - *review_score*: Score given in the review.
    - *review_helpful_votes*: Number of helpful votes the review received.
3. **Matches** – This file contains the true labels for the dataset, representing positive examples of user-accommodation-review relationships. It has over 1.6 million rows and includes the following columns:
    - *user_id*: Unique identifier for the user.
    - *accommodation_id*: Unique identifier for the accommodation.
    - *review_id*: Unique identifier for the review.

Each accommodation in the dataset is associated with a minimum of 10 unique reviews. To ensure data quality, each review is analyzed for at least 3 distinct topics using the text2topic method[2]. Consequently, reviews that are too simplistic, such as those containing only the word ”awesome,” are filtered out because they do not provide sufficient informative content.

3. Model Architectures

Our approach to the RecTour 2024 challenge evolved through several stages. We describe our three main models below:

3.1. Baseline

Our initial approach was a simple ”review score” based ranking model. This model assumed that reviews with higher ratings were more informative and helpful to users. Consequently, the reviews were ranked in the descending order of their scores for each accommodation, without incorporating additional user or accommodation characteristics. As a baseline model, it provided a reference point for evaluating more advanced models.

3.2. Sentence-BERT for User and Review Embeddings

Although the initial review score-based model provided a solid baseline, it did not take advantage of the textual content of the reviews or the detailed profiles of users and accommodations. So, to improve the personalization of review rankings for users, we wanted to incorporate the review content.

To achieve this, we selected Sentence-BERT[3] for its ability to generate high-quality sentence embeddings. Sentence-BERT is particularly well suited for capturing the semantic meaning of sentences, which is crucial for understanding and ranking reviews based on their relevance to user preferences.

We used the following features to generate user and review embeddings:

- **User Features:** guest_type, guest_country, accommodation_type, accommodation_country.
- **Review Features:** review_title, review_positive, review_negative.

For user embeddings, we selected these features because we did not have a direct user history. Instead, we aimed to create user cohorts based on the guest type (e.g., solo, family, etc.), the origin of the guest, and the common travel patterns between guest types and the host countries they usually visit. This allowed us to approximate user preferences more effectively.

We then encoded these user profiles and review texts using the Sentence-BERT model. By computing the cosine similarity between the encoded user profiles and reviews, we ranked the reviews according to their relevance to each user's profile. This method allowed us to capture deeper contextual relevance between users and reviews beyond simple scoring.

3.3. ProfileRec: Sentence-BERT + BallTree

Our final model, ProfileRec, extended the previous approach by incorporating additional user features and utilizing BallTree for efficient nearest-neighbor search. BallTree is a spatial data structure that partitions the data into a binary tree for faster querying of nearest neighbors. This is particularly advantageous when working with embeddings generated by Sentence-BERT, as it speeds up the process of finding similar items.

In addition to the features used in the previous model, we included ”room_nights” and ”month” as additional user features. The decision to incorporate room nights was informed by our observation of a positive trend between review ratings and the number of room nights booked, as illustrated in Figure 1.

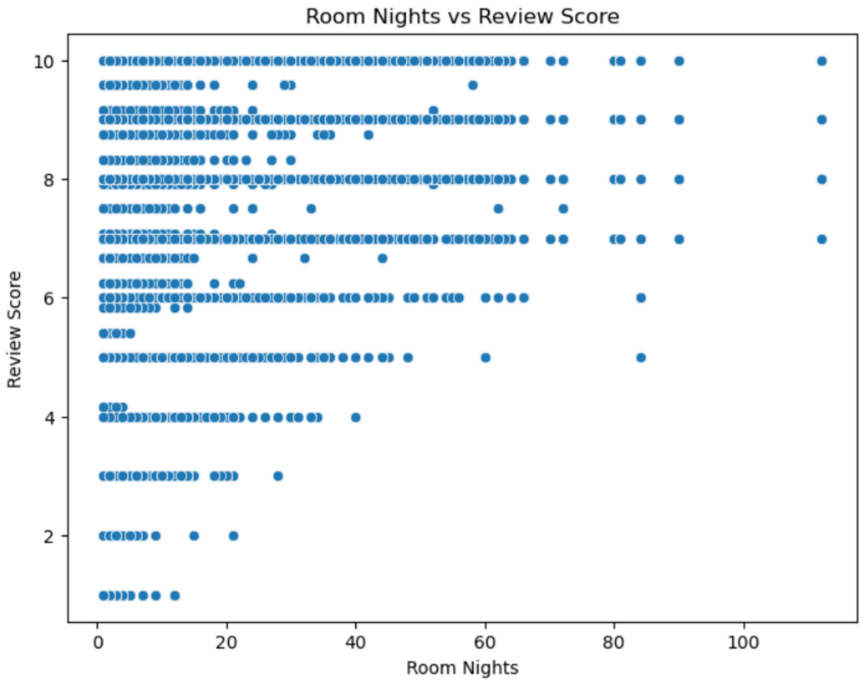

Figure 1: Scatter plot showing the positive trend between review ratings and room nights.

This scatter plot shows that accommodations with higher review scores tend to have more room nights, suggesting that users who stay longer might leave more detailed and potentially higher-rated reviews. We included ”month” as a temporal feature to account for potential seasonal variations in review scores and booking patterns.

The inclusion of these features, combined with the BallTree-based nearest-neighbor search, enhanced the ability of the model to rank reviews based on a more nuanced understanding of user preferences and accommodation characteristics.

4. Results

Table 1 shows the performance of all our models on the test dataset. The results show that the ProfileRec model, which incorporates Sentence-BERT embeddings and BallTree for efficient nearest-neighbor search, outperforms the baseline model and the Sentence-BERT only approach.

Table 1
Performance Comparison on Validation and Test Datasets

Method	MRR@10	Precision@10
ProfileRec	**0.0829**	**0.2679**
SBERT	0.0798	0.2621
Baseline	0.0735	0.2511

5. Conclusion

In this paper, we present our solution for the RecTour 2024 Challenge. Our final approach, ProfileRec, combined Sentence-BERT embeddings with BallTree for an efficient nearest-neighbor search, significantly enhancing ranking accuracy. We hope that our approach provides valuable insights and contributes to the development of cohort-based recommendation systems in the RecSys field.

Acknowledgments

The authors wish to thank the organizers of the RecTour 2024 Challenge for the opportunity to participate in this exciting competition.

References

[1] A. Livne, E. Fainman, Booking.com rectour 2024 challenge, in: ACM RecSys RecTour ’24, Bari, Italy, 2024. Retrieved from https://workshops.ds-ifs.tuwien.ac.at/rectour24/rectour-2024-challenge/.

[2] F. Wang, M. Beladev, O. Kleinfeld, E. Frayerman, T. Shachar, E. Fainman, K. L. Assaraf, S. Mizrachi, B. Wang, Text2topic: Multi-label text classification system for efficient topic detection in user generated content with zero-shot capabilities, in: M. Wang, I. Zitouni (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Singapore, 2023.

[3] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.