Accommodation Review Ranking for Tourism Recommendation

Emrul Hasan[1,2,*], Chen Ding[1], Sajib Saha[3], Neelima Monjusha Preeti[4] and Abdul Halim[5]

[1]Toronto Metropolitan University, Toronto, Canada
[2]Vector Institute, Toronto, Canada
[3]Daffodil International University, Dhaka, Bangladesh
[4]Jahangirnagar University, Dhaka, Bangladesh
[5]RTM Al-Kabir Technical University, Sylhet, Bangladesh

Abstract

The primary goal of tourism management platforms, e.g. booking.com, is to provide the best matches to travelers. User-generated reviews are often leveraged in various ways to influence user's decision-making process. One straightforward approach is ranking historical reviews based on user's preferences by checking the "helpfulness" votes. A major issue with this approach is that many reviews do not receive helpfulness votes, leading to a presentation bias. In this work, we incorporate multiple review features to rank reviews based on user profiles. Reviews are encoded using a state-of-the-art transformer encoder model (e.g., SBERT), and cosine similarity is computed between user profiles and reviews. The ranking performance is assessed with MRR@10 (Mean Reciprocal Rank) and Precision@10. Our results demonstrate that beyond helpfulness votes, leveraging additional features (e.g., accommodation type, review title, positive aspects of reviews, etc.) significantly improves performance. The implementation of our method is available on the GitHub [1]

Keywords

Review Ranking, Helpfulness, Sentence Transformer, Tourism

1. Introduction

The rapid growth and success of e-commerce have sparked extensive research aimed at improving customer engagement [1]. A key strategy to achieve this goal is enabling customers to share their experiences through reviews on product pages. These reviews serve as the primary source of information that bridges the gap between products and consumers [2]. Many e-commerce platforms including tourism, social media, education, health, etc. leverage customer reviews on products to identify user preferences as well as making important business decisions [3, 1, 4]. As noted by Paget et al. [5], 80% of customers rely on past reviews while deciding whether to purchase a product. A single product or service may receive hundreds or even thousands of reviews, it becomes difficult to make a decision by going through all of them. Additionally, the same survey from BrightLocal.com [5] indicates that 90% of consumers read at most ten or fewer reviews before making their decisions on whether to buy a product. Therefore, there is an urgent need to develop such a framework that ranks the most relevant reviews at the top page. Displaying high-quality reviews at the top can save user's time by providing the most valuable information from a few key reviews.

In the traditional approach, reviews are ranked based on either recency or helpfulness. Helpful reviews are those that receive positive feedback from other users who purchased the same product, while recent reviews are sorted by the time they were posted [6]. The primary issue with this approach is that most reviews do not receive their helpfulness votes, leading to a presentation bias. The helpfulness of online reviews is determined by posing a straightforward question: "Was this review helpful to you?" Users provide their feedback by selecting either a "thumbs up" or "thumbs down" button. However, given the vast number of reviews for each product, helpfulness voting does not address all challenges.

[1]https://github.com/emrulhasan-nlp/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.

In fact, sorting reviews based on "helpfulness" votes is subject to the Matthew effect which states that a review positioned at the top of the list tends to stay there because users primarily interact with and vote on the top reviews when making purchase decisions. In contrast, a review placed at the bottom remains unnoticed, as users rarely scroll down to see it[7].

To address this issue, many e-commerce platforms have introduced various methods to filter reviews and enhance customer satisfaction. Hsieh et. al. [8] introduce a SVR (Support Vector Regression)-based online customer review ranker that focuses on leveraging linguistic features to rank the reviews. Amazon has recently introduced a "top reviews" sorting feature, which allows further filtering based on categories like all positive (4- or 5-star ratings) or all negative (1- to 3-star ratings) reviews. However, the exact algorithm that Amazon uses to determine top reviews is not publicly disclosed [9].

In the RecTour 2024 Challenge organized by Booking.com [10], we introduce a feature integration method that leverages Sentence BERT [11] and utilize cosine similarity measures to rank the reviews. We show that incorporating additional review features, beyond just helpfulness votes, can improve the ranking quality. The core contributions of this paper are as follows:

1. We propose a sentence transformer-based feature extraction method for review ranking. Our approach involves generating user and review profiles by concatenating user and review information, respectively, and then encoding both profiles using a sentence transformer. We compute the cosine similarity between the user profile and accommodation reviews to determine ranking.
2. Our method is evaluated using Booking.com datasets and shows improved performance compared to traditional methods based solely on helpfulness votes.

2. Related Work

Review ranking is considered as one of the effective methods to enhance the user engagement in the e-commerce business. Numerous studies have been conducted on personalized review ranking [9, 12]. Traditional methods often rank reviews based on review helpfulness, particularly helpfulness votes [13, 14]. Sunil et. al. [9] utilize textual review to predict the helpfulness score. Helpfulness scores are predicted using features from review text, product descriptions, and customer Q & A data, with a random forest classifier and gradient boosting regressor. Reviews are classified as low or high quality by the classifier, and scores for high-quality reviews are predicted using the regressor. Wu [15] introduces a framework that jointly estimates the impact of review popularity and helpfulness. However, these approaches suffer from presentation bias as many reviews receive no votes and may be overlooked [14].

Integration of diverse user feedback has proven highly effective in understanding user preferences and item characteristics for recommendation systems [16, 17, 18, 19]. These studies harness the power of user reviews to gain insights into user preferences and item features, contributing to a comprehensive understanding of overall user inclinations towards products.

Recent advancements in natural language processing have yielded sophisticated text encoding models. BERT (Bidirectional Encoder Representations from Transformers) [20] and its variants, like Sentence-BERT [11], have markedly enhanced the ability to capture semantic nuances and contextual information from text. These advanced models produce multidimensional embeddings that encapsulate the semantic essence of reviews, enabling more refined analysis and ranking processes. In this paper, we propose a sentence transformer-based feature extraction method for review ranking. In addition, we integrate various forms of reviews to create user and item profiles to enhance the ranking quality.

3. Methodology

3.1. Preliminaries

Consider a user, identified by *user_id*, and an accommodation, identified by *accommodation_id*. Each accommodation is associated with multiple reviews, each with a unique *review_id*. For a specific user and accommodation, the objective is to retrieve the top 10 most relevant reviews. The method involves

three components: 1) user and accommodation profile creation, 2) feature extraction, and 3) similarity search and ranking.

3.2. User and Accommodation Profile Creation

The first step of the review ranking process is the user and accommodation profile construction. As a baseline approach, we rank the review based on the helpfulness vote. To achieve this, for each user and accommodation, we create a list of pairs with review IDs and "helpfulness" vote counts. Then we sort them in descending order and the top-voted review ids are kept. Similar to the "helpfulness" vote, we follow the same strategy when using the review scores for review ranking.

As the next step, we hypothesize that accommodation review profiles are influenced by attributes such as *review_title*, *review_positive*, and *review_negative* while user profiles are linked to *guest_type*, *guest_country*, *accommodation_type*, *accommodation_country*, *location_is_ski*, and *location_is_city_cent*. To build comprehensive user and accommodation profiles, we concatenate these features. We experiment with several combinations of features for both user and accommodation profiles. The detailed combination of different features is discussed in section 4.2.

3.3. Feature Extraction

We apply pre-trained sentence-BERT for feature extraction from the user and accommodation profile. Sentence-BERT is a variant of the pre-trained BERT [20] model that utilizes siamese and triplet network architectures to generate semantically meaningful sentence embeddings, enabling comparison through cosine similarity [11]. It is a cutting-edge text encoder model and a widely used method for generating text embeddings. Both the user and accommodation profiles are encoded using SBERT.

3.4. Similarity Search and Ranking

To compute the similarity score between each user and the list of reviews for accommodation, we use the cosine similarity measure. For a given user u and a review v, the similarity score can be computed as follows.

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|} \tag{1}$$

where u and v are the vector representations of the user and review. The detailed workflow of the method is shown in Figure 1

4. Experimental Details

4.1. Datasets

In this work, we use the dataset provided from the RecTour 2024 Challenge organized by Booking.com [10]. The framework is evaluated with the test dataset that contains the user and accommodation information. The total number of unique users and accommodations are 199138 and 5000 respectively. A detailed description of the dataset is shown in Table in the Appendix.

4.2. Experimental Settings

We perform several experiments using different combinations of features. In all cases, for feature extraction, we use the smaller and faster version of SBERT [11] e.g., "all-MiniLM-L6-v2" which has 6 transformer layers and fewer parameters. Following are the details of each of these experiments.

Experiment 1 and 2: In this case, for each user and accommodation, we sort the reviews based on "review_helpful_votes", and "review_score" respectively. Next, we retain the top 10 reviews for each User ID and accommodation ID pair.

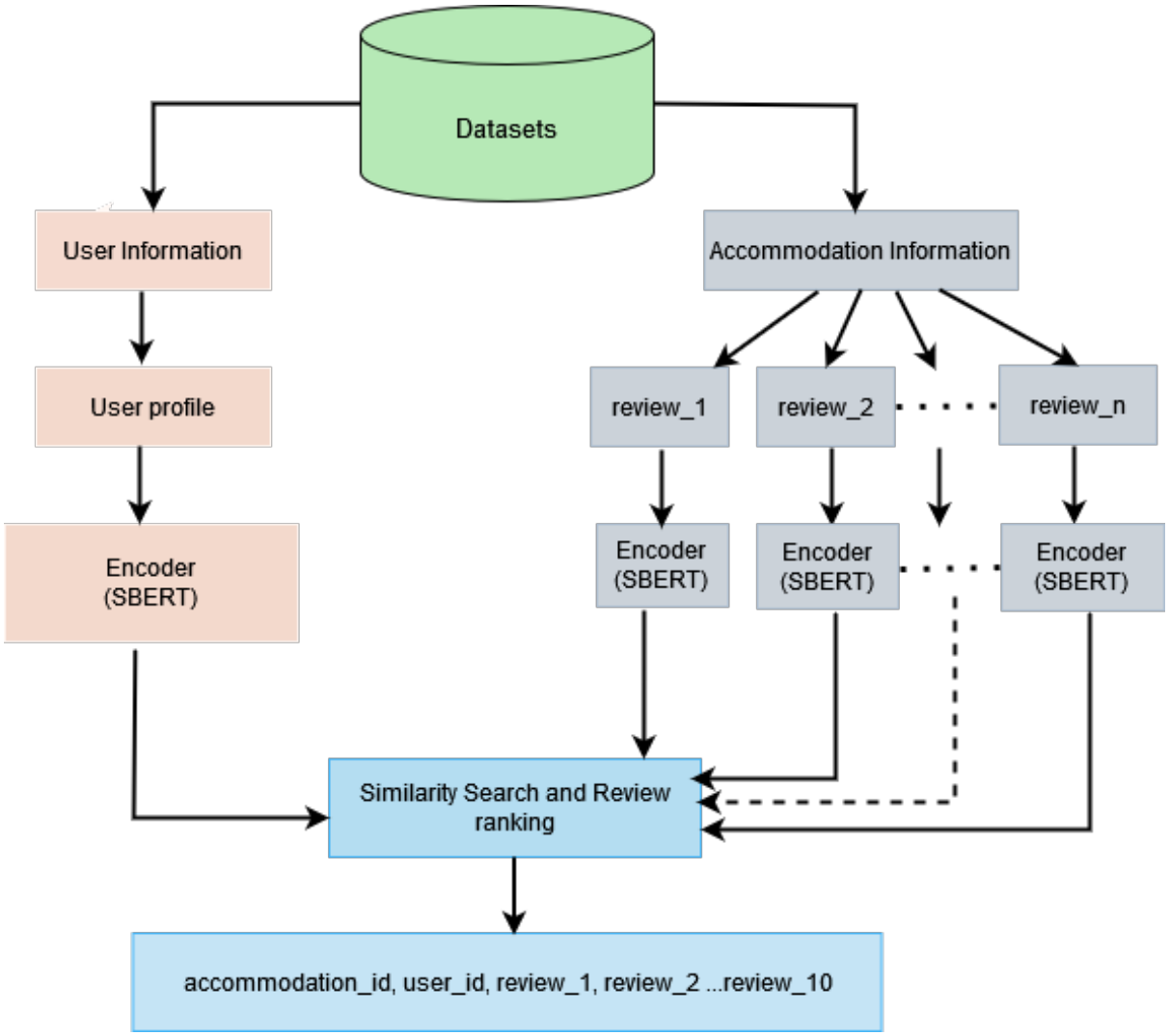

Figure 1: Detailed workflow of ranking review.

Experiment 3: In experiment 3, user profiles are constructed by concatenating the diverse set of review features including “guest_type”, “guest_country”, “room_nights”, “month”, “accommodation_type”, “accommodation_country”, “accommodation_score”, “accommodation_star_rating”, “ location_is_beach ”, “location_is_ski”, and “location_is_city_center” respectively. Similarly, accommodation profiles are created using various review features such as “review_title”, “review_positive”, “review_negative”, “review_score”, and “review_helpful_votes”. Each accommodation has a list of reviews. Using a sentence transformer, both the user profile and all the reviews from an accommodation are embedded. Then, the similarity score between each user profile and all the reviews is calculated. Finally, based on these similarity scores, the top 10 corresponding review IDs are retrieved.

Experiment 4 and 5: Similar strategy is applied for experiment 4 and 5. However, in experiment 4, accommodation profiles are kept the same as in experiment 3, but user profiles are shrunk to only “guest_type” and “guest_country”. On the other hand, experiment 5 keeps the user profile the same but the accommodation profile is added with more features, e.g. “accommodation_type”, “accommodation_country”, “accommodation_score”, “accommodation_star_rating”, “ location_is_beach ”, “location_is_ski”, and “location_is_city_center” respectively.

Table 1
Performance with different scenarios of criteria ratings

Experiment Number	MRR@10	Precision@10
Exp 1	0.0735	0.2511
Exp 2	0.0735	0.2511
Exp 3	0.0775	0.2582
Exp 4	0.0735	0.2511
Exp 5	**0.0787**	**0.2605**

4.3. Evaluation Metrics

We evaluate the performance of our ranking system using MRR (Mean Reciprocal Rank) and Precision, specifically MRR@10 and Precision@10, which indicate the MRR and Precision of the top 10 ranked reviews.

MRR (Mean Reciprocal Rank)@K
MRR@K measures the quality of a ranking system by focusing on the position of the first relevant item from the top K retrieved items. For a user u, the reciprocal rank is:

$$\text{Reciprocal Rank} = \frac{1}{\text{rank}_u}$$

where rank_u is the rank position of the first relevant item.

The Mean Reciprocal Rank across all U is:

$$\text{MRR@K} = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{1}{\text{rank}_u}$$

MRR@K is useful for evaluating how quickly the first relevant result appears. If no relevant item is found within the top K, the reciprocal rank for that user is 0. A higher MRR indicates better performance in placing relevant items early in the ranking. When using MRR@10, we focus specifically on the top 10 positions to evaluate performance within this subset.

Precision@K
Precision@K is defined as the ratio of correctly identified relevant items to the total number of items recommended in a list of length K.

$$\text{Precision@K} = \frac{\text{Total Number of relevant items in the top K}}{|K|}$$

where K is the number of items in the ranked list.

A higher MRR@K score shows that the first relevant result appears early in the ranking, while a higher Precision@k score indicates a higher proportion of relevant items in the top K results. MRR@K is best for situations where finding the first relevant item quickly is important, whereas Precision@K is better for evaluating systems that need to return multiple relevant items.

5. Results and Discussion

The performance of our method is presented in Table 1. It is obvious from the results that when we use only helpfulness vote or review score, both the MRR@10 and the Precision@10 are low compared to the addition of features. In experiment 5, MRR@10 and Precision@10 are 0.0787 and 0.2605, respectively, while in experiment 1, they are 0.0735 and 0.2511, which indicates that MRR and Precision are 0.52% and 0.94% higher in experiment 5 compared to experiment 1. Experiment 1 and 2 give similar performance for both metrics, which indicates that both helpfulness vote and reviews score have similar impact on

review ranking. Experiment 3 shows slightly better performance than Experiment 4, suggesting that the addition of features captures users' comprehensive behavior. This highlights the importance of incorporating textual reviews when modeling user behavior and item characteristics.

Therefore, we can conclude that user-generated textual reviews contain user and item-representative features and leveraging them to review ranking improves the ranking quality compared to traditional helpfulness vote or review score-based ranking methods.

6. Conclusions

In this work, we present a review ranking method for accommodation. We incorporate several forms of review features to create user and accommodation profiles. Sentence Transformer is applied to extract the features from the review profiles, and finally, the cosine similarity score between the user profile and each review is measured to rank the reviews. The performance of the method is evaluated with MRR (Mean Reciprocal Rank) and Precision@K. Our results reveal that integration of linguistic features to create user and accommodation profiles outperforms the vanilla helpfulness vote-based review ranking method. In the future, we aim to leverage state-of-the-art LLMs to rank the reviews.

Acknowledgments

This project is partially sponsored by the Natural Science and Engineering Research Council of Canada (grant 2020-04760).

References

[1] S. Chhabra, Netflix says 80 percent of watched content is based on algorithmic recommendations, Mobile Syrup: Toronto, ON, USA (2017).

[2] K. Heinonen, Consumer activity in social media: Managerial approaches to consumers' social media behavior, Journal of consumer behaviour 10 (2011) 356–364.

[3] L. Sivaranjani, S. K. Rachamadugu, B. S. Reddy, B. R. A, M. Sakthivel, S. Depuru, Fashion recommendation system using machine learning, in: 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), 2023, pp. 1367–1374. doi:10.1109/ICOSEC58147.2023.10275967.

[4] A. K. Lahiri, E. Hasan, Q. V. Hu, C. Ding, Tmu at trec clinical trials track 2023, arXiv preprint arXiv:2403.12088 (2024).

[5] S. Paget, Local consumer review survey 2023, Online verfügbar unter https://www. brightlocal. com/research/local-consumer-review-survey/, zuletzt geprüft am 19 (2023) 2023.

[6] A. Ghose, P. G. Ipeirotis, Designing novel review ranking systems: predicting the usefulness and impact of reviews, in: Proceedings of the ninth international conference on Electronic commerce, 2007, pp. 303–310.

[7] R. K. Merton, The matthew effect in science: The reward and communication systems of science are considered., Science 159 (1968) 56–63.

[8] H.-Y. Hsieh, S.-H. Wu, Ranking online customer reviews with the svr model, in: 2015 IEEE international conference on information reuse and integration, IEEE, 2015, pp. 550–555.

[9] S. Saumya, J. P. Singh, A. M. Baabdullah, N. P. Rana, Y. K. Dwivedi, Ranking online consumer reviews, Electronic commerce research and applications 29 (2018) 78–89.

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

[11] N. Reimers, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[12] N. Adler, L. Friedman, Z. Sinuany-Stern, Review of ranking methods in the data envelopment analysis context, European journal of operational research 140 (2002) 249–265.

[13] J. Qin, P. Zheng, X. Wang, Comprehensive helpfulness of online reviews: A dynamic strategy for ranking reviews by intrinsic and extrinsic helpfulness, Decision Support Systems 163 (2022) 113859.
[14] A. Melleng, A. Jurek-Loughrey, D. Padmanabhan, Ranking online reviews based on their helpfulness: An unsupervised approach, in: International Conference on Recent Advances in Natural Language Processing: Proceedings, 2021, pp. 959–967.
[15] J. Wu, Review popularity and review helpfulness: A model for user review effectiveness, Decision Support Systems 97 (2017) 92–103.
[16] H. Hwangbo, Y. S. Kim, K. J. Cha, Recommendation system development for fashion retail e-commerce, Electronic Commerce Research and Applications 28 (2018) 94–101.
[17] E. Hasan, C. Ding, A. Cuzzocrea, Multi-criteria rating and review based recommendation model, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 5494–5503.
[18] E. Hasan, C. Ding, Aspect-aware multi-criteria recommendation model with aspect representation learning, in: 2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, 2023, pp. 268–272.
[19] E. Hasan, M. Rahman, C. Ding, J. X. Huang, S. Raza, Review-based recommender systems: A survey of approaches, challenges and future perspectives, 2024. URL: https://arxiv.org/abs/2405.05562. arXiv:2405.05562.
[20] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

7. Appendix

Table 1: Dataset description: review, user and accommodation fields

Field name	Description
review_id	Anonymized review ID
review_title	Review title
review_positive	Positive (”liked”) section in review
review_negative	Negative (”disliked”) section in review
review_score	Overall review score for the stay
review_helpful_votes	How many users marked the review as helpful
user_id	Anonymized user ID
guest_type	There are 4 traveller types: Solo traveller (1 adult) / Couple (2 adults) / Group (>2 adults) / Family with children (adults & children)
guest_country	Anonymized country from which the reservation was made
room_nights	The length of the reservation, i.e. number of nights booked
month	The month of the check-in date of the reservation
accommodation_id	Anonymized accommodation ID
accommodation_type	The type of the accommodation, e.g. hotel, apartment, hostel
accommodation_score	The overall average guest review score of the accommodation
accommodation_country	Country of the accommodation
accommodation_star_rating	Accommodation star rating is provided by the property, and is usually determined by an official accommodation rating organisation or another third party
location_is_beach	Is the accommodation located in a beach location
location_is_ski	Is the accommodation located in a ski location
location_is_city_center	Is the accommodation located in the city center