Ranking Roughly Tourist Destinations Using BERT-Based Semantic Search

Myeong Seon Kim, Kang Woo Lee, Ji Won Lim, Da Hee Kim, and Soon-Goo Hong

Abstract With the development of the Internet, extensive online reviews have been generated in the tourism field. To utilize big data and avoid the shortcomings of traditional techniques, we designed a tourist destination ranking system by introducing a semantic search that extracts data related to the input query. For this system, we reviewed tourist spot reviews and pre-processed text reviews. Then, we embed the corpus and query using SBERT, measure their similarity, and leave data similar to the query above the threshold. By implementing a count-based ranking algorithm with the data within the boundary, the tourist destinations are derived in a semantically similar order to the query. We entered three queries to obtain the top 5 relevant tourist destinations. Although there are problems with optimal thresholds and imbalanced data, semantic search derives information of desired conditions and may be referenced in policymaking and recommendation systems.

Keywords Semantic search · BERT · Ranking system · Tourism destination

1 Introduction

With the growth of the Internet, there have been a larger number of online tourism reviews. Tourists' decision-making is influenced by the travel experiences of other individuals that are often form, online review comments [1]. Therefore, the analysis of online review data is essential. It can help tourism practitioners and stakeholders reflect results in the policymaking process and raise policy issues by identifying

M. S. Kim · D. H. Kim
Department of Computer Engineering, Dong-a Univ. 37, Nakdong-daero 550beon-gil, Saha-gu, Busan, South Korea

K. W. Lee · J. W. Lim
Smart Governance Research Center, Dong-a Univ, 225 Gudeok-ro, Seo-gu, Busan, South Korea

S.-G. Hong (✉)
Department of Management Information System, Dong-a Univ. 225 Gudeok-ro, Seo-gu, Busan, South Korea
e-mail: shong@dau.ac.kr

S. Shakya et al. (eds.), *Sentiment Analysis and Deep Learning*,
Advances in Intelligent Systems and Computing 1432,
https://doi.org/10.1007/978-981-19-5443-6_1

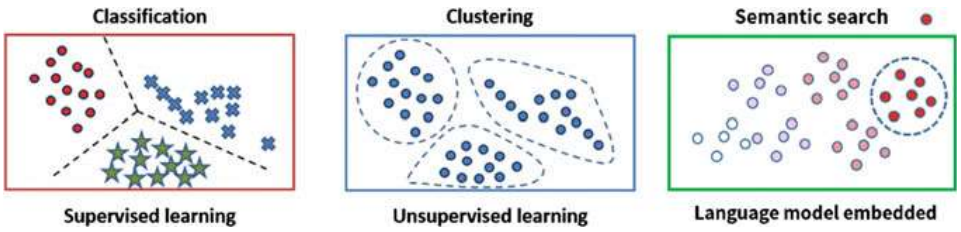

Fig. 1 Comparison of classification, clustering, semantic search [2]

which tourist destinations are popular and which are problematic. Alternatively, it may provide a practical recommendation system that helps tourists have a better travel experience.

Techniques for extracting meaningful information from big data already exist. For classification with these techniques, computers are employed to automatically classify input objects according to their content in predefined classes. This requires the use of labeled data to be used for training in supervised learning. Because it is difficult to obtain large datasets including labels, this method is both demanding and expensive. Clustering is a technique that identifies similarities between objects and groups them according to their characteristics. However, because clusters are ambiguous, the accuracy of these clustering results may be poor.

To avoid the shortcomings of traditional techniques, we introduced a semantic search to effectively extract the necessary information. When a query is entered, the semantic search returns a score for the semantic similarity between the query and corpus. By filtering data below a certain score, it is possible to coarsely define a boundary that contains data related to a common query. This study uses semantic search with this property to design a ranking system that shows tourist attractions that are related to a specific query in order as shown in Fig. 1.

2 Semantic Search with SBERT

2.1 Semantic Search with SBERT

Semantic search, which can be distinguished from lexical search based on literal matches without the semantic understanding of a query, denotes search with an understanding of the query meaning [3]. In traditional search engines, which only find documents based on lexical matches, semantic search engines can also find semantically similar sentences [4]. Therefore, it provides more significant search results by assessing and understanding the search phrase and discovering the most appropriate results in a website, database, or any other dataset [5].

Word embedding is word representation that allows words with similar meanings to have a similar representation. Using bidirectional encoder representations from

transformer (BERT)-based word embedding, a sentence (or word) can be represented in the dense form of a vector. It allows the identification of sentences similar to a search query by measuring the distance in the vector space. Sentence-BERT (SBERT) is a modification of the BERT network using Siamese and triplet networks to improve its efficiency and suitability for semantic searches [6].

2.2 Extracting a Boundary by Applying Different Thresholds

To demonstrate the properties of semantic search based on semantic similarities, we extract a boundary of vectors by applying different threshold values. If the threshold is high, the boundary can omit target sentences (signals) that must be included, whereas non-target sentences (noise) are less likely to be included. On the other hand, if the threshold is low, the boundary contains most of the target sentences but may also include too many non-target sentences. The threshold is key to determining the signal-to-noise ratio. Therefore, it is important to select an appropriate threshold for the search task.

To provide an intuitive understanding of the change in boundary changes according to the threshold values, we applied a semantic search algorithm to a tourist destination, Gwangalli Beach [7]. The results are shown in Fig. 2. They show that the range of boundaries is formed by the query "It is a beautiful beach" with different 3 thresholds (0.7, 0.5, and 0.3). The review data on the beach were embedded in the SBERT. There are 1943 reviews from TripAdvisor [8], and they are labeled according to whether they are similar to the query. In the graph, the colored points indicate queries within the boundary. The blue line is the matched target (signal), and the red line is the non-matched target (noise). The figure shows that the boundary range increased as the threshold decreased.

For a more formal explanation, we measured the signal-to-noise ratio (SNR). The SNR is a measure that compares the detection of the desired signal with the detection of background noise [9]. Table 1 provides detailed information such as the number of signals, noise, and SNR. The SNR was the highest, 14.50, when the threshold was 0.7. The SNR is dramatically reduced to 0.29 and 0.13 when the SNR is set to 0.5 and 0.3, respectively.

3 System Architecture

Figure 3 shows the structure of the count-based ranking system implemented in this study. It consists of dataset construction, data pre-processing, semantic search, and ranking result derivation. We created a dataset by crawling review data for tourist destination on TripAdvisor [8] and performed pre-processing. Subsequently, a semantic search derives the semantic similarity between a specific query and each

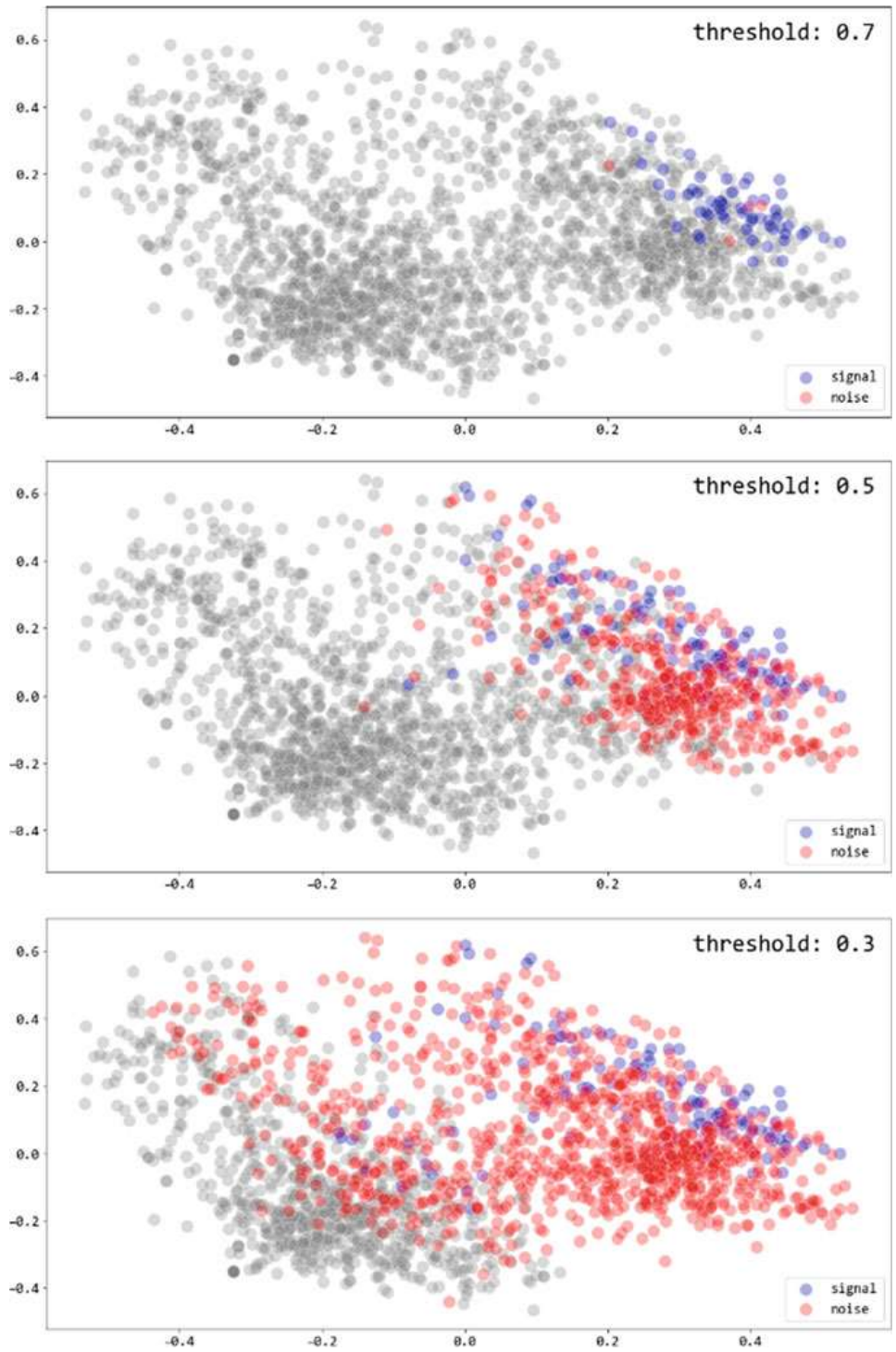

Fig. 2 The boundary at each threshold

Table 1 Numerical Information according to Threshold

Threshold	# of Signal	# of Noise	SNR	# of Signal not included
0.7	58	4	14.50	77
0.5	118	409	0.29	17
0.3	133	1010	0.13	2

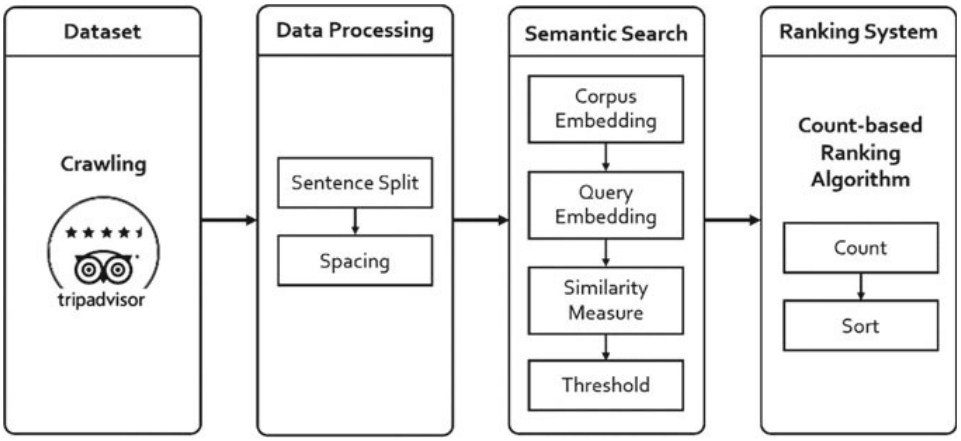

Fig. 3 Ranking system architecture using semantic search

data point in the dataset. Based on the results of the semantic search, tourist destinations with data similar to the query are ranked, making it possible to identify tourist locations suitable for a specific query.

3.1 Data

Reviews of tourist attractions in Busan, South Korea, were collected from TripAdvisor [8]. We scrapped the reviews of 201 tourist spots, including Haeundae, Gamcheon Culture Village, and Haedong Yonggungsa, and stored the data as a comma-separated value (CSV) file. The data consist of "names of tourist destination," "languages," "dates visited," "reviews," and "ratings" attributes. We used 11,751 reviews written in English between July 2007 and May 2021. One review consisted of approximately five sentences on average, and each sentence had an average consisted of approximately 15 words.

3.2 Text Cleaning and Pre-processing

Because reviews include opinions on various aspects of tourist destinations, they must be separated into individual sentences to interpret each opinion more clearly. Using the Natural Language Toolkit (NLTK) library [10] developed for English natural language processing (NLP), sentence segmentation was performed for all reviews. Then, data that had no actual meaning (were not composed of English characters and numbers) were removed. As a result, 58,175 sentences were extracted.

Some sentences had non-spaced phrases such as "activitiesscenerySpiritual" and "housemysonwanted." As they are not separated word by word, they have the potential to lead to inaccurate results. Therefore, the sentences were spaced using the "Wordsegment" library [11] from Apache2.

3.3 Embeddings

In this study, the data were embedded using SBERT. It applies Siamese and triplet networks to the existing BERT to more efficiently express embeddings that reflect the meaning of sentence data and to quickly calculate similarity through cosine similarity. This makes it a suitable embedding model for semantic searches where similarity comparison is important for a large-scale dataset.

Sentences were vectorized by a pre-trained SBERT model named "all- MiniLM-L6-v2." It had sentence embeddings and semantic search performances of 68.06 and 49.54, respectively. This model has a speed of 14,200 sentences per second, which is five times faster while retaining good quality [4].

First, a corpus of review sentences was embedded in vector form. The query of interest is represented as a vector in vector space of the review sentences.

3.4 Similarity Measure

We were able to compute the similarity between the corpus and query by embedding them. Similarity is measured using cosine similarity, and the formula can be obtained as follows:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{AB} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

We set a threshold and assumed that sentences above the threshold value were semantically related to the query. In this study, 0.6 was rather arbitrarily set as the default threshold value. However, it should be noted that an appropriate threshold value can vary depending on the data.

3.5 Ranking Algorithm

The ranking algorithm returns a list of tourist destinations suitable for a specific query. A count-based ranking algorithm considers the number of tourist spots that semantically matched the query. We counted the number of review sentences that were above the threshold value for each tourist destination. Tourist attractions were sorted based on the obtained information.

Algorithm 1 Count-based Ranking Algorithm

1. *ranked_destination* = *dictionary*{}
2. **for each** *tourist_destination* ∈ *dataset* **do**
3. *ranked_destination*[‘*tourist_destination*’] + = 1
4. **end for**
5. *sort_by_value*(*ranked_destination*)
6. **return** *ranked_destination*

4 Result of Ranking System

4.1 Result of Ranking System

A ranking system with three queries was applied.

The first is a positive query “the night view is beautiful,” and the tourist spots suitable for the query are obtained and ordered. The number of reviewed data points was 377. The ranking results of the queries are shown in Fig. 4. Gwangan Bridge [12], Gwangalli Beach [7], Haeundae Beach [13], Taejongdae [14], and Haedong Yonggungsa Temple [15] were placed in order for the query. All tourist destinations were located on the seaside. In particular, the Gwangan Bridge is known to have the best night view because different colors are produced using 7011 LED lights for each season.

The second query concerns the toilet facility, “Toilet is dirty.” A total of 23 cases were included in our study. The ranking results of the queries are shown in Fig. 5. Haeundae Beach [13], Gamcheon Culture Village [16], Haedong Yonggungsa Temple [15], Beomeosa Temple [17], and Busan subway were placed in descending order. Because millions of tourists visit Haeundae Beach annually. Tourists on the beach often complain about the maintenance and cleanliness of toilet facilities.

The third query is concerned with the crowdedness of tourist destinations “it is very crowded.” A total of 219 cases matched our queries. The ranking results of the queries are shown in Fig. 6. Haeundae Beach [13], Haedong Yonggungsa Temple [15], BIFF Square, Gamcheon Culture Village [16], and Gwangalli Beach [7] were ranked for the query. Unsurprisingly, Haeundae Beach is the most popular tourist destination in Busan, where millions of tourists visit each year.

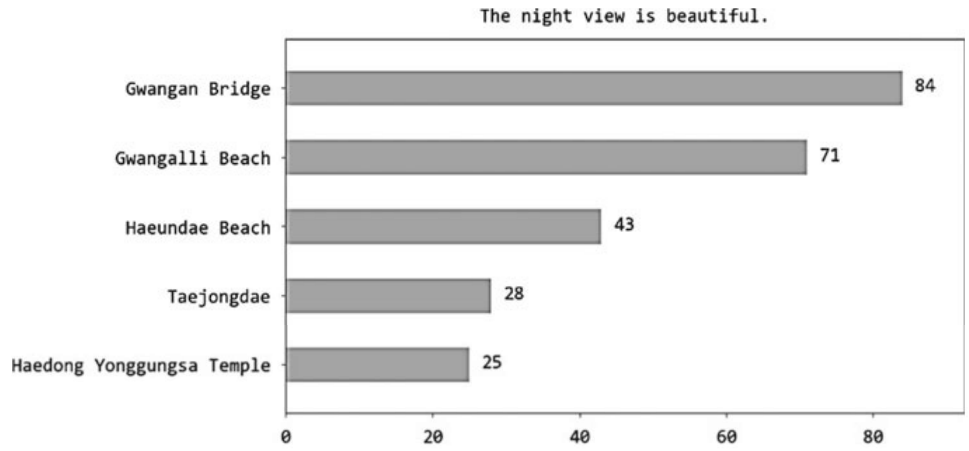

Fig. 4 Result of ranking system for "*The night view is beautiful*"

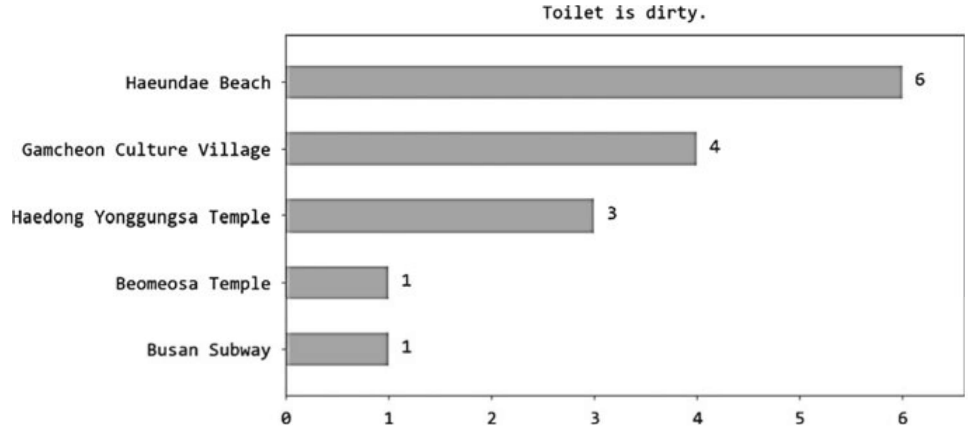

Fig. 5 Result of ranking system for "*Toilet is dirty*"

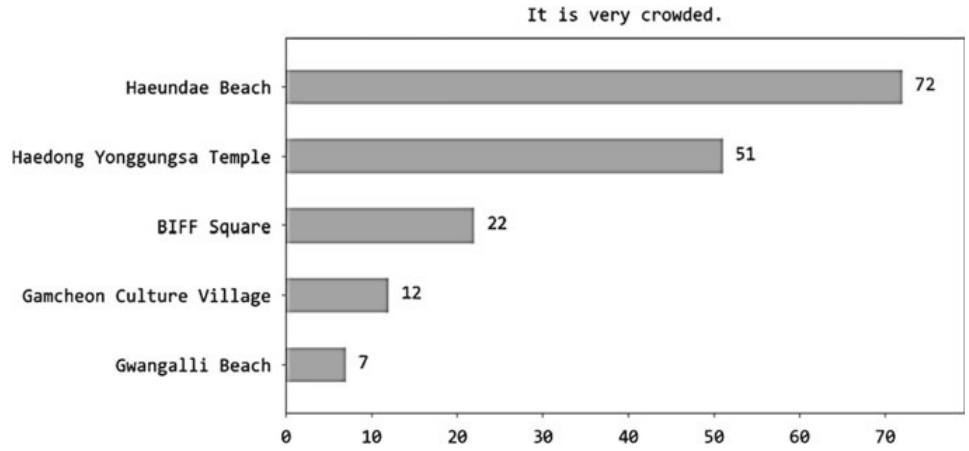

Fig. 6 Result of ranking system for "*It is very crowded*"

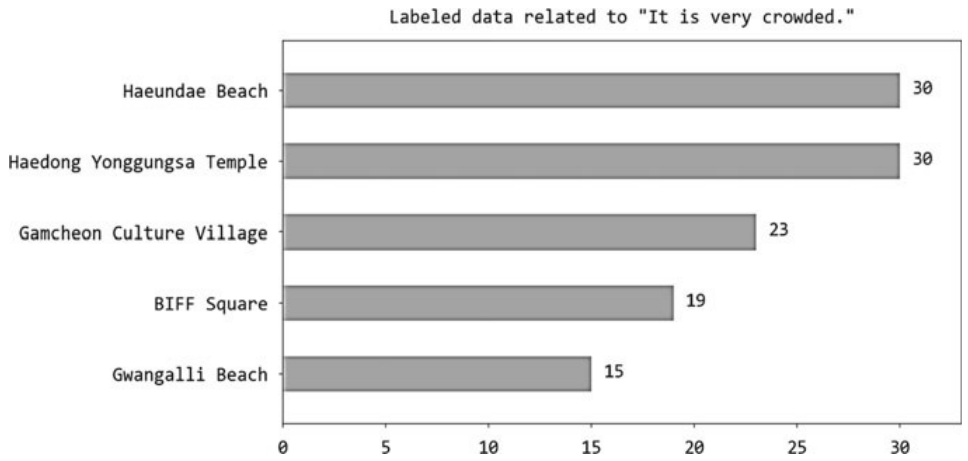

Fig. 7 Ranking for labeled data related to "*It is very crowded*"

4.2 Evaluation of Result

To evaluate the ranking algorithm, we directly labeled 58,175 sentences whether the data were semantically related to the third query on crowdedness. The tourist spots with data labeled as relevant to the query were ranked orderly in Fig. 7.

The top five tourist destinations are Haeundae Beach [13], Haedong Yonggungsa Temple [15], Gamcheon Culture Village [16], BIFF Square, and Gwangalli Beach [7]. Compared with Fig. 6, which is the result of the ranking algorithm, they yield the same top places, although in different rankings.

The labeling process guarantees accuracy but requires a lot of time and effort. The ranking algorithm produces similar results and is even extremely fast and convenient. The ranking algorithm of this study is competitive as an economical and efficient method to roughly derive tourist attractions related to a query.

5 Discussion and Conclusion

To extract meaningful information and provide insight from online tourism big data, this study ranks tourist destinations according to a specific topic using semantic search. For the three queries on night view, toilet facilities, and crowdedness, we derived a ranking of tourist destinations with high semantic relevance to the query.

This study extracts information limited to a particular subject (query), which facilitates access to the desired information. This makes it possible to intuitively rank and recommend tourist attractions according to the query. Because this recommendation is information generated based on the opinions of real tourists, not companies or governments, it is likely to be more reliable and appropriate. This might help tourists optimize their destination choices and find unexpectedly lovely places. From the point

of view of a policymaker or relevant stakeholder, tourist sites with related issues can be recognized and reflected in policy decisions by using appropriate queries, and more strategic solutions can be devised by analyzing the properties of the obtained tourist destinations. For example, budgets can be assigned deliberately based on the ranking of tourist destinations. When planning a local tour product, they can find suitable places for the theme of the tour. Practitioners and stakeholders can refer to the results of tasks, such as policy decision-making, budget allocation, and new business development, allowing them to improve their services.

Some of the shortcomings of this study include optimal thresholds and imbalanced data. The appropriate threshold value is different for each query and depends on the corpus (data). Determining the appropriate threshold is a difficult and important task for achieving good performance. As the tourist destinations that appear frequently in the ranking of the results show, there is an imbalanced data problem, in which the review data of some tourist spots dominate. Among the data used in this study, Gamcheon Culture Village, Haeundae Beach, and Haedong Yonggungsa Temple, which commonly appear in the above results, have 7977 (39.7%), 6379 (31.7%), and 5431 (27.01%) sentences, respectively, which are large percentages of the total data. The imbalanced data make the count-based algorithm biased because tourist spots with more data are likely to have more data within the boundary.

In future research, we plan to develop algorithms other than the count-based ranking algorithm to solve the unbalanced data problem. We plan to develop an algorithm based on the average ratings and an algorithm that weighs the portion, which is relatively insensitive to the amount of data.

Semantic search, compared to classification, is economical and simpler but has relatively poor accuracy. Compared to the clustering method, this method is less ambiguous. Just as precise measurements are necessary for science, it is equally important to make rough estimates of quantities using rudimentary ideas and common observations. In practical terms, a semantic search is a useful tool for obtaining specific results, although the boundary is arbitrary.

Acknowledgements This study was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A3A2075240).

References

1. Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior, 2*(27), 634–639.
2. Analyst Prep. https://analystprep.com/study-notes/cfa-level-2/quantitative-method/supervised-machine-learning-unsupervised-machine-learning-deep-learning/
3. Wikipedia Semantic search. https://en.wikipedia.org/wiki/Semantic_search
4. SBERT. https://www.sbert.net
5. Roy, S., Modak, A., Barik, D., & Goon, S. (2019). An overview of semantic search engines. *Int. J. Res. Rev, 10*(6), 73–85.

6. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *2019 Conference on empirical methods in natural language processing* (pp. 671–688). Association for Computational Linguistics, Hong Kong.
7. Wikipedia Gwangalli Beach. https://en.wikipedia.org/wiki/Gwangalli_Beach
8. Tripadvisor. https://www.tripadvisor.co.kr/
9. Wikipedia Signal-to-noise ratio. https://en.wikipedia.org/wiki/Signal-to-noise_ratio
10. Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. In *ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics* (Vol. 1, pp. 63–70).
11. Wordsegment. https://pypi.org/project/wordsegment/
12. Wikipedia Gwangan Bridge. https://en.wikipedia.org/wiki/Gwangan_Bridge
13. Wikipedia Haeundae Beach. https://en.wikipedia.org/wiki/Haeundae_Beach
14. Wikipedia Taejongdae. https://en.wikipedia.org/wiki/Taejongdae
15. Wikipedia Haedong Yonggungsa. https://en.wikipedia.org/wiki/Haedong_Yonggungsa
16. Wikipedia Gamcheon Culture Village. https://en.wikipedia.org/wiki/Gamcheon_Culture_Village
17. Wikipedia Beomeosa. https://en.wikipedia.org/wiki/Beomeosa