Journal of Korea Society of Industrial Information Systems Vol.29 No.4, Aug. 2024 :91-103
http://dx.doi.org/10.9723/jksiis.2024.29.4.091 ISSN:1229-3741

BERT 기반 의미론적 검색을 활용한 관광지 순위 시스템 개발+

(Development of a Ranking System for Tourist Destination Using BERT-based Semantic Search)

이 강 우[1], 김 명 선[2], 홍 순 구[3]*, 노 수 경[4]
(KangWoo Lee, MyeongSeon Kim, Soon Goo Hong, and SuGyeong Roh)

요 약 본 연구의 목적은 시맨틱 검색 기법을 활용하여 사용자 쿼리 기반의 타당한 정확도를 가진 관광지 랭킹시스템을 설계하는 것이다. 이를 위해 관광지에 대한 텍스트 리뷰 데이터 수집, 데이터 전처리 및 SBERT를 활용한 임베딩 과정을 거쳤다. 이후 유사도를 측정하고 임계값을 충족하는 데이터를 필터링한 후 카운트 기반 랭킹 알고리즘을 적용하여 쿼리와 의미적으로 유사한 순서로 관광지 순위를 도출하였다. 제안된 랭킹 알고리즘의 평가를 위해 4개의 쿼리로 실험을 진행하여 연관성이 높은 상위 5개 관광지를 도출하였다. 도출된 결과값의 비교를 위해 58,175개의 문장에 직접 라벨을 붙여 세 번째 쿼리인 혼잡도와 의미적으로 연관성이 있는지를 확인하였다. 두 결과값이 유사하여 본 연구에서 제시된 랭킹 알고리즘의 효율성이 검증되었다. 임계값 최적화, 데이터 불균형 등의 문제에도 불구하고 이 연구는 시맨틱 검색 기법을 이용하여 적은 비용과 시간으로도 사용자의 의도를 파악하여 관광지를 추천하는 것이 가능하다는 것을 보여주었다.

핵심주제어: BERT 기반 시맨틱 검색, 카운트 기반 랭킹 알고리즘, 관광지

Abstract A tourist destination ranking system was designed that employs a semantic search to extract information with reasonable accuracy. To this end the process involves collecting data, preprocessing text reviews of tourist spots, and embedding the corpus and queries with SBERT. We calculate the similarity between data points, filter out those below a specified threshold, and then rank the remaining tourist destinations using a count-based algorithm to align them semantically with the query. To assess the efficacy of the ranking algorithm experiments were conducted with four queries. Furthermore, 58,175 sentences were directly labeled to ascertain their semantic relevance to the third query, 'crowdedness'. Notably, human-labeled data for crowdedness showed similar results. Despite challenges including optimizing thresholds and imbalanced data, this study shows that a semantic search is a powerful method for understanding user intent and recommending tourist destinations with less time and costs.

Keywords: BERT-based semantic search, count-based ranking algorithm, tourist destination

* Corresponding Author: hongsoongoo@dtu.edu.vn
+ This research is based on the study presented at the 2022 ICSADL conference.
+ This study was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A3A2075240).
Manuscript received August 02, 2024 / revised August 18, 2024 / accepted August 21, 2024

1) Smart Governance Research Center, Dong-A University, 제1저자
2) Department of Computer Engineering, Dong-A University, 제2저자
3) International School, Duy Tan University, Vietnam, 교신저자
4) Department of Management Information Systems, Dong-A University, 제3저자

1. Introduction

The growth of the Internet has led to a significant increase in the number of online reviews on tourism and other industries (Ryu and Cho, 2020; Lee, 2020; Kim et al., 2022). Since online reviews significantly influence tourists' decision-making (Ye et al., 2011), the analysis of this data is indispensable for the tourism industry. It allows practitioners and stakeholders to consider the implications of the findings in the policy development process and identify potential policy concerns, such as accessibility concerns or a lack of amenities at tourist spots. Furthermore, such analysis may offer a practical suggestion system that will enhance the travel experience for tourists (Bauernfeind, 2003).

Online reviews serve as a rich data source for enhancing marketing strategies. By analyzing review sentiments and keywords, tourism businesses can tailor their marketing messages to better address the desires and concerns of potential tourists (Sun et al., 2024). Reviews often contain detailed feedback on various aspects of tourist experiences, such as customer service, cleanliness, and the quality of accommodations. By systematically analyzing this feedback, businesses can identify areas needing improvement (Zheng et al., 2023).

Methods for the extraction of meaningful information from large data sets have been created. Two methods in particular have gained popularity. One is classification, which automatically categorizes input objects (e.g., text or images) into predefined categories based on their content. A classification model learns patterns from labeled examples to predict the class of new, unseen input objects (Hand, 2012). However, the process of labeling tourist reviews with these categories can be subjective and time-consuming, which makes it challenging to obtain large labeled datasets (IBM, 2024). Another popular approach is clustering, which groups objects based on their similarities and characteristics (Sun et al., 2021). This technique also has limitation in that the inherent ambiguity of classification may result in less accurate outcomes (Hennig, 2015).

To address the limitations of traditional methods like classification and clustering, this research aims to design a tourist destination ranking system that uses semantic search to extract information based on user queries with reasonable accuracy. To this end, we use the semantic search method.

Semantic search is designed to understand the meaning and context of words, rather than simply matching keywords. Its goal is to provide more accurate search results by understanding the user's intent.

For this reason, this study employs semantic search to develop a tourist attraction ranking system based on a particular query by users. In general, user-generated ratings is frequently used to rank tourist attractions. Each tourist attraction is ranked based on the average rating it receives from visitors (Sun et al., 2021). Another simple approach is to rank tourist attractions according to their popularity, which can be gauged by the amount of visitor traffic, frequency of search queries, or the number of social media mentions (Parry, 2012). This approach operates on the premise that attractions with higher levels of traffic or mentions will rank higher.

In contrast, semantic search-based ranking systems offer several advantages for enhancing both the user experience and the accuracy of search results. Semantic search understands the context and meaning behind search queries rather than relying solely on keyword

matching (Shelke et al., 2023). This allows the system to deliver more relevant results that align closely with the user's intent. Thus, by providing more accurate and contextually appropriate search results, semantic search increases user satisfaction.

This paper is organized as follows. The chapter 2 reviews previous studies on semantic search. The chapter 3 explains the architecture of our ranking system based on SBERT (Sentence Bidirectional Encoder Representations from Transformer). The evaluation of the ranking system is described in the chapter 4 and contribution along with limitations is followed in chapter 5.

2. Literature on Semantic Search

2.1 Semantic Search and Sentence-BERT

The limitations of classification and clustering discussed in the previous section highlight the need for a more sophisticated way to analyzing tourist reviews. Semantic search provides a powerful solution by taking into account the meaning behind the words. Unlike traditional search techniques, semantic search engines can catch the hidden meaning of a query and find semantically similar sentences in the reviews, even if they don't use the exact keyword matching (Shelke et al., 2023). Consequently, it offers more substantial search results by identifying the search term and locating the most suitable results within the database. (Roy et al., 2019).

The use of BERT-based word embedding allows for the representation of a sentence or word as a dense vector. Vector representations assist in the sentences identification with comparable semantic content to a search query by analyzing the distance between them in vector space (Birunda and Devi, 2021). In other words, instead of relying on exact word matches, it identifies reviews that express semantically similar ideas, even if they use different words. Sentence-BERT (SBERT), a special type of the BERT, is designed for semantic search. It is an advanced training technique like processing sentences in pairs (Siamese networks) or triplets (anchor, similar, dissimilar) to learn which sentences have similar meanings, even if they don't use the same words. A Siamese network consists of two or more identical networks with shared parameters, which process two input sentences simultaneously and then compare their outputs. Triplet networks expand on this idea by evaluating three sentences simultaneously: an anchor, a positive example, and a negative example. These concepts enable SBERT to find relevant tourist reviews that match travelers' preferences, even if they expressed differently (Reimers et al., 2019).

In summary, SBERT was chosen as the fundamental model for this study due to its superior ability to understand the context and meaning of words within sentences in tourism industry.

2.2 Previous Studies on BERT in Tourism Industry

BERT has been employed in tourism-related researches, in particular for analyzing online review data and improving recommendation systems. Several studies have demonstrated the effectiveness of BERT in understanding complex, contextual tourist review.

Firstly, BERT was utilized in tourist sentiment analysis. For example, Zhang et al. (2023) used BERT to analyze Airbnb review data. It identifies unique tourism experiences that create value, highlighting four dimensions

of uniqueness within the service-dominant logic framework.

Secondly, BERT was also used for development of recommendation Systems. Zhuang and Kim (2021) refines the BERT model using TripAdvisor review data to predict ratings for six criteria (e.g, value, service) and proposes a multi-criteria recommender system. They stated that their system outperforms traditional single-criteria recommender systems and provides more effective recommendations to potential hotel customers.

These studies have successfully shown the potential of BERT in tourism-related applications. However, our research differs in that it creates a count-based ranking system that ranks tourist spots based on how closely they match user queries with reviews. This provides users with a relevant ranking of tourist attractions with less time and effort.

2.3 Setting a Boundary with Various Thresholds

It is necessary to apply a range of threshold values to create boundaries within the vector space. A high threshold may exclude relevant reviews (signals) while reducing the likelihood of including irrelevant reviews (noise). On the other hand, a low threshold can capture most relevant reviews, but it may also include many irrelevant ones. The threshold value is a critical component in the filtering out of noise and the enhancement of signal clarity.

To show how boundary changes occur in accordance with threshold values, a semantic search algorithm is applied to the reviews of 'Gwangalli Beach'. Fig. 1 demonstrates the scope of boundaries evaluated for the following query, "It is a beautiful beach" at three distinct values: 0.7, 0.5, and 0.3. Gwangalli Beach data was embedded in the SBERT. Total 1,943 TripAdvisor review data was examined and classified based on their relevance to the search query.

Typically, the review embedded in the SBERT are represented in a form of high-dimensional vector (768 dimensions). It is not easy to present the results in the high-dimensional vector space. Thus, principal component analysis (PCA) is applied to reduce the dimensions of high-dimensional vectors into the two dimensions. After performing PCA, the reviews were plotted in a 2D space.

Fig. 1 illustrates the distribution of queries within the defined boundary. The blue line indicates the matched target, referred to as the signal, while the red one represents the unmatched target, categorized as noise.

For a more quantitative evaluation, the signal-to-noise ratio (SNR) is measured. SNR compares the number of relevant reviews ("signal") detected by our system to the number of irrelevant reviews ("noise") included in the results. As shown in the Table 1, the SNR reached its peak value of 14.50 when the threshold was set to 0.7. It is noteworthy that the SNR drops significantly to 0.29 and 0.13 when the threshold value is 0.5 and 0.3, respectively.

Table 1 Results with 3 Different Threshold Vales

Threshold Value	Number of Signal	Number of Noise	SNR	Number of Signal Excluded
0.7	58	4	14.50	77
0.5	118	409	0.29	17
0.3	133	1010	0.13	2

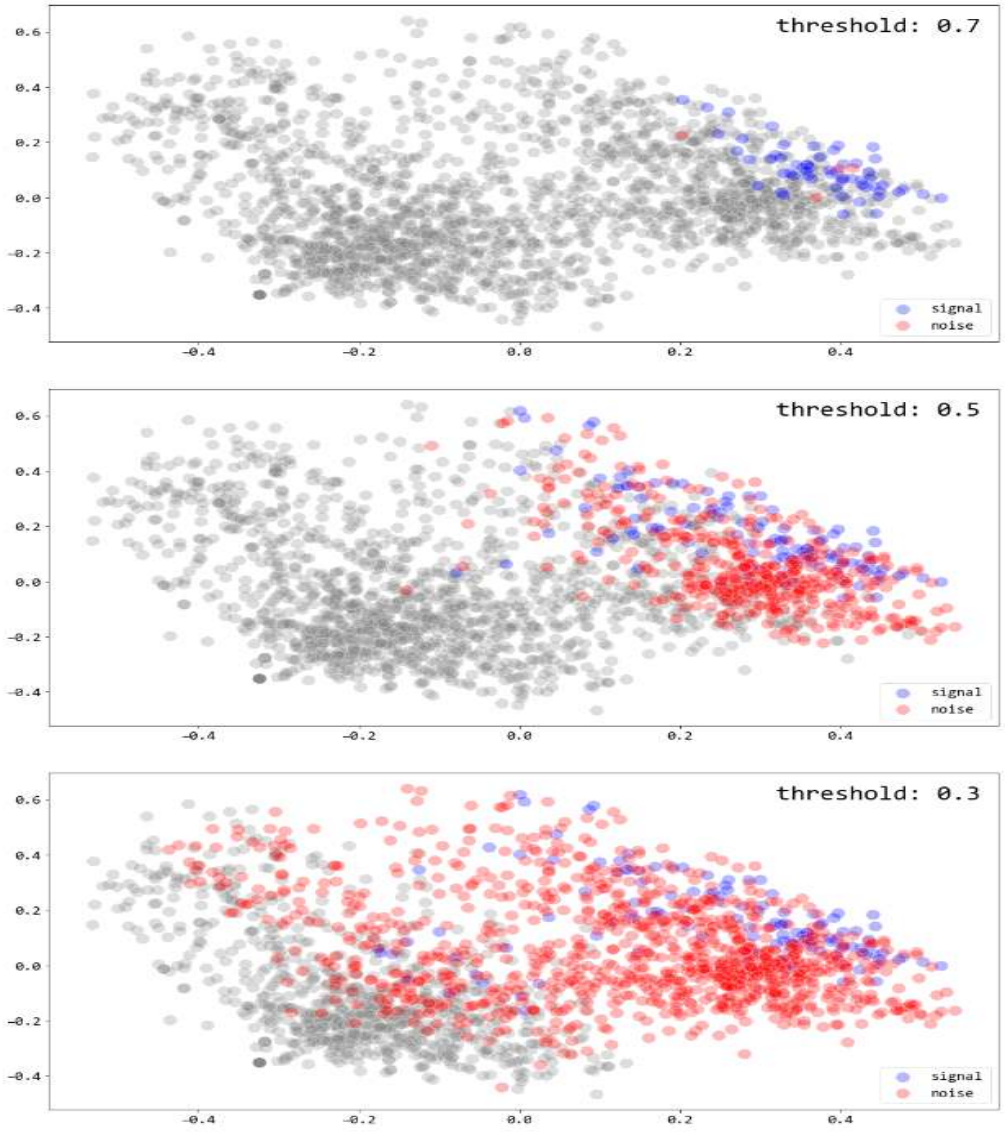

Fig. 1 The Boundary at 3 Threshold Value

3. Architecture of a Ranking System

Fig. 2 shows the process for developing a ranking system suggested in this study. The process includes the creation of datasets, preprocessing, semantic search, and the derivation of ranking results. The initial step involved the review of data for tourist destinations on TripAdvisor (TripAdvisor, 2022), which was then crawled and preprocessed to create a dataset. Subsequently, a semantic search is conducted to evaluate the degree of meaning-based similarity between the specified query and each point within the dataset. The semantic search results rank tourist destinations with data based on the user query, enabling users to identify suitable tourist destinations for their specific needs.

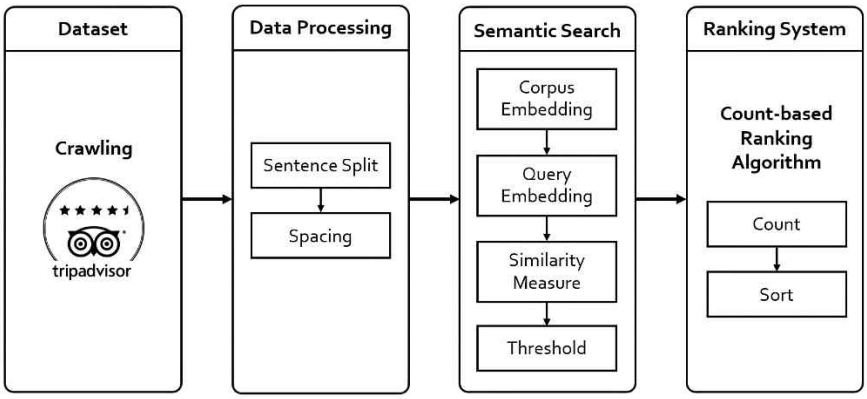

Fig. 2 Architecture of a Ranking System with Semantic Search

3.1 Data Collection

A total of 201 tourist spots including Haeundae, Gamcheon Culture Village, etc. in Busan, Korea, were collected from the TripAdvisor reviews. The data was subsequently stored in a Comma-Separated Values file format, which includes the following attributes: the names of tourist destinations, the languages spoken in those destinations, the dates on which the destinations were visited, reviews of the destinations, and the ratings assigned to them. The present study analyzed 11,751 English reviews from July 2007 to May 2021. Each review data has 5 sentences, with 15 words in average per sentence.

3.2 Data Preprocessing

Tourist reviews typically encompass a range of perspectives on different facets of tourist attractions, providing valuable insight for potential visitors. The reviews include when and how they went to a tourist attraction, what they saw and did, and what was good and bad. Thus, it is recommended that they be segmented into individual sentences to facilitate more accurate interpretation of each opinion. NLTK, a widely used library for English NLP tasks (Loper and Bird, 2002), is utilized for segmentation. After filtering out sentences containing non-English characters (e.g., symbols), a set of 58,175 sentences was obtained for a further analysis.

Some sentences included non-spaced phrases that were not separated by punctuation, such as "activitiesscenerыspiritu-al" and "housemysonwanted." To prevent any potential inaccuracies, these sentences were separated into individual words using the Apache2 'Wordsegment' library (Grant, 2018).

3.3 Word Embedding

The data were analyzed using SBERT. The BERT model has been enhanced by the application of Siamese and triplet networks, enabling more efficient expression of sentence embedding and rapid similarity calculations via cosine similarity measure. The use of Siamese and triplet networks creates a powerful embedding model, particularly suited for large-scale semantic search applications where rapid and precise similarity comparisons are crucial.

Subsequently, the sentences were vectorized using the pre-trained SBERT model, known as "all-MiniLM-L6-v2 (Hugging Face, 2022)"

The corpus was transformed into a vector format, and the query was then mapped into the resulting vector space. Finally, the query is subsequently embedded within the same vector space.

3.4 Measurement of Similarity

The similarity between the sentences and the query is measured with the process of embedding. The degree of similarity between the sentence and the query is calculated using cosine similarity, with the following formula:

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

Thresholds have been set, and sentences surpassing these cutoff points are assumed to be contextually aligned with the user query. For this research, the default threshold value was arbitrarily set to 0.6.

3.5 Count-based Ranking Algorithm

To generate tourist attractions that align with the specified query, the count-based ranking algorithm is employed. For each tourist destination, the number of sentences with a similarity score above the specified threshold was recorded. The tourist attractions were then classified according to this information.

Ranking Algorithm

```
ranked_destination = dictionary{}
for each tourist_destination ∈ dataset do
    ranked_destination['tourist_destination'] += 1
end for
sort_by_value(ranked_destination)
return ranked_destination
```

4. Evaluation of the Suggested System

4.1 Experiments with Four Queries

To demonstrate the performance of the count-based ranking algorithm for various search intentions, the system was tested using four different queries. The results are presented in Fig. 3, 4, 5 and 6. It is noted that the top 5 tourist attractions corresponding to a query are presented.

"The night view is beautiful," the first query, was employed to analyze 377 data points for identifying and prioritizing tourist destinations. Shown in Fig. 3, Gwangan Bridge has the most numbers which is 84. It is particularly famous for its night view, which is enhanced by 7,011 LED lights that change colors with each season.

Secondly, the query "Toilet is dirty" was related to the condition of toilet facilities and was addressed on 23 occasions. As shown in Fig. 4, the findings indicated that Haeundae Beach was ranked first. Due to the high volume of tourists visiting Haeundae Beach on an annual basis, complaints regarding the maintenance of its toilet facilities are frequent.

The third query, "It is very crowded," is focused on the crowdedness of tourist destinations with a total of 219 cases. As illustrated in Fig. 5, Haeundae Beach was identified as the most highly ranked destination,

followed by Haedong Yonggungsa Temple. Indeed, Haeundae Beach is the most frequently visited location in Busan.

The final query, "The food is delicious" is focused on the food taste of tourist destinations with a total of 231 cases. The results show that Jagalchi Market, with the highest number of mentions, stands out as the top destination for delicious food, followed by BIFF Square and other markets and cultural spots known for their unique and tasty offerings. The rankings reflect the importance of diverse and high-quality food experiences in attracting tourists to these destinations.

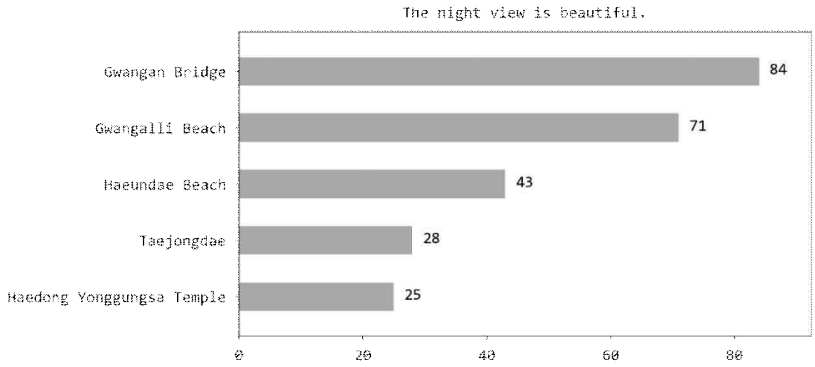

Fig. 3 Query result: "The night view is beautiful."

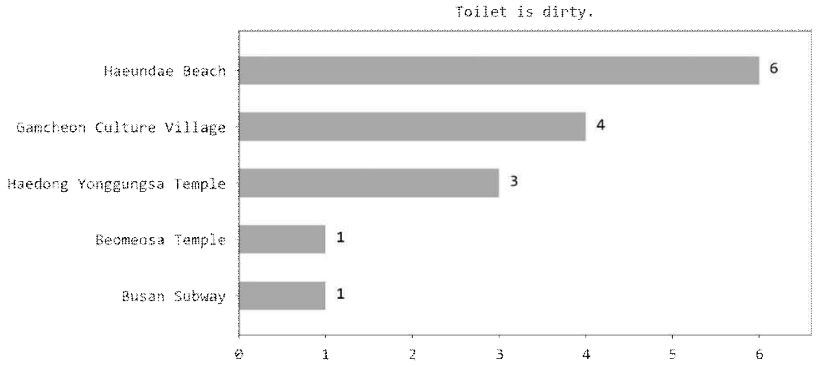

Fig. 4 Query result: "Toilet is dirty."

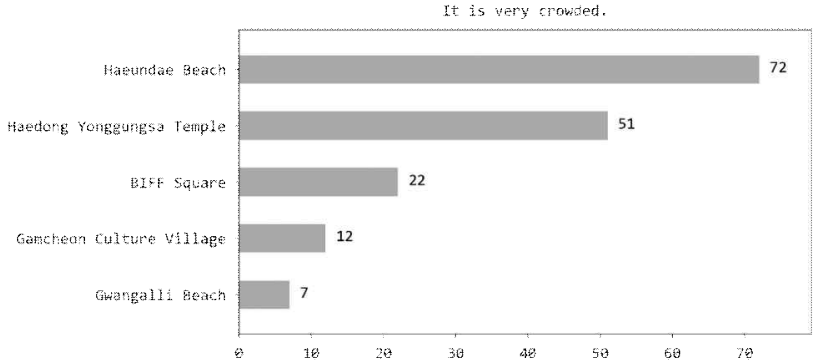

Fig. 5 Query result: "It is very crowded."

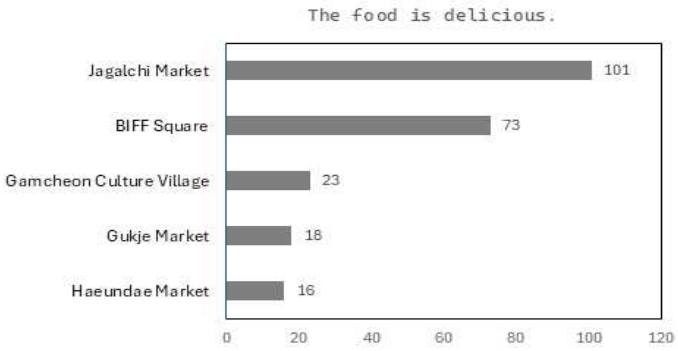

Fig. 6 Query result: "The food is delicious."

4.2 Evaluation

To assess the effectiveness of the ranking algorithm, a human reviewer manually labeled 58,175 sentences for semantic relevance to the "crowdedness" query.

As illustrated in Fig. 5, our ranking system revealed that the top five locations (e.g., Haeundae Beach) were consistently ranked as the top-five spots, but in a different order.

The process of labeling is an effective means of ensuring accuracy, although it is a time-consuming and labor-intensive approach. In contrast, the ranking algorithm yields similar results and is notably quick and convenient. The ranking system developed in this research may be a competitive method for approximately identifying tourist destinations in an economical and efficient manner.

5. Conclusions

5.1 Summary

As online reviews become so popular, analysis of this data is essential for understanding tourist preferences. To overcome the limitations of traditional classification and clustering methods, this study employs semantic search with SBERT to enhance the relevance and efficiency of identifying tourist destinations. By applying different thresholds for semantic similarity, this study shows how these thresholds impact the signal-to-noise ratio in search results. The ranking system, which organizes tourist destinations based on semantic matches to queries, demonstrates itself as a cost-effective and efficient tool for identifying relevant tourist spots.

5.2 Implications and Contributions

The implications of this research are as follows. Classification relies on accuracy and clustering focuses on intra-cluster coherence and inter-cluster separation (Pfitzner et al., 2009; Powers, 2011), whereas our semantic search based on a ranking system evaluates relevance to a query (Valcarce et al., 2020). In this work, we aimed to develop a flexible, practical, and low-cost system at the expense of a certain degree of accuracy. Not like precise measurement in the scientific research, our approach is focused on developing a low-cost and practical tool with moderate accuracy. Achieving higher accuracy requires more data, more high-performance hardware, and more complex algorithms. This is a

time-consuming and costly task. On the other hand, aiming for moderate accuracy reduces the resources needed, thereby reducing costs.

Our research has shown that semantic search is effective in ranking tourist destinations, making a significant insights to the academician. In fact, it enables users to state their preferences in different words through their queries, while the system can still identify relevant destinations based on the underlying meaning.

Secondly, recommendations directly from actual travelers tend to provide a more accurate and relevant perspective than those provided by businesses or official bodies. These insights assist tourists in making more informed destination choices and in discovering unexpected and appealing locations. The results of this research provide valuable practical implications for policymakers and tourism businesses, such as policy development, budget allocation, and new tour services.

For policy development, the ranking system enables policymakers to make evidence-based decisions on tourist preferences and new trends. By analyzing which tourist spots are highly preferred for specific attributes, policymakers can develop strategic or long-term plans to address areas for improvements.

For budget allocation, the system's ability to rank destinations enables more data-driven and efficient budget allocation. For example, the most visited spots can be prioritized for funding and infrastructure.

Tourism related organizations can use our ranking system to guide the development of new tour services by identifying which destinations excel in specific attributes (e.g., accessibility). This makes sure that new tour services build on current strengths and weaknesses.

5.3 Limitations and Future Research

This study has several limitations, including determining an optimal threshold value and addressing imbalanced data.

Firstly, the optimal threshold is dependent upon the specific query and dataset, which presents a challenge in achieving consistently good results. In our algorithm, the threshold was set arbitrarily, which is a limitation. In further work, our research will extend to applying more effective thresholding methods using Histogram or Kernel Density Estimation (KDE) (Xu et al., 2015) or Receiver Operating Characteristic Curve (Carter et al., 2016).

Secondly, there is an imbalance in the dataset, in which certain tourist destinations dominate the review data. For instance, Gamcheon Culture Village, Haeundae Beach, and Haedong Yonggungsa Temple frequently are presented in the ranking results, constituting large proportions of the total data－39.7%, 31.7%, and 27.01% respectively. The presence of imbalanced data in the data set introduces a bias into the count-based algorithm, resulting in an overrepresentation of tourist locations with a greater quantity of data. Future research will concentrate on the creation of alternative algorithms with the objective of reducing the impact of imbalanced data. To achieve this, we will investigate the potential of average rating-based and weighted portioning approaches.

The third limitation relates to the validation of our algorithm. Out of four queries, one (crowdness) was randomly selected and the researcher performed labeling on 58,175 sentences. However, there is insufficient theoretical basis to determine whether this number is adequate. Further research is therefore needed to address this issue.

Finally, future research could explore the

application of our proposed ranking system in different domains. In other words, research could investigate how semantic search and a ranking algorithm perform in other industries, such as hospitality or retail, to evaluate its generalizability and effectiveness.

References

Analyst Prep. (2022). https://analystprep.com/study-notes/cfa-level-2/quantitative-method/supervised-machine-learning-unsupervised-machine-learning-deep-learning/ (Accessed on Nov. 1th, 2022)

Carter, J. V., Pan, J., Rai, S. N. and Galandiuk, S. (2016). ROC-ing along: Evaluation and Interpretation of Receiver Operating Characteristic Curves, *Surgery*, 159(6), 1638-1645.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011). *Cluster Analysis*, Chichester, Wiley.

Grant, J. (2018). wordsegment 1.3.1 https://pypi.org/project/wordsegment/ (Accessed on Jul. 30th, 2022)

Hand, D. (2012). Assessing the Performance of Classification Methods, *International Statistical Review*, 80(3), 400-414.

Hugging Face. (2024). all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (Accessed on Jul. 30th, 2022)

IBM. (2024). What is Data Labeling? https://www.ibm.com/topics/data-labeling (Accessed on Jul. 30th, 2024)

Kim, S. H., Kim, M. G. and Ryu, M. H. (2022). Importance-Performance Analysis for Korea Mobile Banking Applications: Using Google Playstore Review Data, *Journal of Korea Society of Industrial Information Systems*, 27(6), 115-126.

Lee, T. W. (2020). A Study on Analysis of Topic Modeling using Customer Reviews based on Sharing Economy: Focusing on Sharing Parking, *Journal of Korea Society of Industrial Information Systems*, 25(3), 39-51.

Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit, *In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 63-70.

Parry, F. (2012). Website Visibility: The Theory and Practice of Improving Rankings, *Aslib Proceedings*, 64(2), 215-215.

Pfitzner, D., Leibbrandt, R. and Powers, D. (2009). Characterization and Evaluation of Similarity Measures for Pairs of Clusterings, *Knowledge and Information Systems*, 19(3), 361-394. doi:10.1007/s10115-008-0150-6. S2CID 6935380.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, *Journal of Machine Learning Technologies*, 2(1), 37-63.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982-3992.

Roy, S., Modak, A., Barik, D. and Goon, S. (2019). An Overview of Semantic Search Engines, *International Journal of Research & Review*, 6(10), 73-85.

Ryu, M. H. and Cho, H. S. (2020). An Analysis of IoT Service using Sentiment Analysis on Online Reviews: Focusing on the Characteristics of Service Providers, *Journal of Korea Society of Industrial Information Systems*, 25(5), 91-102.

SBERT. (2022). https://www.sbert.net (Accessed

on Nov. 1th, 2022)

Shelke, P., Shewale, C., Mirajkar, R., Dedgoankar, S., Wawage, P. and Pawar, R. (2023). A Systematic and Comparative Analysis of Semantic Search Algorithms, *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(11s), 222-229.

Sun, H. L., Liang, K. P., Liao, H. and Chen, D. B. (2021). Evaluating User Reputation of Online Rating Systems by Rating Statistical Patterns, *Knowledge-Based Systems*, 219.

Sun, X., Wang, Z., Zhou, M., Wang, T. and Li, H. (2024). Segmenting Tourists' Motivations via Online Reviews: An Exploration of the Service Strategies for Enhancing Tourist Satisfaction, *Heliyon*, 10(1).

TripAdvisor. (2024). BUSAN Reviews. In TripAdvisor. from https://www.tripadvisor.com/Tourism-g297884-Busan-Vacations.html (Accessed on Jul. 30th, 2024)

Valcarce, D., Bellogín, A., Parapar, J. and Castells, P. (2020). Assessing Ranking Metrics in Top-N Recommendation, *Information Retrieval Journal*, 23(4), 411-448.

Wikipedia Semantic Search. (2022). https://en.wikipedia.org/wiki/Semantic_search (Accessed on Nov. 1th, 2022)

Wikipedia Signal-to-noise Ratio. (2022). https://en.wikipedia.org/wiki/Signal-to-noise_ratio (Accessed on Nov. 1th, 2022)

Xu, X., Yan, Z. and Xu, S. (2015). Estimating Wind Speed Probability Distribution by Diffusion-based Kernel Density Method, *Electric Power Systems Research*, 121, 28 - 37.

Ye, Q., Law, R., Gu, B. and 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*, 27(2), 634-639.

Zhang, H., Liu, R. and Egger, R. (2023). Unlocking Uniqueness: Analyzing Online Reviews of Airbnb Experiences Using BERT-based Models, *Journal of Travel Research*, 63(7), https://doi.org/10.1177/00472875231197381.

Zheng, X., Huang, J., Wu, J., Sun, S. and Wang, S. (2023). Emerging Trends in Online Reviews Research in Hospitality and Tourism: A Scientometric Update (2000–2020), *Tourism Management Perspectives*, 47.

Zhuang, Y. and Kim, J. K. (2021). A BERT-Based Multi-Criteria Recommender System for Hotel Promotion Management, *Sustainability*, 13(14), 8039.

이 강 우 (KangWoo Lee)
- 정회원
- 부산대학교 심리학과 문학사
- Birmingham university 인지과학과, MSc
- Sussex university, 인공지능 전공, DPhil
- 동아대학교 스마트거버넌스연구센터 연구원(전)
- (현재) ㈜ SmartGo 공동대표
- 관심분야: 기계학습, 텍스트마이닝, 빅데이터

김 명 선 (MyeongSeon Kim)
- 동아대학교 컴퓨터공학과 공학사
- 관심분야: 텍스트마이닝, 사용자인터페이스

홍 순 구 (Soon Goo Hong)
- 종신회원
- 영남대학교 경영학과 경영학사
- University of Nebraska-Lincoln, 경영학석사
- University of Nebraska-Lincoln, 경영정보학박사
- Texas A&M International University 조교수(전)
- 동아대학교 경영대학 교수(전)
- (현재) Duy Tan University, International School 교수
- 관심분야: 빅데이터, 관광정보시스템, Smart Goverance

노 수 경 (SuGyeong Roh)
- 신라대학교 경영정보학과 경영학사
- 동아대학교 경영정보학과 경영학석사
- 동아대학교 경영정보학과 박사수료
- (현재) 동아대학교 경영정보학과 박사수료생
- 관심분야: 네트워크 분석, 토픽모델링, 물류정보시스템