Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

A Content-Based Filtering Approach Using TF-IDF and Cosine Similarity for Hotel Recommendation Based on Traveloka Accommodation Data: A Case Study of Jakarta

Abdul Latif[1], Siti Khotimatul Wildah[2*], Sarifah Agustiani[3], Ego Oktafanda[4]

[1]*Information Systems, Universitas Bina Sarana Informatika, Indonesia*
[2]*Computer Technology, Universitas Bina Sarana Informatika, Indonesia*
[3]*Informatics, Universitas Bina Sarana Informatika, Indonesia*
[4]*Computer Science, Universitas Rokania, Indonesia*
abdul.bll@bsi.ac.id[1], siti.ska@bsi.ac.id[2]*, sarifah.sgu@bsi.ac.id[3], ego.oktafanda@rokania.ac.id[4]

Abstract

The development of Online Travel Agents (OTA) has generated a large and diverse volume of accommodation data, which often makes it difficult for users to select hotels that match their preferences in terms of location, facilities, price, and service reputation. This study is a continuation of a previous work on Traveloka accommodation data acquisition using web scraping based on the data-testid attribute. The focus of this research is to utilize the scraped data for developing a machine learning-based hotel recommendation system using a content-based filtering approach. The dataset used consists of 1,809 hotel records in the Jakarta area with attributes including hotel name, property type, star rating, rating score, location, price, facilities, and image URL. Data preprocessing includes price cleaning, separating rating scores and number of reviews, and combining location, facilities, and property type as textual content representation. The recommendation model is built using Term Frequency–Inverse Document Frequency (TF-IDF) to construct text feature vectors, followed by Cosine Similarity to measure similarity between hotels. In addition, this study introduces a weighted popularity score that combines rating values and the number of reviews to ensure that recommendations are not only based on content similarity but also reflect the credibility of hotel popularity. Experimental results produce a TF-IDF matrix of size 1,364 × 371 and a similarity matrix of 1,364 × 1,364. Functional testing shows that the system is capable of generating ten relevant hotel recommendations based on similarity in location, facilities, and property type, which are then ranked according to the popularity score.

Keywords: *Content-Based Filtering, Cosine Similarity, Hotel Recommendation System, Popularity Score, TF-IDF,*

1. Introduction

The development of Online Travel Agents (OTA) has transformed the marketing and distribution strategies of hotel services from conventional approaches to digital channels. The utilization of OTA platforms can increase hotel visibility, simplify the booking process, and support improvements in hotel occupancy rates [1]. In this research context, Traveloka is also considered a source of accommodation data as it provides hotel information such as hotel name, location, rating, number of reviews, facilities, property type, price, and images. In previous research, such data was successfully collected automatically using a web scraping method based on the data-testid attribute, which is more stable compared to class-based selectors on dynamic websites. This previous study produced 1,809 valid hotel records in the Jakarta area that are structured and ready for further analysis [2].

Although hotel data has been successfully collected, its practical value is not yet optimal when stored only in tabular form. The main problem in OTA platforms is the large number of hotel options that users must compare. This condition leads to information overload, where users are exposed to excessive information, making it difficult to find the most relevant accommodation options efficiently. This condition highlights the necessity of information filtering mechanisms that can reduce cognitive overload and support efficient decision-making in large-scale accommodation search environments. This issue aligns with findings that in the big data era, users increasingly struggle to find relevant information from massive amounts of data on online platforms [3].

Recommendation systems are a relevant approach to address this issue because they act as information filtering systems that provide item suggestions to users. In digital platforms, recommendation systems have become an important part of human-centered artificial intelligence as they support decision-making processes. Recommendation systems have also been widely applied in various web services and play a role in facilitating human decision-making [4]. This implies that scraped accommodation data holds potential beyond descriptive analysis and can be transformed into structured representations for recommendation modeling.

In the tourism and hospitality domain, recommendation systems face more complex challenges compared to general product recommendations. Accommodation choices are not determined by a single attribute but by a combination of factors such as location, price, facilities, property type, rating, and number of reviews. Conventional tourism recommendation systems often rank items based on popularity, reputation, and category similarity; however, such approaches may not adequately represent the actual experience users may obtain from a destination or item [5]. In the hotel context, this means that popular hotels are not necessarily the most suitable for users if their attributes do not match user preferences.

The dominance of item-level attributes over user interaction data makes content-based filtering a more appropriate approach for modeling similarity among accommodation entities. The scraped Traveloka data contains hotel content information such as location, facilities, property type, rating, and price, but does not include click history, transactions, or user preferences. In tourism recommendation systems, recommendations can start from content-based approaches and later be extended with popularity, demographic information, or collaborative filtering when user data becomes more mature [6]. Thus, content-based filtering is a logical approach for early-stage research because it can generate recommendations based on similarities among hotels without requiring user history data.

In addition to content attributes, review and rating data are also important as they reflect previous user experiences. Online hotel platforms provide direct information regarding customer preferences, and user reviews can be used to extract service aspects relevant to hotel customers [7]. This strengthens the argument that rating and number of reviews should not be ignored in hotel recommendation systems, as they serve as indicators of hotel quality trustworthiness.

Recent research in hotel recommendation also shows that artificial intelligence and natural language processing (NLP) are increasingly used to understand user preferences from review text. AI-based algorithms in hotel recommendation systems are becoming more popular, particularly NLP models that extract semantic knowledge from user reviews [8]. Although this study does not use full review text as the main input, hotel attributes such as facilities, location, and property type can still be treated as textual representations that can be processed using term-weighting techniques.

To construct numerical representations of hotel attributes, this study uses Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF is used to assign weights to words or terms that represent hotel characteristics. After that, inter-hotel similarity is calculated using Cosine Similarity. TF-IDF can determine term relevance in documents, while Cosine Similarity measures similarity between contents [9]. In this study, each hotel is treated as a document composed of location, facilities, and property type attributes. Hotels with similar attribute combinations obtain higher similarity scores and are recommended to users.

However, content similarity-based recommendation alone is not sufficient. Two hotels may have similar facilities and locations but differ in reliability due to differences in ratings and number of reviews. Tourism recommendation systems can integrate popularity and seasonal demand indicators to enhance the recommendation process [10]. Furthermore, tourism recommendation systems should consider diverse data types because tourism recommendations may include accommodation, transportation, and attraction contexts with varying user needs [11]. Therefore, this study introduces a weighted popularity score that considers both rating and number of reviews so that recommendations are not only based on content similarity but also reflect user-generated trust signals.

Based on the above discussion, this study is a continuation of previous research focusing on hotel data acquisition from the Traveloka platform using a data-testid-based web scraping method [2]. The previous study produced a structured dataset suitable for further analysis but did not yet utilize it for machine learning-based recommendation modeling. Therefore, this study reuses the dataset to build a hotel recommendation system based on content-based filtering using TF-IDF, Cosine Similarity, and a weighted popularity score. The main contribution of this study is transforming scraped data into a recommendation system capable of generating Top-N hotel recommendations based on attribute similarity and user review credibility.

2. Literature Review

2.1. Online Travel Agents and Accommodation Data

Online Travel Agents (OTA) can be utilized as a data source in tourism studies because they provide accommodation information relevant for tourism marketing analysis purposes [12]. In the hospitality industry, OTA platforms also serve as digital channels that support hotel marketing activities and online room reservations [1]. In previous research, the Traveloka platform was used as a data source for accommodation extraction because it contains hotel information such as hotel name, location, rating score, number of reviews, price structure, property type, facilities, and image URL. The study applied Selenium for browser automation, Beautiful Soup for HTML parsing, and the data-testid attribute as the primary selector in the hotel data extraction process. The extraction process resulted in 1,809 valid hotel records in the Jakarta area, which were cleaned, structured into a DataFrame, and stored in CSV format for further analysis [2].

In this study, OTA data is positioned as an item data source for the hotel recommendation system. Each hotel is represented as an item with descriptive attributes. Location, facilities, and property type are used as content features because these attributes are available in the extracted dataset and represent the main characteristics of each hotel [2].

2.2. Content-Based Filtering

Content-based filtering is a recommendation system approach that uses item characteristics or attributes as the basis for generating recommendations to users [13]. In this study, the input hotel is compared with other hotels based on a combination of location, facilities, and property type attributes. These attributes are used because they are available in the Traveloka hotel dataset extracted in previous research and represent the descriptive characteristics of each hotel [2]. Thus, hotels with similar location, facilities, or property type can be considered as alternative recommendations with similar characteristics.

Content-based filtering has been widely applied in hotel and tourism recommendation studies. In the hotel domain, research in Palangka Raya applied content-based filtering to build a hotel recommendation system based on hotel attributes [14]. Another study on hotels in Yogyakarta also applied a content-based filtering approach to generate hotel recommendations based on available hotel descriptions or information [15]. In the tourism domain, a study on tourist attractions in Aceh Tamiang applied content-based filtering to generate recommendations based on attraction characteristics [16]. Based on these studies, item attribute representation can be used as a foundation for developing content-based hotel recommendation systems.

2.3. TF-IDF and Cosine Similarity

Term Frequency–Inverse Document Frequency (TF-IDF) is a text weighting method that calculates the importance of a term based on its frequency in a document and its rarity across the entire corpus. In content-based recommendation systems, TF-IDF can be used to transform textual item attributes into numerical representations, enabling computational comparison between items [9]. In this study, each hotel is treated as a document composed of location, facilities, and property type attributes. These attributes are processed using the TfidfVectorizer to generate a TF-IDF feature matrix.

In general, TF-IDF weighting is calculated as the product of Term Frequency (TF) and Inverse Document Frequency (IDF) [9].

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \tag{1}$$

Since this study uses default parameters in TfidfVectorizer, the IDF component follows the smooth IDF formulation implemented in Scikit-learn, as follows [17].

$$IDF(t) = log\left((1+n)/\left(1+df(t)\right)\right) + 1 \tag{2}$$

where t is a term, d is a hotel document, n is the total number of documents in the corpus, and $df(t)$ is the number of documents containing term t. The addition of 1 in both numerator and denominator prevents division by zero during IDF computation.

After each hotel is represented as a TF-IDF vector, the next step is to compute inter-hotel similarity using Cosine Similarity. Cosine Similarity is a similarity measurement method that compares two vectors based on the angle between them. In content-based recommendation systems, Cosine Similarity is used to measure similarity between items based on their vector representations [9].

$$\cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} \tag{3}$$

where A and B are TF-IDF vectors, $A \cdot B$ is the dot product of the two vectors, and ||A|| and ||B|| represent their magnitudes. In Scikit-learn implementation, Cosine Similarity is defined as the normalized dot product between two vectors [18]. Since TF-IDF vectors are non-negative, similarity values range from 0 to 1, where higher values indicate stronger content similarity between hotels.

2.4. Weighted Popularity Score

Hotel ratings alone are not sufficient when used directly for ranking. A hotel with a high rating but a low number of reviews may not be more reliable than a hotel with a slightly lower rating but a significantly higher number of reviews. To reduce this bias, this study adopts the concept of a prior-weighted rating system, which balances item ratings with a prior or global average. In this approach, the system defines a prior quality estimate and updates it based on user ratings received over time [19].

In addition to weighted rating, recommendation systems can also utilize popularity as an additional signal. Alhadlaq et al. proposed a similarity-popularity-based recommendation approach, showing that both similarity and popularity can be used in recommendation prediction [20]. Therefore, this study does not only rely on inter-hotel content similarity but also incorporates a weighted popularity score to reflect signals from ratings and review counts.

In this study, the prior is represented by the average rating across all hotels in the dataset, while the confidence level is represented by a minimum review threshold. Thus, hotels with a higher number of reviews are more influenced by their actual ratings, whereas hotels with fewer reviews are pulled closer to the global average rating. The weighted popularity score used in this study is defined as follows:

$$WR = \left(\frac{v}{v+m}\right)R + \left(\frac{m}{v+m}\right)C \tag{4}$$

where WR is the weighted popularity score, R is the hotel rating score, v is the number of reviews, C is the global average rating, and m is the minimum review threshold defined using the 75th percentile. The use of the 75th percentile is an operational decision to ensure that higher weight is given to hotels with a relatively large number of reviews compared to most hotels in the dataset.

2.5. Recommendation Systems

A recommendation system is an information filtering approach designed to help users find relevant items from a large set of alternatives. In the literature, common recommendation approaches include content-based filtering, collaborative filtering, knowledge-based filtering, and hybrid filtering [13]. Another survey categorizes recommendation techniques into content-based, collaborative filtering, knowledge-based, and hybrid approaches in the context of big data applications [21].

Collaborative filtering generates recommendations based on user–item interaction patterns and therefore requires data such as ratings, bookings, or user interactions. In contrast, content-based filtering uses item attributes as the basis for recommendation and is more suitable when user interaction history is not available.

In the tourism domain, recommendation systems are not only concerned with accuracy but also need to consider multiple stakeholders such as tourists, service providers, and digital platforms [22]. In the hotel domain, a study in Palangka Raya applied content-based filtering to build a hotel recommendation system based on hotel attributes [14]. Another study in Yogyakarta also implemented a content-based filtering approach to generate hotel recommendations based on available hotel information or descriptions [15]. Additionally, item similarity-based approaches have also been applied in restaurant recommendation systems using Cosine Similarity to assist users in selecting suitable restaurants [23].

3. Methodology

The research methodology is structured following a data science experimental workflow, starting from data loading, preprocessing, feature engineering, modeling, and evaluation of the recommendation function. The research workflow is illustrated in Figure 1.

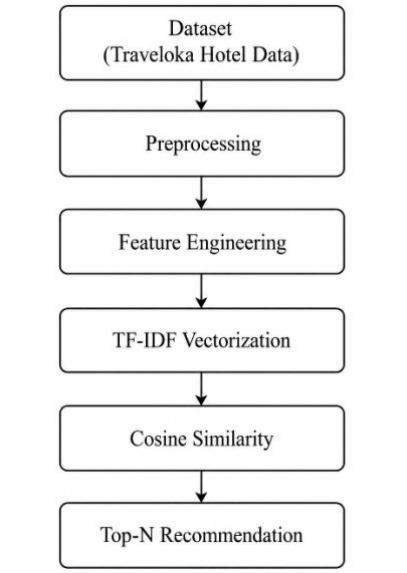

Fig. 1: Research Methodology Flow

3.1. Dataset

The dataset used in this study is a continuation of previous research focused on the acquisition of accommodation data from the Traveloka platform using a web scraping technique based on the data-testid attribute (Latif et al., 2025). In that study, a structured dataset containing hotel data in the Jakarta area was obtained and reused in this research as the basis for developing a recommendation system.

The dataset consists of 1,809 records with nine main attributes, namely Hotel Name, Property Type, Hotel Star Rating, Rating, Location, Discount Price, Original Price, Facilities, and Image URL. Based on initial inspection, not all attributes have complete values. The attributes Hotel Name, Property Type, Hotel Star Rating, Location, Discount Price, and Original Price contain 1,809 complete values, while the Rating attribute has 1,364 values, Facilities has 924 values, and Image URL has 1,048 values.

These differences in data completeness indicate that several attributes require further preprocessing before being used in the recommendation model. To clarify the role of each attribute in the study, Table 1 presents the data types and their roles in the hotel recommendation modeling process.

Table 1: Dataset Attributes and Their Roles in Modeling

Attribute	Data Type	Role in the Study
Hotel Name	Text	Hotel identifier and recommendation input
Property Type	Text	Content feature distinguishing hotels, apartments, guest houses, etc.
Hotel Star Rating	Numeric	Formal quality indicator and supporting description
Rating	Text	Source for extracting rating score and number of reviews
Location	Text	Main feature for spatial similarity representation
Discount Price	Text	Cleaned and converted into numerical price
Original Price	Text	Price comparison information
Facilities	Text	Main content feature representing hotel services
Image URL	Text	Additional attribute for system interface development

3.2. Data Preprocessing

The preprocessing stage is performed to convert raw data into numerical and textual formats suitable for modeling. Discount price is cleaned by removing the currency symbol “Rp”, thousand separators, and spaces, and then converted into integer format. The Rating column is split into two variables, namely Rating Score and Number of Reviews. The rating score is used as a hotel quality indicator, while the number of reviews is used as one component in the weighted popularity score.

Records that do not contain Rating Score, Number of Reviews, Numerical Price, or Property Type are removed from the dataset, as these attributes are required for feature construction and recommendation computation. After preprocessing, the final dataset used for model development consists of 1,364 hotel records.

3.3. Feature Engineering

The feature engineering stage is conducted to construct new features used in the recommendation process. In the first step, a weighted popularity score is computed by combining Rating Score and Number of Reviews. This score ensures that the system does not rely solely on rating values but also considers the reliability derived from review volume.

In the second step, content features are constructed by combining Location, Facilities, and Property Type. These attributes are selected because they represent the main characteristics of hotels relevant for accommodation selection. Location describes the hotel area, facilities represent available services, and property type distinguishes accommodation forms such as hotels, apartments, and other property types.

Before modeling, the content features are normalized into a consistent text format. This includes converting all text into lowercase, replacing commas in location attributes with spaces, and replacing semicolons in facilities with spaces. The resulting output is a unified feature representation that describes each hotel profile and serves as input for TF-IDF weighting.

3.4. TF-IDF Model Construction

After the TF-IDF matrix is constructed, Cosine Similarity is used to compute the similarity between each hotel and all other hotels. This process produces a similarity matrix of size 1,364 × 1,364, representing pairwise relationships among all hotels in the dataset.

The recommendation function takes a hotel name as input and retrieves its index from the DataFrame. After the index is identified, the system extracts similarity scores between the input hotel and all other hotels, sorts the scores in descending order, and removes the input hotel from the recommendation list. The top ten hotels with the highest similarity scores are then selected as recommendation candidates.

The selected candidates are then re-ranked based on the Weighted Popularity Score. This step ensures that the final recommendation results consider not only content similarity based on location, facilities, and property type, but also the strength of user-generated signals such as ratings and review counts. As a result, the system produces Top-N hotel recommendations that are similar in content while also reflecting stronger user evaluation signals.

3.5. Similarity Measurement and Recommendation Function

After the TF-IDF matrix is constructed, Cosine Similarity is used to compute the similarity between each hotel and all other hotels. This process produces a similarity matrix of size 1,364 × 1,364, representing pairwise relationships among all hotels in the dataset.

The recommendation function takes a hotel name as input and retrieves its index from the DataFrame. After the index is identified, the system extracts similarity scores between the input hotel and all other hotels, sorts the scores in descending order, and removes the input hotel from the recommendation list. The top ten hotels with the highest similarity scores are then selected as recommendation candidates.

The selected candidates are then re-ranked based on the Weighted Popularity Score. This step ensures that the final recommendation results consider not only content similarity based on location, facilities, and property type, but also the strength of user-generated signals such as ratings and review counts. As a result, the system produces Top-N hotel recommendations that are similar in content while also reflecting stronger user evaluation signals.

3.6. System Evaluation

The system evaluation is conducted functionally by selecting a single hotel as input, running the recommendation function, and checking whether the system can generate a list of alternative hotels. The evaluation focuses on the consistency of output attributes, particularly location, property type, facilities, similarity score, and popularity score.

Since the dataset does not include explicit ground truth such as user preferences, booking history, or manual relevance labels, the evaluation is limited to functional validation and qualitative interpretation of the recommendation results. Functional validation ensures that the system can accept hotel input, compute inter-hotel similarity, rank candidate recommendations, and display Top-N recommendation results according to the proposed method.

4. Results and Discussion

4.1. Data Cleaning Results

After all preprocessing stages were applied, the original dataset consisting of 1,809 hotel records was reduced to 1,364 records used for modeling. The reduction occurred due to missing essential attributes such as rating score, number of reviews, numerical price, and property type. The retention rate of 75.40% suggests that the dataset maintains sufficient structural completeness for reliable recommendation modeling despite missing attribute distributions.

4.2. Popularity Score Results

The weighted popularity score is used to balance rating and review count in determining hotel credibility. Table 2 presents the top 5 hotels based on this scoring method.

Table 2: Top 5 Hotels Based on Popularity Score

No	Hotel Name	Rating Score	Number of Reviews	Popularity Score
1	BW Express Jakarta Tanah Abang	9.4	415	9.152.884
2	Ashley Tugu Tani Menteng	9.5	238	9.093.935
3	Ashley Tanah Abang	9.1	983	9.010.223
4	Ashley Tang Menteng	9.1	874	9.000.178
5	Citadines Gatot Subroto Jakarta	9.3	288	8.997.872

Table 2 shows that hotels with a higher number of reviews tend to achieve more stable ranking positions compared to hotels with high ratings but very few reviews. This confirms that review volume plays an important role in stabilizing hotel popularity estimation.

4.3. Dataset Representation Results

The processed dataset consists of 1,364 hotels with structured attributes used for modeling, including location, facilities, and property type. These attributes form the basis for content representation in the recommendation system. The final dataset structure enables uniform representation of hotel characteristics, allowing similarity-based ranking and recommendation generation.

4.4. Recommendation Testing Results

The recommendation function was tested using “Pejaten Valley Residence” as the input hotel. The Top-10 recommendation output demonstrates that the model effectively preserves similarity constraints while maintaining ranked diversity among accommodation alternatives. The final results were ranked based on the popularity score, as shown in Table 3.

Table 3: Sample Recommendation Results for “Pejaten Valley Residence”

No	Hotel Name	Property Type	Location	Rating Score	Number of Reviews	Popularity Score	Price
1	D'rhea Syariah	Hotel	Pejaten Barat, Jakarta	8.8	316	8.638.615	266175
2	Cove Arimbi	Hotel	Pejaten Barat, Jakarta	8.7	231	8.527.468	260101
3	Super OYO Capital O 133 Griya Ciaji	Guest House	Pejaten Barat, Jakarta	8.3	352	8.262.297	392335
4	Pejaten Valley Residence By Zuzu	Hotel	Pejaten Barat, Jakarta	10.0	1	8.148.382	356620
5	Comfort And Modern Look 3Br Apartment Royal Olive Residence	Apartment	Pejaten Barat, Jakarta	8.5	1	8.133.567	1211417
6	Hester Basoeki (HB) Garden Guest House	Guest House	Cilandak Barat, Jakarta	8.1	106	8.114.539	314880
7	RedDoorz Syariah near Prasetya Mulya Cilandak	Guest House	Cilandak Barat, Jakarta	8.0	112	8.061.360	225184
8	Cove Birah at Senopati	Guest House	Rawa Barat, Jakarta	8.0	121	8.058.864	288777
9	Residence 21 Syariah	Hotel	Pejaten Barat, Jakarta	6.1	6	8.015.282	144186
10	OYO 121 Rumah Ayub Syariah Near Rumah Sakit JMC	Hotel	Pejaten Barat, Jakarta	7.5	498	7.605.556	225713

Table 3 shows that most recommended hotels are located in nearby areas such as Pejaten Barat and Cilandak Barat, indicating that location is a dominant factor in the recommendation process. In addition, recommended hotels also share similar property types, such as hotels and guest houses, indicating consistency in accommodation categories. Interestingly, hotels with extremely high ratings but very low review counts do not dominate the top rankings, as the popularity score adjusts the final ranking.

4.5. Discussion

The experimental results show that the proposed recommendation system is capable of generating meaningful hotel recommendations based on structured accommodation data extracted from OTA platforms. The observed consistency in recommendation patterns suggests that the feature representation effectively captures structural similarities among accommodation entities. The integration of popularity-based scoring improves ranking stability by reducing bias toward hotels with insufficient review data.

However, the system still has limitations. The evaluation does not yet include ground truth or user interaction data, making quantitative performance measurement unavailable. In addition, missing values in several attributes may affect recommendation quality. The model also does not yet incorporate personalized user preferences such as budget constraints or travel purpose.

5. Conclusion

This study successfully developed a hotel recommendation system based on a content-based filtering approach using Traveloka accommodation data from the Jakarta region. The initial dataset consisted of 1,809 hotel records, and after the data preprocessing stage, 1,364 records were used for model development. The content features were constructed by combining location, facilities, and property type attributes, which were then represented using TF-IDF. The modeling process produced a TF-IDF matrix of size 1,364 × 371 and a Cosine Similarity matrix of size 1,364 × 1,364.

Functional testing indicates that the system is capable of generating a list of alternative hotels with content similarity to the input hotel. In the case study using "Pejaten Valley Residence," the recommended hotels were predominantly located in nearby areas and shared relatively similar property types. The addition of a weighted popularity score improved the ranking process, as the final results not only considered attribute similarity but also incorporated user rating signals through a combination of rating scores and the number of reviews. Therefore, this study extends previous work from the data acquisition stage to the utilization of data for developing a machine learning-based hotel recommendation system.

Several future improvements are suggested. First, the review count parsing process should be improved so that abbreviated formats such as "5k" and "17k" can be accurately converted into numerical values. Second, quantitative evaluation metrics such as Precision@K, Recall@K, NDCG, or MAP should be applied when ground truth data such as user preferences, interaction logs, or expert relevance assessments are available. Third, the model can be extended into a hybrid recommendation approach by combining content-based filtering and collaborative filtering when user interaction data becomes available. Fourth, additional features such as price, hotel star rating, and geographic distance can be incorporated as reranking components to better align recommendations with user needs. Fifth, sentiment analysis of hotel reviews can be integrated to enrich content profiles and improve recommendation personalization.

References

[1] N. Khairunnisa, A. Hermawan, and R. G. Guntara, "Strategi Pemasaran Untuk Meningkatkan Occupancy Kamar Hotel Melalui Online Travel Agent Di Indies Hotel Bandung," vol. 13, no. 2023, pp. 2417–2423, 2025.

[2] A. Latif, S. K. Wildah, S. Agustiani, and E. H. Juningsih, "Implementation of a Data-Testid Attribute-Based Web Scraping Method for Accommodation Data Extraction from a Dynamic E-Commerce Website (Case Study : Traveloka)," vol. 5, no. 1, 2025.

[3] I. Hossain *et al.*, "a survey o f recommender system techniques and the e − commerce domain - hossain 2023.pdf," 2022.

[4] Y. Ge *et al.*, "A Survey on Trustworthy Recommender Systems," vol. 1, no. 1, pp. 1–67, 2024.

[5] K. Yi, R. Yamagishi, T. Li, Z. Bai, and Q. Ma, "Recommending POIs For Tourists By User Behavior Modeling and Pseudo-Rating," 2021.

[6] V. T. Camacho and J. Cruz, "Ontology-based Context Aware Recommender System Application for Tourism .," pp. 1–41, 2022.

[7] V. Vargas-Calderón, A. M. Ochoa, G. Y. C. Nieto, and J. E. Camargo, "Machine learning for assessing quality of service in the hospitality sector based on customer reviews," no. 40, 2021.

[8] L. Aravani, E. Pintelas, C. Pierrakeas, and P. Pintelas, "A Natural Language Processing Framework for Hotel Recommendation based on user's text reviews.," 2024.

[9] P. Khadka and P. Lamichhane, "Content-based Recommendation Engine for Video Streaming Platform," 2025.

[10] A. Banerjee, T. Mahmudov, and E. Adler, "Modeling Sustainable City Trips : Integrating CO 2 e Emissions , Popularity , and Seasonality into Tourism Recommender Systems," pp. 1–38, 2024.

[11] Z. Wang and D. Jannach, "A Survey on Point-of-Interest Recommendations Leveraging Heterogeneous Data," pp. 1–51, 2024.

[12] S. Suzuki, "Use of online travel agencies as a data source for tourism marketing," *J. Glob. Tour. Res.*, vol. 5, no. 2, pp. 167–171, 2020, doi: 10.37020/jgtr.5.2_167.

[13] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *J. Big Data*, 2022, doi: 10.1186/s40537-022-00592-5.

[14] N. N. K. Sari, Licantik, and M. Zahra, "Pemanfaatan Sistem Rekomendasi Menggunakan Content- Based Filtering pada Hotel di Palangka Raya," vol. 15, no. 4, pp. 754–763, 2024.

[15] C. A. Melyani, A. Kesumawati, R. Bagus, and F. Hakim, "Hotel Recommendation System with Content-Based Filtering Approach (Case Study : Hotel in Yogyakarta on Nusatrip Website) Department of Statistics , Universitas Islam Indonesia," vol. 15, no. 1, pp. 152–157, 2022.

[16] D. Pratiwi, Asrianda, and L. Rosnita, "Penerapan Metode Content-Based Filtering dalam Sistem Rekomendasi Objek Wisata di Aceh Tamiang," vol. 4, no. 2, pp. 85–96, 2024.

[17] Scikit-learn Developers, "TfidfVectorizer," Scikit-learn Documentation. Accessed: May 14, 2026. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[18] Scikit-learn Developers, "cosine_similarity," Scikit-learn Documentation. Accessed: May 14, 2026. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html

[19] T. Ma, M. S. Bernstein, R. Johari, and N. Garg, "Balancing Producer Fairness and Efficiency via Prior-Weighted Rating System Design," 2025.

[20] A. Alhadlaq, S. Kerrache, and H. Aboalsamh, "A Recommendation Approach based on Similarity-Popularity Models of Complex Networks," pp. 1–12, 2022.

[21] Z. Xia, A. Sun, J. Xu, Y. Peng, and M. Cheng, "Contemporary Recommendation Systems on Big Data and Their Applications : A Survey," vol. 12, no. July, 2024.

[22] A. Banerjee, P. Banik, and W. Wörndl, "Towards Individual and Multistakeholder Fairness in Tourism Recommender Systems," 2023.

[23] F. Christyawan, A. N. Rohman, and A. D. Hartanto, "Application of Content-Based Filtering Method Using Cosine Similarity in Restaurant Selection Recommendation System," vol. 6, no. 3, pp. 1559–1576, 2024, doi: 10.51519/journalisi.v6i3.806.