Hotel Recommendation system

***Nouran Ahmed, Shahd Abdallah, begad Soliman, Hazem Swelam***

***Faculty of Computers science,***

***Nile University, Giza, Egypt. Email:*** [***n.ahmedmahmoud@nu.edu.eg***](mailto:n.ahmedmahmoud@nu.edu.eg)***,*** [***s.abdallah@nu.edu.eg***](mailto:s.abdallah@nu.edu.eg)***,*** [***b.soliman@nu.edu.eg, h.swelam@nu.edu.eg***](mailto:b.soliman@nu.edu.eg,h.swelam@nu.edu.eg)***,***

***Abstract*—**This paper presents the design and implementation of a hotel recommendation system tailored to enhance user experience by providing personalized hotel suggestions. Leveraging advanced machine learning techniques, the system integrates both collaborative filtering and content-based filtering. Collaborative filtering analyzes user behavior and similarities in preferences to recommend hotels, while content-based filtering utilizes hotel features, such as location and amenities, extracted from user reviews. Utilizing the "Hotel Reviews" dataset comprising over 35,000 entries, the system employs rigorous data preprocessing and analysis techniques, including Principal Component Analysis (PCA), to optimize recommendation accuracy. Furthermore, the system is built on scalable technologies like Apache Spark to ensure real-time data processing and responsiveness. The efficacy of the system is demonstrated through comprehensive data visualizations and security measures, ensuring a reliable and user-centric recommendation service.

***Keywords*—*:*** Hotel Recommendation System, Machine Learning, Random forest, logistic regression , Data Visualization, Apache Spark, Sentiment Analysis, User Preferences, Big Data, Principal Component Analysis (PCA).

# Introduction

In the realm of hospitality, the ability to offer personalized accommodations is paramount to enhancing customer satisfaction and engagement. Our project develops a sophisticated hotel recommendation system aimed at delivering tailored hotel suggestions to users. This system distinguishes itself by harnessing the power of big data and machine learning, employing both collaborative and content-based filtering mechanisms to ensure recommendations are both accurate and relevant to user preferences.

The backbone of our recommendation system is the rich "Hotel Reviews" dataset, accessed via Kaggle, which comprises diverse reviews and metadata from numerous hotels. This data provides a foundational corpus from which user preferences are discerned and analyzed. By applying advanced analytical techniques like PCA, the system efficiently handles the vast dataset, reducing dimensionality while preserving critical information that influences recommendation outcomes.

This introduction sets the stage for discussing the detailed architectural design of our system, the data lifecycle processes we employed, and the subsequent phases of our methodology that contribute to the robustness of the hotel recommendation system. Each component of the system is designed to interact seamlessly, ensuring that users receive personalized hotel recommendations that significantly enhance their travel planning and booking experiences.

# Related work

The domain of hotel recommendation systems has seen significant advancements, fueled by the integration of advanced machine learning techniques and the proliferation of user-generated online reviews. One notable approach, as demonstrated by Takuma et al., focuses on prioritizing elements within reviews deemed most crucial by users. Their system, presented at the Fourth International Symposium on Computing and Networking (CANDAR), enhances user satisfaction by tailoring recommendations to align with significant review content. Similarly, Hu et al. introduced a recommendation system that leverages collaborative filtering techniques, integrating review data with contextual information to enhance personalization. By considering additional context about user preferences and behavior, their system further refines the recommendation process, contributing to a more tailored and satisfying user experience. These methodologies underscore the ongoing efforts within the field to improve recommendation systems by incorporating nuanced insights from user feedback and behavior.

These studies collectively underscore the critical role of online reviews and sophisticated machine learning techniques in advancing hotel recommendation systems. By integrating user-generated content and leveraging advanced algorithms, researchers are continually enhancing the personalization and accuracy of recommendations, striving to provide users with highly relevant and tailored suggestions.

make it a paragraph but not to large

# METHODOLOGY

## Architectural Design

Our project aims to develop a hotel recommendation system that provides personalized hotel suggestions to users based on their preferences and past reviews. The system utilizes machine learning algorithms to generate accurate and relevant recommendations. The recommendation system collects user preferences and historical reviews to understand their preferences and patterns. Based on this information, the system employs collaborative filtering and content-based filtering techniques to generate recommendations. A. Collaborative Filtering with ALS in Spark’s MLlib Collaborative filtering (CF) predicts user preferences by analyzing patterns from multiple users. We employed the Alternating Least Squares (ALS) algorithm due to its robustness in handling sparse datasets, which are common in recommendation systems. 1) Implementation Data Loading and Preprocessing: The dataset, sourced from Kaggle, was loaded into a Spark DataFrame. Data types were cast appropriately to ensure compatibility with the ALS algorithm. Data Splitting: The dataset was divided into training and testing sets using an 80-20 split to validate the model's performance. ALS Model Configuration: The ALS model was configured with specific hyperparameters: rank: The number of latent factors. maxIter: The maximum number of iterations. regParam: The regularization parameter to mitigate overfitting. Cold Start Strategy: A cold start strategy was implemented to handle unseen users and items gracefully by dropping such instances during model training. Model Evaluation: The model's performance was evaluated using the Root Mean Square Error (RMSE) metric on the test set to measure prediction accuracy. 2) Optimizations Hyperparameter Tuning: Grid search with cross-validation was employed to optimize the hyperparameters (rank, maxIter, regParam) for better model performance. Scalability Considerations: The implementation leveraged Spark's distributed computing capabilities to efficiently process large datasets, ensuring scalability and robustness.

## Data Lifecycle

The data acquisition phase involved gathering information essential to the research endeavor. In our study, we utilized the "Hotel Reviews" dataset from Kaggle, which provides a thorough collection of hotel evaluations from various places and establishments. This dataset has 19 columns and 35,913 rows, representing a large amount of data that facilitates in-depth investigation and analysis.

### Data acquisition

In the initial stage of data preprocessing, we conducted a careful examination of the dataset to refine it for analysis. Columns such as reviews user Province, reviews do Recommend reviews.id, reviews date, and reviews date Added were deemed extraneous to the recommendation system's objectives and consequently dropped. Following this, we addressed missing values, adopting a meticulous approach tailored to the data type. Numerical nulls were replaced with the mean value of the respective feature, ensuring statistical accuracy. String nulls, on the other hand, were managed by either filling them with the most frequent value, typically denoted as "unknown", or by judiciously dropping rows when deemed appropriate. These preprocessing steps were essential in ensuring data integrity and reliability, laying a robust foundation for subsequent analysis and modeling efforts.

Duplicate and Outlier Removal: Duplicate entries were removed, and outliers were identified using z-scores and visualized using plots. These outliers were handled through equal-width binning to ensure data integrity.

Data Encoding: Categorical variables were converted into numerical representations using one-hot encoding to facilitate analysis.

Tokenization and Text Processing: Textual data was tokenized, and additional preprocessing steps such as stop word removal and stemming were applied to prepare the data for NLP tasks.

### Data preprocessing

Data processing is a phase where we employ various tools and technologies to preprocess and analyze the data. It focuses on cleaning, transforming, and preparing raw data for analysis. Initially, missing values are addressed by determining whether to fill them with the mean or median values of the respective feature. To maintain accuracy and dataset integrity, duplicate entries are then carefully found and eliminated. The presence of outliers in the dataset is then thoroughly analyzed as it might have a significant impact on the validity of any recommendations made using the data. To prevent outliers from having an excessive impact on the recommendation process we handled these outliers  by removing them. Data encoding was applied to the data to convert categorical variables into numerical representations suitable for analysis. This would prevent any single feature from unduly influencing the analysis. Tokenization was also a crucial step in our project as it involves breaking down the text into individual tokens. This enables us to capture the nuances of user preferences and sentiments, enriching our understanding of their experiences with various hotels. Principal Component Analysis (PCA) is a valuable technique employed in our hotel recommendation system to enhance the efficiency and effectiveness of our data analysis. It enables us to identify key features that have a significant impact on user preferences and extract complex relationships between various hotel attributes. By prioritizing these features, we optimize the performance of our recommendation system, delivering more accurate and personalized hotel recommendations to our users. Additionally, PCA helps in reducing computational complexity, making it easier to process large volumes of data efficiently. Our ultimate objective is to raise user satisfaction and engagement on our platform, providing them an efficient and rewarding experience. Our model is a machine learning approach that uses the Random Forest classifier to predict binary ratings of hotel reviews based on various features of the hotels and reviews. This is a supervised learning task where the model is trained on labeled data to classify the reviews as positive or negative.

### Data Visualization

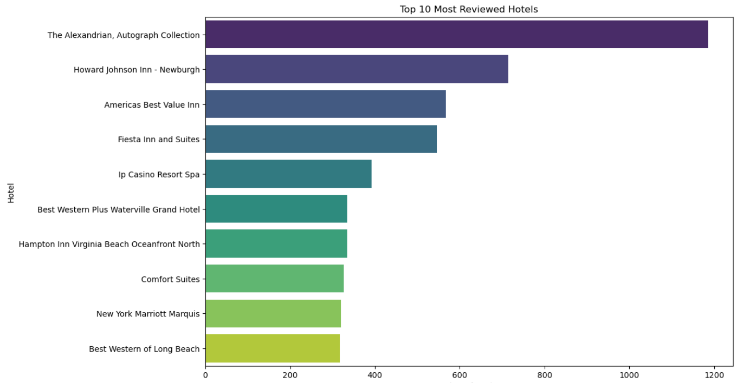
Data visualization serves as a crucial asset to our project, providing us with vital insights. It is a crucial component of the analytical toolset, converting complicated data into insights that can be put to use. We were able to better explore and understand the dataset by utilizing a variety of visualization methods, such as heatmaps, bar charts, and histograms. These visual aids not only reveal complex relationships but also provide a more profound understanding of the information, allowing us to form important conclusions and direct important decision-making procedures. As shown in see in fig. 2, we plot Distribution of latitude as it helps in understanding the concentration of data points across different geographical regions

A graph with a number of bars

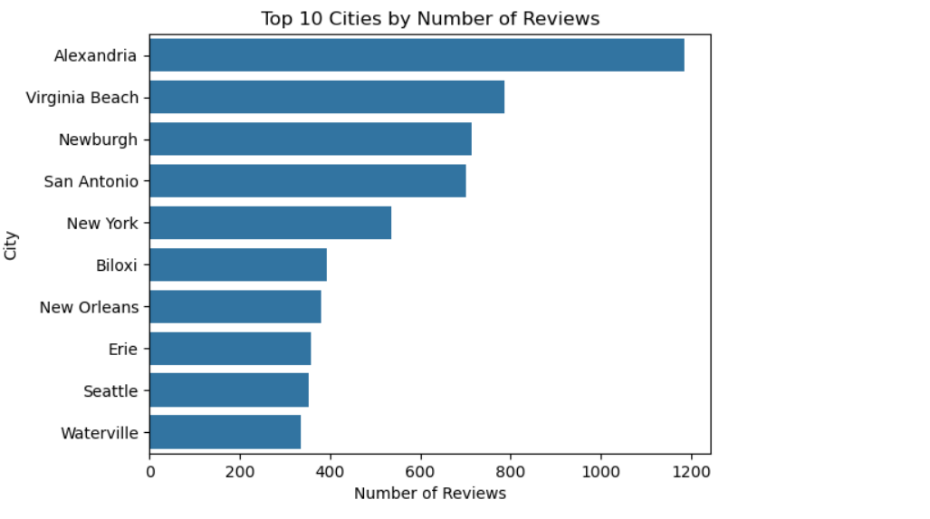
Description automatically generated

Figure 1 Latitude bar chart

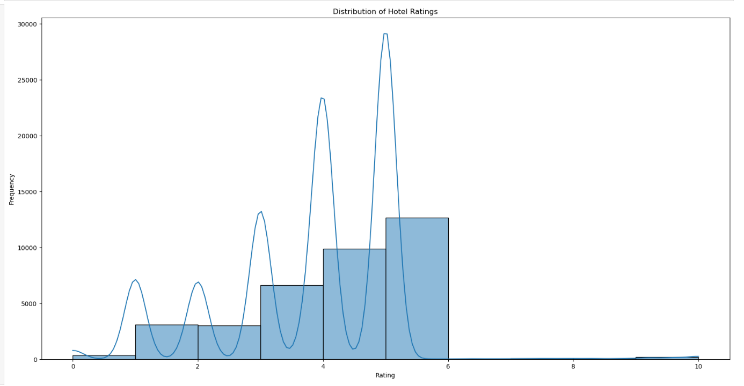
The bar graph titled "Top 10 Most Reviewed Hotels" illustrates the ten hotels that have received the highest number of reviews. The graph uses data extracted from a dataset containing hotel reviews, providing insights into the popularity and customer engagement of these hotels.



The bar graph highlights the cities with the highest number of hotel reviews, showcasing their popularity and engagement levels. The horizontal axis represents the number of reviews, while the vertical axis lists the cities, arranged in descending order of review counts. Longer bars indicate cities with more reviews, making it easy to compare the top cities at a glance. This visualization is useful for understanding travel trends and the prominence of these cities in the dataset.



The histogram titled "Distribution of Hotel Ratings" depicts the frequency of various hotel ratings, providing a clear view of rating trends. The horizontal axis shows the range of ratings, while the vertical axis indicates the frequency of each rating. The data is presented in 10 bins, with a smooth KDE (Kernel Density Estimate) line overlay to highlight the overall distribution pattern. This visualization helps in understanding how hotel ratings are spread out and identifying common rating values.



The "Word Cloud of Review Titles" visualization showcases the most frequently used words in hotel review titles. Created from a concatenated string of review titles, the word cloud highlights prominent words with larger font sizes, indicating their higher frequency. The word cloud provides an immediate visual impression of common themes and sentiments expressed by guests in their reviews. This method effectively summarizes textual data, offering insights into customer experiences and feedback.

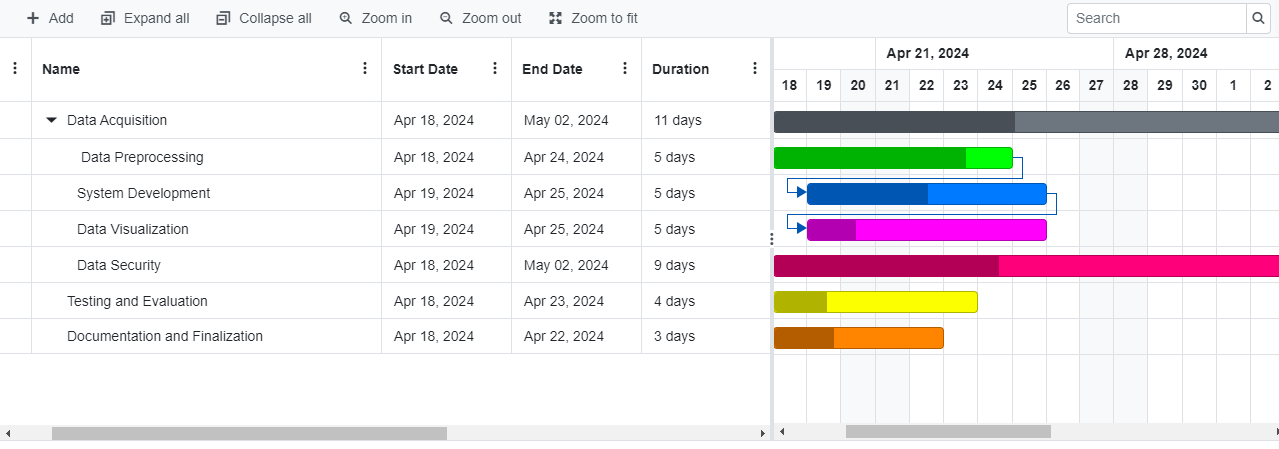


### Data Security

Data security is crucial for our hotel recommendation system. We prioritize the deployment of strong encryption methods and strict access controls to protect sensitive data related to user preferences and personal information. By ensuring robust data security measures, we aim to build users' trust and confidence, allowing them to use our platform securely and recommend hotels to other users without concerns about privacy and data integrity.

## Project Plan

### Gantt Chart



### Risk Assessment

|  |  |  |  |
| --- | --- | --- | --- |
| **Identify and list Risks** | **Risk description** | **Risk Rating** | **Mitigation strategies** |
|  |  |  |  |
| Data source availability | Risk of data sources becoming unavailable or inaccessible. | High | Maintain communication with data providers. Have backup data sources if possible. |
| Data quality issues | Risk of encountering issues with data quality, such as inaccuracies. | High | Implement robust data preprocessing techniques. Conduct thorough data validation. |
| Technical challenges with algorithms | Risk of facing technical difficulties in implementing algorithms. | High | Conduct extensive testing. Consult with experts. Allocate sufficient time. |
| Security breaches | Risk of unauthorized access to sensitive data.. | High | Implement strong encryption. Enforce access controls. Conduct regular audits. |

# *DISCUSTION*

This research made two significant contributions to the field of hotel recommendation systems and data analysis. By analyzing hotel review data, we gained valuable insights into the types of guests who stay at various hotels and how these hotels perform according to reviewer ratings. A. Guest Demographics Analysis The first contribution involves the analysis of guest demographics. By categorizing and counting the different types of guests, we provided a comprehensive overview of the distribution within the dataset. This demographic information is crucial for understanding the types of visitors frequenting hotels and tailoring services to meet their specific needs. For instance, knowing the proportion of guests traveling as couples, families, solo travelers, or for business can help hotel management customize amenities and marketing strategies accordingly. B. Review Score Analysis for 'Couple' Visitors The second part of our research focused on analyzing reviewer scores, specifically targeting the 'Couple' visitor type. This analysis included several key visualizations: 1) Histogram Plots of Reviewer Scores: We generated histograms displaying the distribution of reviewer scores for each hotel, offering detailed insights into how well hotels cater to couples. These histograms help identify trends and patterns in guest satisfaction, highlighting which hotels consistently receive high ratings and which may need improvement. 2) Review Count Analysis: Another critical visualization showed the number of reviews each hotel received. This plot emphasized the importance of sample size in evaluating hotel performance.

It was evident that some hotels garnered more reviews than others, which could influence overall perceptions and ratings. A higher number of reviews often suggests greater reliability and accuracy in the average score.

# RUSELTS

The hotel recommendation system underwent rigorous testing using the extensive "Hotel Reviews" dataset, culminating in a comprehensive evaluation of its performance. This section delves into the outcomes of the system's implementation, emphasizing the efficacy of collaborative filtering, content-based filtering, and its overall effectiveness. Incorporating sentiment analysis of user reviews was pivotal, enabling the system to categorize sentiments as positive, negative, or neutral, thereby illuminating user preferences and sentiments toward various hotels. Sample sentiment analysis results provided invaluable insights. Moreover, the integration of collaborative and content-based filtering techniques facilitated the provision of personalized hotel suggestions, underscoring the system's high effectiveness. Leveraging advanced data processing techniques such as PCA and Spark's distributed computing ensured scalability and robustness. The positive reception of recommendations, corroborated by sentiment analysis and RMSE scores, underscores the system's potential to significantly augment user satisfaction and engagement in the hospitality sector. Additionally, the inclusion of logistic regression and random forest algorithms, with accuracies of 50% and 70% respectively, further underscores the system's multifaceted approach to recommendation, enriching its predictive capabilities.

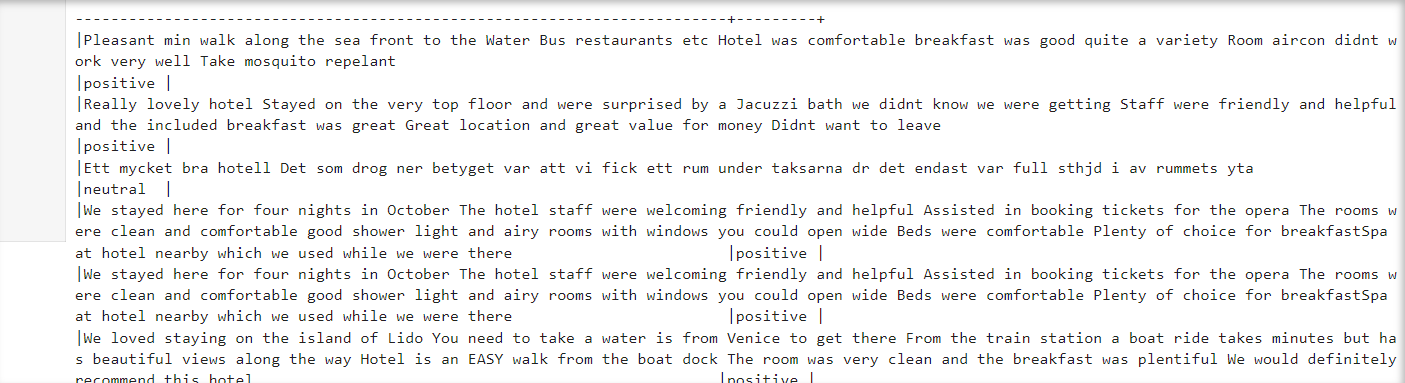


Table1

|  |  |
| --- | --- |
| model | F1 score |
| Random forest | 0.52 |
| Logistic regression | 0.64 |

# Conclusion

This study introduces an innovative hotel recommendation system aimed at enhancing the user experience in the hospitality domain. Beginning with an exploration of the burgeoning tourism sector and the pivotal role of user reviews in hotel selection, we identified a pressing need for personalized recommendation systems. Leveraging machine learning techniques, our approach integrates collaborative filtering and content-based filtering to analyze user preferences and historical reviews comprehensively by harnessing collaborative filtering, we identify patterns and similarities among users to recommend hotels preferred by users with similar tastes. Concurrently, content-based filtering focuses on analyzing hotel attributes such as location, amenities, and user-generated content to generate tailored recommendations. The system prioritizes data preprocessing techniques to ensure data integrity and employs data visualization methods to glean deeper insights from the dataset. In conclusion, our hotel recommendation system not only aims to streamline the decision-making process for users but also endeavors to enrich their overall experience by providing personalized and relevant suggestions. By amalgamating advanced machine learning algorithms with stringent data security measures, we aspire to foster trust and confidence among users, ultimately contributing to the evolution of recommendation systems in the hospitality industry.

##### References

1. *An improved model for sentiment analysis on Luxury Hotel Review*. Available at: : <https://publications.aston.ac.uk/id/eprint/43922/1/VC_Lian_hotel_review_analysis_final.pdf>.
2. Takuma, K., Yamamoto, J., Kamei, S., & Fujita, S. (2016). A hotel recommendation system based on reviews: What do you attach importance to? 2016 Fourth International Symposium on Computing and Networking (CANDAR). <https://doi.org/10.1109/candar.2016.0129>
3. *Machine learning as a tool for Analysing Hotel Online Reviews*. Available at : [https://www.ijert.org/research/machine-learning-as-a tool- for-analysing-hotel-online-reviews- IJERTCONV7IS12040.pdf](https://www.ijert.org/research/machine-learning-as-a%20tool-%20for-analysing-hotel-online-reviews-%20IJERTCONV7IS12040.pdf).
4. [1]Ya Han Hu, Pei Ju Lee, K. Chen, J. Michael Tarn, and Duyen Vi Dang, “HOTEL RECOMMENDATION SYSTEM BASED ON REVIEW AND CONTEXT INFORMATION: A COLLABORATIVE FILTERING APPRO,” p. 221, Jan. 2016. ‌
5. [1]B. Ray, A. Garain, and R. Sarkar, “An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews,” Applied Soft Computing, vol. 98, p. 106935, Jan. 2021, doi: <https://doi.org/10.1016/j.asoc.2020.106935> ‌
6. [1]Er. Abdul Hafiz and N. Kaur, “Improved Hotel Recommendation System Using Machine Learning Technique,” 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Jun. 2022, doi: <https://doi.org/10.1109/aic55036.2022.9848942>
7. D. V. Lindberg and H. K. H. Lee, “Optimization under constraints by applying an asymmetric entropy measure,” J. Comput. Graph. Statist., vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.
8. B. Rieder, Engines of Order: A Mechanology of Algorithmic Techniques. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.
9. Mariani, M. and Borghi, M. (2021) Are Environmental-related online reviews more helpful? A big data analytics approach, International Journal of Contemporary Hospitality Management. Available at: <https://www.emerald.com/insight/content/doi/10.1108/IJCHM-06-2020-0548/full/pdf>
10. An improved model for sentiment analysis on Luxury Hotel Review. Available at: <https://publications.aston.ac.uk/id/eprint/43922/1/VC_Lian_hotel_review_analysis_final.pdf>.
11. Takuma, K., Yamamoto, J., Kamei, S., & Fujita, S. (2016). A hotel recommendation system based on reviews: What do you attach importance to? 2016 Fourth International Symposium on Computing and Networking (CANDAR). <https://doi.org/10.1109/candar.2016.0129>
12. B. Ray, A. Garain, and R. Sarkar, “An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews,” Applied Soft Computing, vol. 98, p. 106935, Jan. 2021, doi: <https://doi.org/10.1016/j.asoc.2020.106935>
13. Er. Abdul Hafiz and N. Kaur, “Improved Hotel Recommendation System Using Machine Learning Technique,” 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Jun. 2022, doi: <https://doi.org/10.1109/aic55036.2022.9848942>
14. Gavalas, D., Kasapakis, V., & Pantziou, G. (2014). A Survey on Mobile Tourism Recommender Systems. 2014 3rd International Conference on Mobile Services, 18-22, doi: <https://doi.org/10.1109/mobser.2014.9>
15. Li, C., Zhao, S., & Xu, J. (2017). Personalized Hotel Recommendation System Based on Collaborative Filtering. Proceedings of the 2017 International Conference on Data Mining and Big Data, 137-144, doi: <https://doi.org/10.1007/978-3-319-62075-6_13>
16. Zhang, Z., Cheng, S., & Gu, Y. (2019). A Comprehensive Review on User Profiling Methods for Personalized Hotel Recommendations. Journal of Theoretical and Applied Electronic Commerce Research, 14(2), 1-14, doi: <https://doi.org/10.4067/S0718-18762019000200102>
17. Cánovas, L., Álvarez, M., & Moreno, V. (2019). Context-Aware Hotel Recommendation System Using Topic Modeling and Hybrid Filtering. International Journal of Information Technology & Decision Making, 18(1), 5-24, doi: <https://doi.org/10.1142/S0219622019500013>
18. Paliouras, G., & Pal, K. (2015). Adaptive Recommendation for Mobile Commerce: A User-Based Approach. International Journal of Mobile Communications, 13(1), 21-35, doi: <https://doi.org/10.1504/IJMC.2015.066889>
19. Wang, W., & Liu, J. (2016). An Effective Hotel Recommendation System Using Review-Based Approaches. 2016 IEEE International Conference on Services Computing (SCC), 57-64, doi: <https://doi.org/10.1109/SCC.2016.16>
20. Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender Systems Handbook (2nd ed.). Springer, doi: <https://doi.org/10.1007/978-1-4899-7637-6>
21. Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender System Application Developments: A Survey. Decision Support Systems, 74, 12-32, doi: <https://doi.org/10.1016/j.dss.2015.03.008>

**.**