From Reviews to Reservations: A Data Science Approach to Hotel Recommendations

1st Bakht Singh Basaram
Dept. of Data Science
University of North Texas
Texas, US
bakhtsinghbasaram@my.unt.edu

2nd Kavya Kasthuri Dodle
Dept. of Data Science
University of North Texas
Texas, US
kavyakasthuridodle@my.unt.edu

3rd Srinath Reddy Kotha
Dept. Data Science
University of North Texas
Texas, US
srinathreddykotha@my.unt.edu

4th Sanjay Kumar Machanapally
Dept. Data Science
University of North Texas
Texas, US
sanjaykumarmachanapally@my.unt.edu

***Abstract*—Making decisions on which hotels to choose remains one of the major challenges for customers due to the widespread review systems with the vast amount of unstructured textual data. The diversity of opinions and experiences shared by users often makes the final decision on the hotel choice impossible and causes dissatisfaction with any chosen options. To mitigate the consequences of this issue, we introduce a recommender system based on machine learning. Our project strives to find the most important bottom-lines in Trip Advisor hotel reviews by scraping them and munging the gained textual data. The further implementation uses clusterization algorithms to cluster reviews of one type, use sentence transformers and further simplify customer choice of hotels. The implementation of the recommender system would take businesses to another level of customer centricity and strengthen their reputation.**

***Index Terms*—Hotel, Machine Learning, Clustering, NLP, K-means, TripAdvisor**

I. Introduction and Statement of the Problem

Customer satisfaction is crucial in the hotel industry. Discovering customer preferences and making personalized recommendations can greatly improve the overall customer experience. Customer reviews and opinions about different hotels are now available thanks to online reviews like TripAdvisor. If used effectively, this information can provide valuable information to hotel management and potential customers.

With so many choices, customers often can't decide. One way to facilitate customer decision-making is to provide a system that effectively filters reviews and recommends the most suitable options. Hotels that can use recommendation engines gain a competitive advantage in the market. Hotels can do more business by offering personalized recommendations based on customer preferences.

The emergence of online review platforms has given customers a strong voice to share their experiences with other people. Customers can rate and review hotels restaurants and attractions in-depth on sites like TripAdvisor. Thanks to recent developments in text mining and natural language processing (NLP) techniques, valuable insights can now be extracted from textual data. Techniques such as topic modeling, sentiment analysis, and text summarization are crucial for understanding and analyzing customer reviews. Finding patterns and putting related data points in a group is possible by using machine learning algorithms like K-means clustering. Clustering algorithms can be used to find groups of hotels that share traits or customer opinions in the context of hotel reviews.

Across all online platforms, including streaming services and e-commerce websites, personalized recommendations are now commonplace. Businesses can increase customer satisfaction and engagement by customizing recommendations based on individual preferences.

II. Review of Literature

In the field of hotel recommendation systems, recent research has explored various ways to help travelers make informed decisions based on online reviews.

An ensemble-based hotel recommender system that involves sentiment analysis and aspect categorization reviews of hotels are introduced by the notable study done by Ray et al. (2021). The authors propose a systematic approach, utilizing Random Forest classifiers for aspect-based classification and Bidirectional Encoder Representations from Transformers (BERT) models for sentiment analysis. With a macro F1-score of 84% and a test accuracy of 92.36% in sentiment classification, their approach delivered impressive results. The ability of machine learning techniques to extract information from online reviews to help in the process of decision-making is highlighted in this study.

Additionally, the research paper Akhtar et al. (2017) focuses on aspect-based, sentiment-oriented summarization of hotel reviews. The authors would like to provide users with clear and informative analyses of review aspects and sentiments through the combination of classification and topic modeling that uses Latent Dirichlet Allocation (LDA), sentiment analysis, and summarization techniques. Their results demonstrate the important advantages of review summarization in improving the decision-making process of the users and also highlight the significance of summarization techniques in assisting users in decision-making processes.

Additionally, the opinion mining and summarization approach for hotel reviews Raut and Londhe (2014) provide users or customers with their decision-making processes. The authors obtain an accuracy of around 87% in categorizing hotel review content as positive or negative comments by utilizing machine learning techniques and SentiWordNet for opinion classification and summarization. This study contributes to the expanding field of research on opinion mining and summarization techniques, which offer to extract information from online reviews by providing useful insights into effective strategies.

In addition, Chang et al. (2020) use deep learning and visual analytics technologies to analyze hotel reviews and relevant responses obtained from TripAdvisor. With this method, they intend to identify the response strategies and also provide insights into the decision-making processes of hotel management. The authors represent an innovative perspective on the complicated relationships between reviews and managerial responses through the use of computational linguistics, visual analytics, and deep learning techniques. This also highlights the use of important applications of smart technologies that can assist hotel representatives in decision-making processes.

All these studies collectively highlight the wide range of techniques and strategies that were used in the field of hotel recommendation systems. Researchers are always looking to explore new innovative strategies that can enhance the usage and efficiency of online review platforms in supporting traveler's decision-making processes. So, all these methods range from sentiment analysis and aspect categorization to opinion mining and summarization.

III. Objectives of the Study

The project's main goal is to build a sophisticated recommender system by utilizing NLP and machine learning methods. This system will evaluate textual hotel evaluations via an easy-to-use interface, categorize them according to semantic similarities, forecast customer preferences, and provide customized hotel recommendations. The main goals are enhancing customization in hotel suggestions, expediting the vacation planning process, and providing consumers with well-informed decision-making tools.

- Utilize NLP techniques: Employ Natural Language Processing (NLP) methodologies to extract valuable insights from textual hotel reviews. This involves preprocessing text, analyzing sentiment, identifying topics, and extracting key features to capture the essence of user feedback.
- Implementing Text Clustering: Utilize text clustering methods to combine comparable hotel reviews according to their semantic content.
- Develop a Predictive Model: Build a predictive model with the ability to suggest hotels to users based on their reviewers' specified values and interests. By using sentence transformers that have been pre-trained on one billion sentence pairs, this model will be able to predict which hotels will most likely match the interests of individual users.
- Design an Intuitive User Interface: Make the recommender system's interface easy to use so that users can engage with it without any difficulty. Features that improve the user experience overall, such as search, filtering, and tailored suggestions, should be included in the UI.

IV. Data Collection

TripAdvisor, the leading source of hotel ratings and travel information, was deliberately scraped to carefully compile the dataset. HTML pages that mirrored the format of hotel listings on the network were methodically created by using web scraping techniques. Once the HTML pages were parsed, important information including hotel characteristics, geographic locations, review titles, content, and traveler categories, are extracted by the BeautifulSoup library. The dataset contains nearly 2,000 rows, each row representing a unique review from 20 different hotels. This dataset comprises features like the name of the hotel, location, traveler type, and finally, customer review, which which can be found in every row of the dataset. The Dataset can be accessed using a Github link (Singh 2024). This vast data repository serves as a solid basis for further analysis and the creation of a cutting-edge hotel recommendation system. Through utilizing the knowledge gained from this meticulously selected dataset, scholars may explore trends in consumer inclinations, attitudes, and journey practices, eventually enabling the development of customized hotel suggestions that enhance the customers' entire trip-planning experience.

V. Data Preprocessing

Thorough pre-processing of the raw text data was necessary to guarantee consistency and remove unnecessary parts before analysis could begin. To clean up and standardize the textual material, this required a methodical process with many phases. In order to eliminate online scraping artifacts, HTML elements and URLs were removed. Then, to reduce noise, frequent stop words were removed, and superfluous punctuation marks were removed. To preserve grammatical coherence, emojis and reviews written in languages other than English were methodically removed, and throughout the corpus, homogeneity was guaranteed by changing all text to lowercase. To finish the thorough pre-processing routine required to optimize the dataset for further analysis, stemming was used to minimize word variants.

By diligently carrying out these preprocessing procedures, the raw text data was thoroughly refined, making it appropriate for additional analysis and modeling efforts. This thorough preprocessing method created a strong basis for carrying out reliable and perceptive evaluations, which in turn made it easier to create a hotel recommender system that is both extremely accurate and highly effective.

VI. Exploratory data analysis (EDA) and Hypotheses for the Study

The sentiment analysis chart for hotel reviews shows that most of the reviews are extremely positive, with the majority scoring a perfect 1.0. This suggests that either the hotels are doing a great job, or that the reviews are mostly written by very satisfied customers. In terms of travel type, the data is evenly distributed with each category—Business, Friends, Couple, Family, and Solo—making up exactly 20% of the reviews. This shows that the dataset fairly represents different types of travelers.

Fig. 1. Number of reviews per Hotel

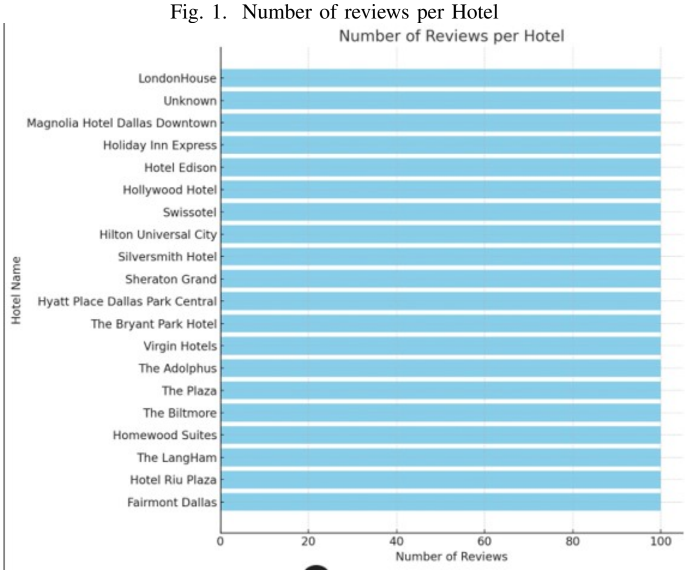
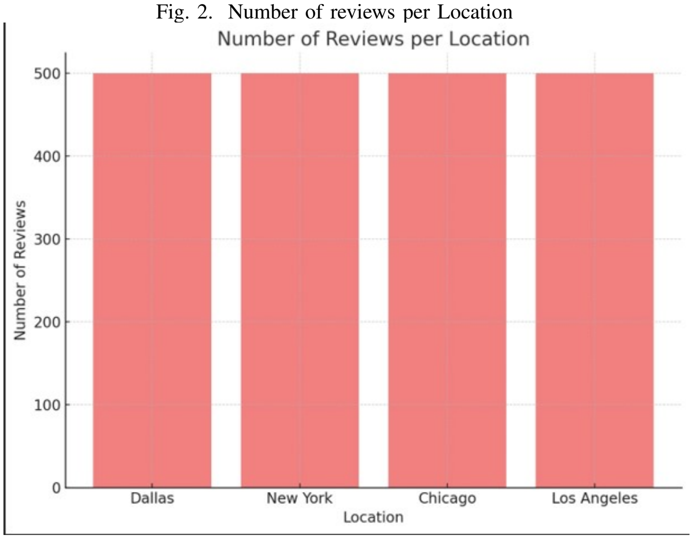
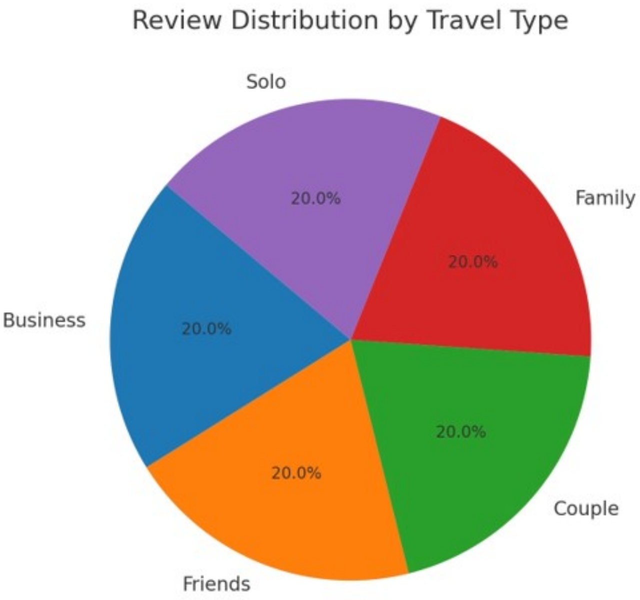


Fig. 2. Number of reviews per Location



Additionally, each hotel and each of the four major cities—Dallas, New York, Chicago, and Los Angeles—has an equal number of reviews, with 100 and 500 reviews respectively. This uniform distribution ensures that no single hotel or city influences the analysis more than any other, which helps prevent any geographical or property-specific bias in the system's recommendations. This balanced approach supports the hypothesis that clustering hotels based on review similarity can lead to more personalized and satisfying recommendations for users.

Fig. 3. Review Distribution by Travel Type

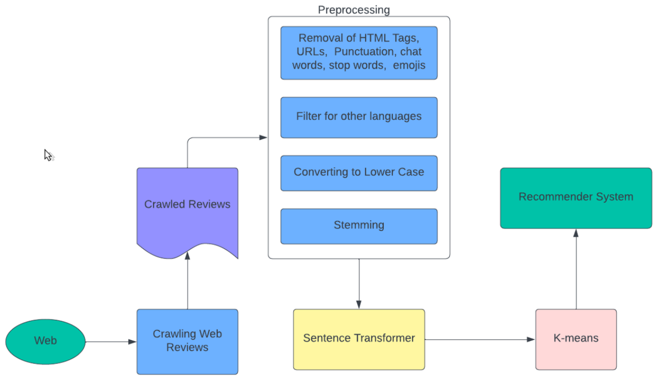


VII. Model Description

Before the Sentence Transformer we used, there are several major points to be considered since the development of NLP. RNNs were the first technology to solve the sequence-to-sequence task, like translating complete sentences from one language to another. However, RNNs were slow to train and had trouble capturing the long-range dependencies in the text, often losing the "context" of the longer sentence. Transformers replaced the RNN with more efficient architecture featuring an "attention" mechanism allowing the model to "pay attention" to different parts of the input sequence dynamically, rather than fixating on the local context. Transformers are both faster and more capable of understanding the context of a language.

Based on transformers, BERT uses stacks of transformer encoders to read words in a sentence related to all other words in a sentence and therefore profound understanding of context. It was not, however, designed to understand sentence-level similarities but in improving the model's overall performance in NLP works. Therefore, to address this gap, there are technologies that adapt BERT for high-quality sentence embeddings. In summary, sentence transformers essentially adjust the standard BERT transformer network to learn the language model in a way that makes embeddings suitable for measuring language similarity, and similarity between two sentences. In this project, we specifically used an all-MiniLM-

Fig. 4. Model Design

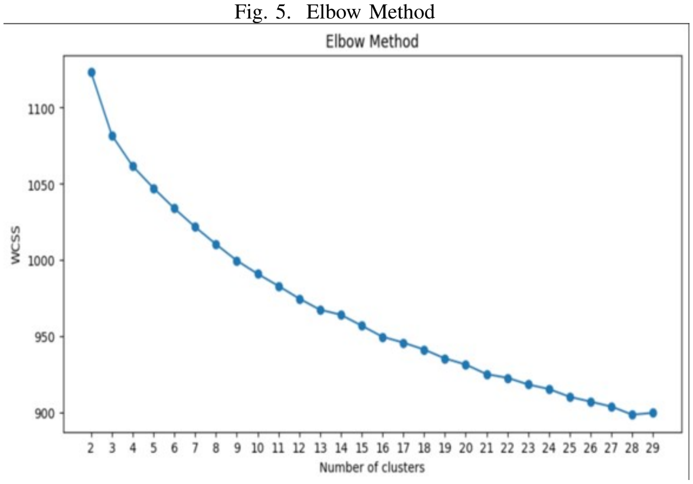


L6-v2 sentence transformer tailored for efficient performance on semantic similarity tasks. This model utilizes the MiniLM architecture, which is a lighter and faster version of BERT with fewer parameters but maintains comparable performance. The ”L6” indicates that it has 6 layers, making it less complex and quicker at processing while still capturing a deep understanding of the text.

It is specifically optimized for creating sentence and paragraph-level embeddings. This model transforms text into a 384-dimensional vector space, ideal for tasks requiring a nuanced understanding of language such as clustering, semantic search, and direct sentence comparison.

This model was fine-tuned using a self-supervised contrastive learning approach on over 1 billion sentence pairs, enhancing its ability to discern and encode semantic similarities accurately. This training included diverse datasets, ensuring that the model is robust and adaptable across various contexts.

Fig. 5. Elbow Method



VIII. Data Visualization and Results Report

In our project, we employed K-means clustering to organize the hotel reviews into groups based on their semantic similarities, which were visualized through detailed cluster charts. Determining the optimal number of clusters was a critical step, for which we utilized the Elbow Method and Silhouette Scores. These methods helped us assess the cohesion and separation of clusters. Despite the challenges in pinpointing the perfect number of clusters, we chose to set k=3 for practical implementation, ensuring that the segmentation was meaningful. To demonstrate our findings and make the recommender system user-friendly, we developed a basic web application using Flask. Structured under the FlaskApp directory, the app features a simple yet effective interface that allows users to input their preferences and receive hotel recommendations based on the cluster analysis.

Fig. 6. SilhouetteScore

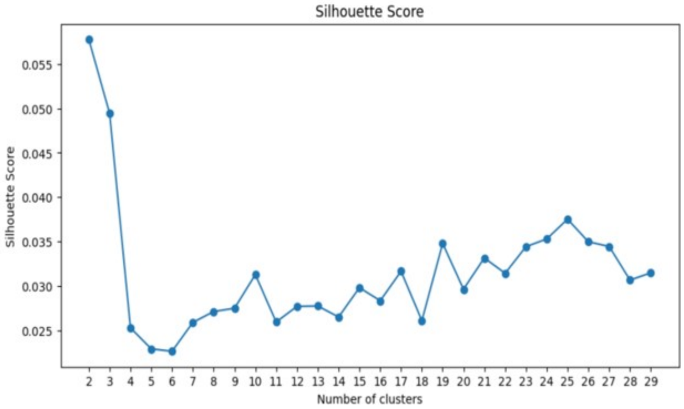
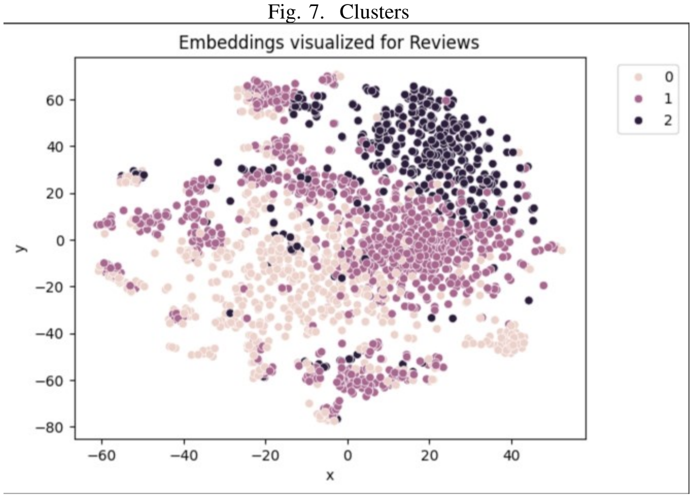


Fig. 7. Clusters



IX. Conclusion

In conclusion, our study demonstrates how machine learning and natural language processing (NLP) may improve hotel selection. We developed a sophisticated recommender system by utilizing clustering methods and doing a comprehensive study of textual hotel evaluations. Our technology, which is built on the MiniLM-L6-v2 sentence transformer architecture, improves the customer experience by providing tailored recommendations based on semantic similarities. The need of utilizing cutting-edge technology to meet the changing demands of travelers and industry stakeholders is highlighted by these findings. In the future, these recommendation systems will play a major role in boosting customer happiness and financial success in the hotel sector.

References

[1] Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). Aspect-based sentiment-oriented summarization of hotel reviews. Procedia Computer Science, 115, 563–571.

[2] Chang, Y.-C., Ku, C.-H., & Chen, C.-H. (2020). Using deep learning and visual analytics to explore hotel reviews and responses. Tourism Management, 80, 104129.

[3] Raut, V. B., & Londhe, D. D. (2014). Opinion mining and summarization of hotel reviews. In 2014 International Conference on Computational Intelligence and Communication Networks, (pp. 556–559).

[4] Ray, B., Garain, A., & Sarkar, R. (2021). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. Applied Soft Computing, 98, 106935.

[5] Singh, B. (2024, April). Hotel Recommendation System. GitHub. Retrieved from https://github.com/bakhtsingh/hotel-recommendation-system

[6] Code Emporium. (2022, February 28). Sentence Transformers - EXPLAINED! [Video]. YouTube. https://youtu.be/O3xbVmpdJwU.

[7] Sentence Transformers - all-MiniLM-L6-v2. (n.d.). Hugging Face. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[8] Shi, H. X., & Li, X. J. (2011, July). A sentiment analysis model for hotel reviews based on supervised learning. In 2011 International Conference on Machine Learning and Cybernetics (Vol. 3, pp. 950-954). IEEE.

[9] Muhammad, P. F., Kusumaningrum, R., & Wibowo, A. (2021). Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews. Procedia Computer Science, 179, 728-735.

[10] Tran, T., Ba, H., & Huynh, V. N. (2019). Measuring hotel review sentiment: An aspect-based sentiment analysis approach. In Integrated Uncertainty in Knowledge Modelling and Decision Making: 7th International Symposium, IUKM 2019, Nara, Japan, March 27–29, 2019, Proceedings 7 (pp. 393-405). Springer International Publishing.

[11] Sharma, Y., Bhatt, J., & Magon, R. (2015, October). A multi-criteria review-based hotel recommendation system. In 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (pp. 687-691). IEEE.

[12] Kaya, B. (2020). A hotel recommendation system based on customer location: a link prediction approach. Multimedia Tools and Applications, 79(3), 1745-1758.

[13] Wahyudi, K., Latupapua, J., Chandra, R., & Girsang, A. S. (2020, March). Hotel content-based recommendation system. In Journal of Physics: Conference Series (Vol. 1485, No. 1, p. 012017). IOP Publishing.