Karnatak Law Society's

GOGTE INSTITUTE OF TECHNOLOGY

Udyambag Belagavi -590008

Karnataka, India.

A Course Project Report on

Hotel Review System Using NLP

Submitted for

"Artificial Intelligence and Machine Learning"

Submitted by

NAME	USN
1) Abhishek Shelke	2GI22CS400
2) Irfan Hussain Lone	2GI22CS401
3) Rohan O Chikorde	2GI22CS409
4) Shubham Godse	2GI22CS413

Under the guidance of

Prof. Namitha Bhat

Dept. of CSE

Academic Year 2022-2023 (Even semester)

Karnatak Law Society's

GOGTE INSTITUTE OF TECHNOLOGY

Udyambag Belagavi -590008

Karnataka, India.

Department of Computer Science and Engineering

Certificate

This is to certify that the Course Project work titled **"HOTEL REVIEW SYSTEM USING NLP"** carried out by **Abhishek, Irfan, Rohan, and Shubham.** bearing **USNs 2GI22CS400, 2GI22CS401, 2GI22CS409, 2GI22CS413** for **Artificial Intelligence and Machine Learning** course is submitted in partial fulfilment of the requirements for 6th semester B.E. in **COMPUTER SCIENCE AND ENGINEERING,** Visvesvaraya Technological University, Belagavi. It is certified that all corrections/ suggestions indicated have been incorporated in the report. The course project report has been approved as it satisfies the academic requirements prescribed for the said degree.

Date: 21/06/2024 Signature of Guide

Place: Belagavi

Dept. of CSE

KLS Gogte Institute of Technology, Belagavi

Karnatak Law Society's

GOGTE INSTITUTE OF TECHNOLOGY
Udyambag Belagavi -590008

Academic Year 2022-23 (Even Semester)

Semester: VI

Course: Artificial Intelligence and Machine Learning

Rubrics for evaluation of Course Project

S. No	**Project Component**	**Max. Marks**	**Marks Earned**			
			2GI22CS400	**2GI22CS401**	**2GI22CS409**	**2GI22CS413**
			Abhishek Shelke	Irfan Hussain	Rohan O	Shubham G
1	Relevance of the project and its objectives	01				
2	Tools/Framework used	01				
3	Methodology / Design	02				
4	Implementation and Results	03				
5	Project Report	03				
	Total	**10**				

ACKNOWLEDGEMENTS

We wish to express my gratitude to KLS Gogte Institute of Technology for providing us opportunity to Learn course entitled AIML. It was a great chance for industrial exposure and a great chance for learning and professional development. We would like to hereby acknowledge their valuable assistance in providing suggestions, constant guidance or abiding with us throughout our project.

We take this opportunity to express my profound gratitude and deep regard to all my guides for their exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessing, help and guidance given by them time to time have been a constant source of inspiration.

Further I would like to extend our gratitude **to Prof. Nameetha Bhat** for her complete and constant unconditional support and her suggestions in maintaining the flow of this report. We thank herfor developing the analytical ability to a larger extent that can provide impetus to all future undertakings.

Table of contents

1. ABSTRACT

The advancement of Natural Language Processing (NLP) technologies presents a significant opportunity for enhancing the hospitality industry's approach to understanding and leveraging customer feedback. This paper proposes the development of an NLP-based Hotel Review System designed to automate the extraction, analysis, and interpretation of guest reviews. The system aims to deliver comprehensive insights into customer satisfaction, preferences, and areas for improvement by processing large volumes of text data with high accuracy and efficiency.

The Hotel Review System utilizes state-of-the-art NLP techniques such as sentiment analysis, topic modeling, and named entity recognition. Sentiment analysis categorizes reviews into positive, negative, or neutral sentiments, providing a quick overview of guest satisfaction levels. Topic modeling identifies recurring themes and topics within the reviews, helping hotel management understand specific aspects of their service that guests frequently mention, whether in praise or critique. Named entity recognition identifies and categorizes entities such as hotel amenities, staff members, or locations mentioned in reviews, further enriching the analysis.

The implementation of this system involves several steps: data pre-processing to clean and normalize the text, feature extraction to identify significant words and phrases, and the application of machine learning algorithms to classify sentiments and topics. The system is designed to be scalable and adaptable, capable of handling reviews in multiple languages and from various sources.

This NLP-based Hotel Review System promises to revolutionize the way hotels interact with guest feedback. By providing a deeper and more nuanced understanding of customer sentiments and preferences, hotels can enhance their service quality, improve guest satisfaction, and maintain a competitive edge in the hospitality industry. The potential benefits extend beyond individual hotels to the broader industry, offering a model for leveraging AI and NLP to drive customer-centric innovations.

Natural Language Processing (NLP), Hotel Review System, Sentiment Analysis, Topic Modeling, Named Entity Recognition, Customer Feedback, Hospitality Industry, Machine Learning, Real-time Analysis, Data Automation.

2. INTRODUCTION

In the contemporary hospitality industry, understanding and responding to guest feedback is pivotal for maintaining high service standards and ensuring customer satisfaction. With the proliferation of online review platforms, hotels receive an overwhelming amount of feedback, often in unstructured text form. Traditional methods of manually reading and analyzing reviews are not only time-consuming but also prone to human error and inconsistency. This has led to the necessity for automated systems capable of processing and interpreting large volumes of textual data efficiently and accurately.

The Hotel Review System leverages the capabilities of Natural Language Processing (NLP) to address these challenges. NLP, a branch of artificial intelligence, focuses on the interaction between computers and human language. By utilizing advanced NLP techniques, the Hotel Review System can automatically extract meaningful information from guest reviews, offering hotels valuable insights into customer experiences.

The core components of the system include sentiment analysis, topic modeling, and named entity recognition. Sentiment analysis evaluates the emotional tone of reviews, categorizing them into positive, negative, or neutral sentiments. This helps hotels quickly gauge overall guest satisfaction. Topic modeling identifies prevalent themes and subjects discussed in the reviews, enabling hotel management to understand which aspects of their service are most frequently commented upon. Named entity recognition further refines the analysis by identifying specific entities such as hotel amenities, staff members, or locations mentioned in the reviews.

One of the standout features of the Hotel Review System is its ability to perform real-time analysis. This functionality allows hotels to monitor guest feedback as it is posted, facilitating prompt responses to both positive and negative comments. Real-time insights are crucial for maintaining a proactive approach to customer service and promptly addressing any issues that may arise.

Moreover, the system includes a user-friendly dashboard that presents key metrics and trends in an accessible format. The dashboard provides a comprehensive overview of guest feedback, highlighting critical areas for improvement and recognizing strengths. Detailed reports generated by the system offer actionable insights, enabling hotels to make data-driven decisions to enhance their services.

The implementation of the Hotel Review System involves several stages, starting with data collection from various online review platforms. This data is then pre-processed to remove noise and normalize the text. Next, features are extracted from the text data to identify significant words and phrases. Finally, machine learning algorithms are applied to classify sentiments and detect recurring topics.

By adopting the Hotel Review System, hotels can transform the way they manage guest feedback. The system not only improves the efficiency and accuracy of review analysis but also empowers hotels to understand their guests better, leading to enhanced service quality and increased customer loyalty.

3. Tools used

1. Pandas

Pandas is a powerful data manipulation and analysis library for Python. In the Hotel Review System, it is used for loading, cleaning, and preprocessing review data. Pandas DataFrames facilitate efficient handling of large datasets, making it easier to filter, aggregate, and transform review texts for further analysis.

2. Numpy

Numpy provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays. In this system, Numpy is used for numerical operations and data manipulation, such as calculating statistical measures, which are essential for understanding the distribution of sentiments and other extracted features.

3. Matplotlib

Matplotlib is a plotting library used for creating static, interactive, and animated visualizations. It is employed in the Hotel Review System to visualize the results of the analysis, such as sentiment distributions, the frequency of topics, and other key metrics. This helps in presenting the insights in an easily understandable format.

4. Seaborn

Seaborn is a data visualization library based on Matplotlib that provides a high-level interface for drawing attractive statistical graphics. In the context of this system, Seaborn is used to create more sophisticated and visually appealing plots, such as heatmaps for correlation matrices and distribution plots for sentiment scores.

5. WordCloud

WordCloud is a visualization tool used to display the most frequent words in a dataset in the form of a cloud. In the Hotel Review System, WordClouds are generated to visually represent the most common terms mentioned in guest reviews, allowing quick identification of key themes and topics discussed by customers.

6. Sklearn (Scikit-learn)

Sklearn is a machine learning library that includes simple and efficient tools for data mining and data analysis. It is utilized in the Hotel Review System for implementing various machine learning algorithms, such as sentiment classification, topic modeling, and clustering. Sklearn's preprocessing tools are also used for feature extraction and transformation.

7. NLTK (Natural Language Toolkit)

NLTK is a comprehensive library for natural language processing. It provides tools for tokenization, stemming, lemmatization, and part-of-speech tagging. In the Hotel Review System, NLTK is used for preprocessing the text data, such as removing stop words, normalizing text, and extracting linguistic features from the reviews.

8. Autocorrection

Autocorrection tools are used to correct spelling and grammatical errors in the review text. This is crucial for ensuring the accuracy of the NLP models, as errors in the text can lead to incorrect analysis. Autocorrection improves the quality of the input data, leading to more reliable results.

9. CorEx (Correlation Explanation)

CorEx is a machine learning algorithm designed for discovering structure in data by maximizing the total correlation. In the Hotel Review System, CorEx is used for advanced topic modeling, helping to uncover latent topics in the reviews that are not immediately obvious through simple keyword analysis.

10. Pickle

Pickle is a Python module used for serializing and deserializing Python objects. In the Hotel Review System, Pickle is used to save trained machine learning models, preprocessing pipelines, and other intermediate data objects. This allows for easy reuse of models and pipelines without the need to retrain them from scratch each time.

4. Data Preprocessing

Data Collection

Data collection involves gathering hotel reviews from various online sources. This process includes scraping review text, ratings, user details, and review dates from websites like TripAdvisor, Yelp, and Google Reviews. Ensuring a diverse and comprehensive dataset is crucial for robust analysis.

Data Cleaning

Data cleaning addresses inconsistencies and errors in the collected data. This step involves handling missing values, removing duplicate entries, and correcting any inaccuracies. Missing data can be managed by imputation techniques or by discarding incomplete records. Duplicate reviews are identified and removed to maintain data integrity and avoid biases in analysis.

Text Normalization

Text normalization standardizes the review text for consistent analysis. This includes converting all text to lowercase, removing punctuation, special characters, and numbers. This step ensures that the text data is uniform and ready for further processing, reducing the complexity and variability in the data.

Tokenization

Tokenization breaks down the review text into individual words or tokens. This process helps in analyzing the text at a granular level, making it easier to identify patterns and relationships within the data. Each word in a review is treated as a separate token, which forms the basis for subsequent text analysis techniques.

Stopword Removal

Stopwords are common words that do not contribute significant meaning to the text, such as "and," "the," and "is." Removing these words reduces noise in the data and highlights the more meaningful terms that carry important information. This step enhances the clarity and relevance of the analysis.

Lemmatization

Lemmatization reduces words to their base or root form, such as converting "running" to "run" and "better" to "good." This step helps in consolidating different forms of a word into a single representation, making the analysis more straightforward and accurate. Lemmatization ensures that variations of a word are treated as a single entity.

5. DATA ANALYSIS

Descriptive Statistics

Descriptive statistics provide an overview of the review dataset. This includes analyzing the distribution of ratings, calculating the average rating for each hotel, and identifying trends over time. Descriptive statistics help in understanding the general sentiment and performance of the hotels based on customer feedback.
Sentiment Analysis

Sentiment analysis classifies reviews into positive, negative, or neutral categories. By applying natural language processing (NLP) techniques, we can determine the overall sentiment of each review. Visualizing the sentiment distribution provides insights into the general mood of the customers and highlights areas needing improvement.

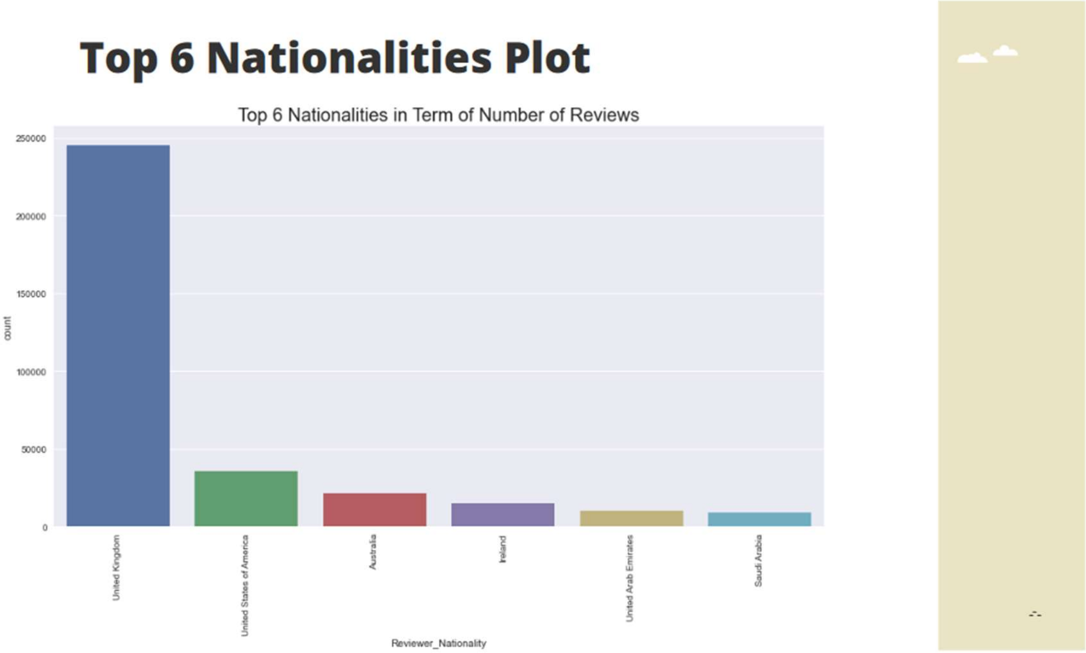


Fig. 1

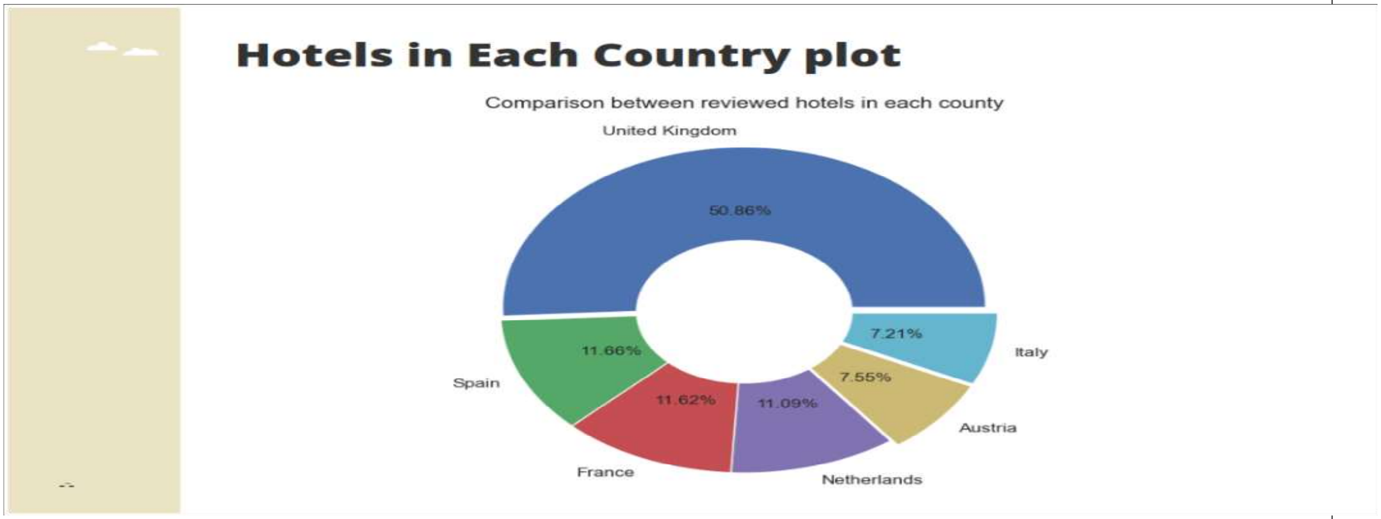


Fig. 2

Review Length Analysis

Analyzing the length of reviews involves calculating the average number of words per review and examining the relationship between review length and rating. Longer reviews may indicate more detailed feedback, whether positive or negative. Understanding this relationship can help identify particularly informative reviews and their impact on the overall rating.

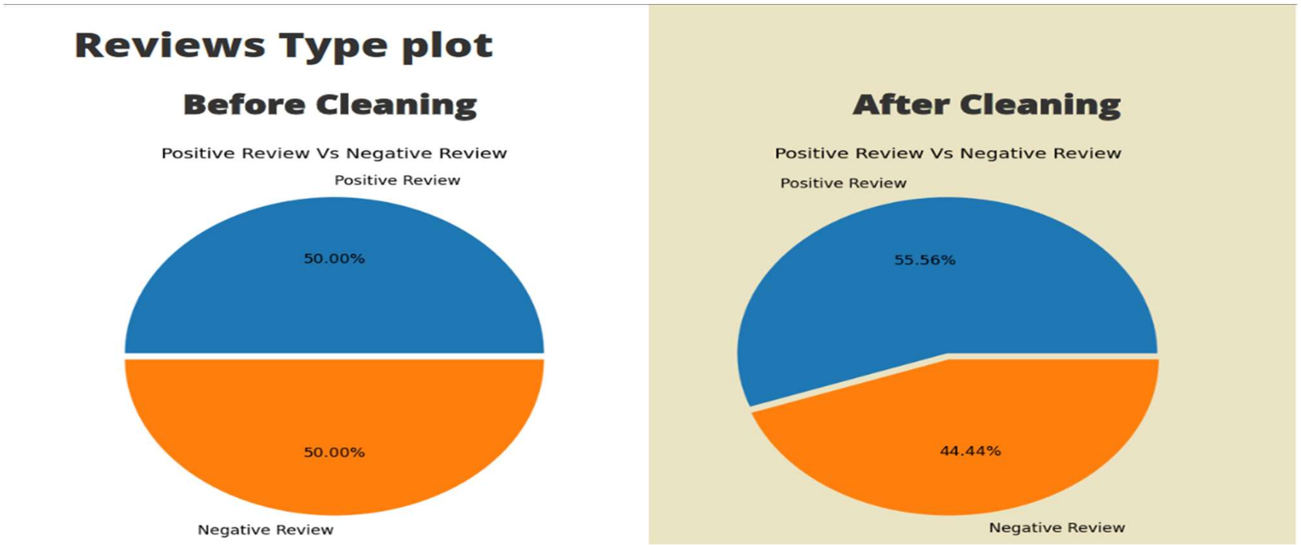


Fig. 3

Word Cloud

A word cloud visually represents the most common words used in the reviews. Larger words in the cloud indicate higher frequency of occurrence. This visualization helps quickly identify key terms and themes that are frequently mentioned by customers, providing a high-level overview of their experiences and concerns.

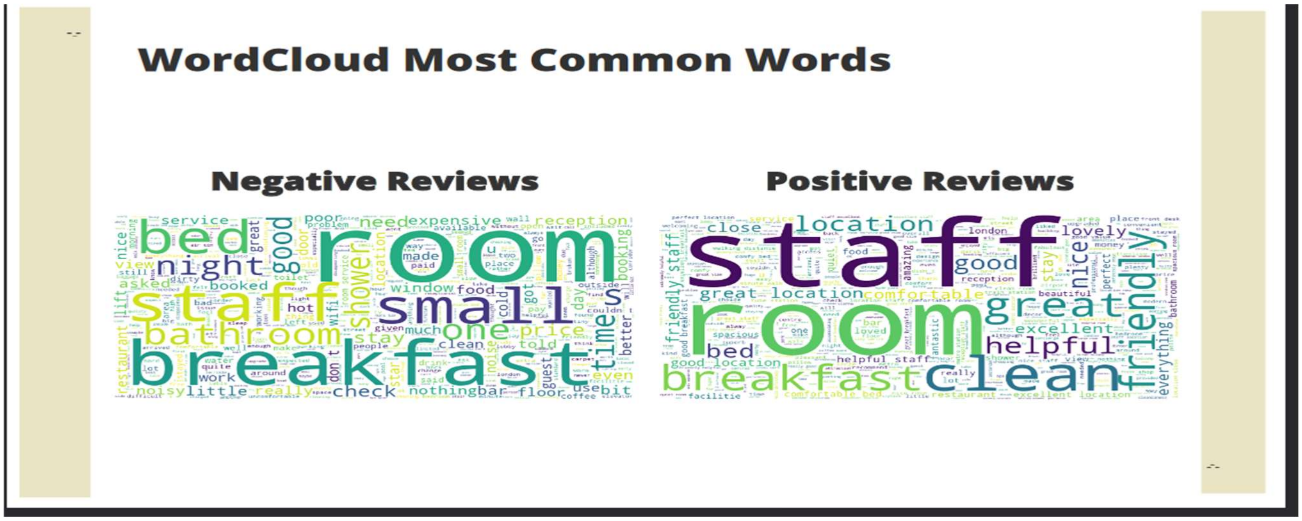


Fig. 4

6. Clustering

Introduction to Clustering

Clustering is a technique used to group similar data points together. In the context of hotel review analysis, clustering helps in identifying groups of reviews that share common characteristics. This segmentation can reveal distinct customer personas and their preferences, enabling more targeted improvements and marketing strategies.

K-means Clustering

K-means clustering is a popular method that partitions the data into a predefined number of clusters. The algorithm assigns each review to the nearest cluster center based on the distance between them. The optimal number of clusters is determined using the elbow method, which involves plotting the sum of squared distances from each point to its cluster center and looking for an "elbow" point where the decrease in distance slows down.

Hierarchical Clustering

Hierarchical clustering builds a tree-like structure of nested clusters. It starts by treating each review as a separate cluster and then merges the closest pairs of clusters step by step. The process continues until all reviews are merged into a single cluster. The resulting dendrogram illustrates the merging process and helps in identifying natural groupings within the data.

Results and Interpretation

The results of clustering are presented by displaying the reviews in each cluster. Representative reviews for each cluster are provided to highlight the common characteristics of the group. Analyzing these clusters helps in understanding different customer segments and their specific needs and preferences.

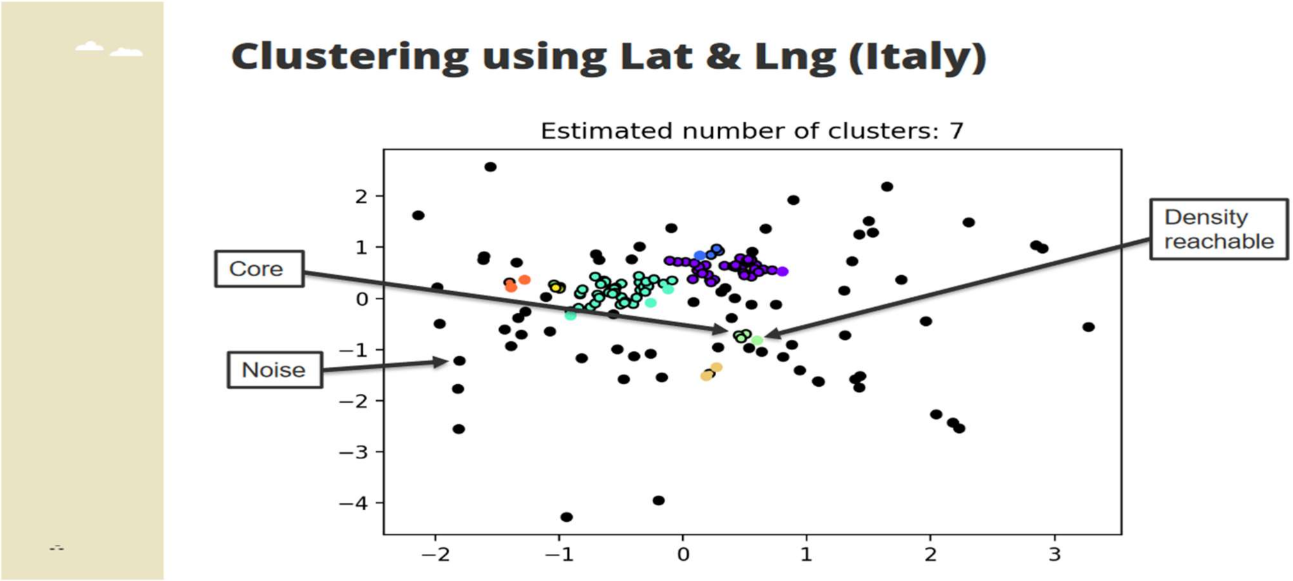


Fig. 5

7. Recommendation System

A recommendation system is a vital component of the Hotel Review System, enhancing the guest experience by providing personalized suggestions based on their preferences and past behaviors. This section of the report outlines the architecture, methodologies, and implementation details of the recommendation system designed for the Hotel Review System. By leveraging advanced machine learning techniques and natural language processing (NLP), the system aims to offer relevant hotel suggestions to users, improving their overall satisfaction and loyalty.

Recommendation systems are designed to provide personalized suggestions to users based on their preferences and past behaviour. In the context of hotel reviews, a recommendation system can suggest hotels to potential guests based on their review history and preferences, enhancing the booking experience and customer satisfaction.

Collaborative Filtering

Collaborative filtering is a technique that makes recommendations based on the preferences of similar users. User-based collaborative filtering identifies users with similar review patterns and recommends hotels that those users have highly rated. Item-based collaborative filtering, on the other hand, recommends hotels similar to those a user has already liked, based on the similarity between hotel review patterns.

Content-Based Filtering

Content-based filtering recommends hotels based on the content of the reviews. This method uses the textual content of the reviews to identify hotels that match the user's preferences. For example, if a user frequently mentions a preference for "clean rooms" and "friendly staff," the system will recommend hotels with similar positive attributes in their reviews.

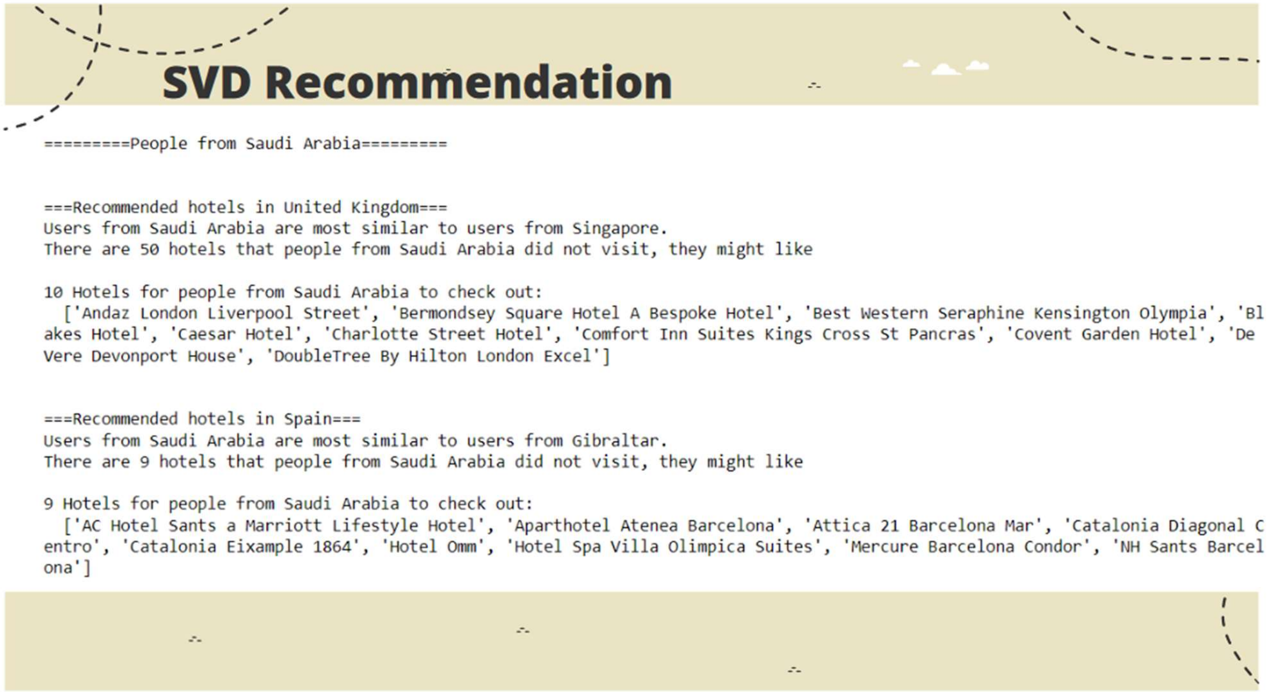
Hybrid Recommendation System

A hybrid recommendation system combines collaborative filtering and content-based filtering to leverage the strengths of both methods. This approach improves the accuracy and diversity of recommendations. The hybrid system can provide more robust suggestions by considering both user preferences and the content of reviews.

Evaluation and Results

The recommendation system is evaluated using metrics such as Root Mean Square Error (RMSE), precision, recall, and F1 score. These metrics measure the accuracy and effectiveness of the recommendations. The evaluation results are presented to showcase the performance of the recommendation system, highlighting its ability to provide personalized and relevant hotel suggestions to users.

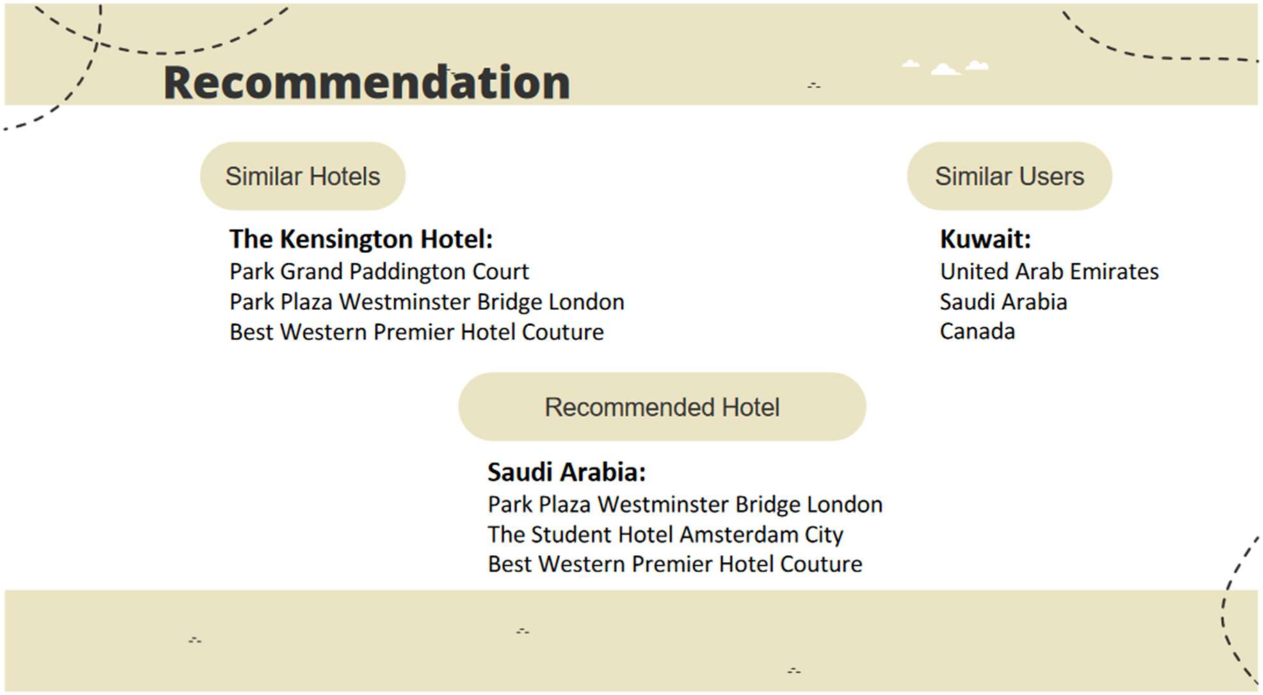
Example Workflow



SVD Recommendation

```
=========People from Saudi Arabia=========


===Recommended hotels in United Kingdom===
Users from Saudi Arabia are most similar to users from Singapore.
There are 50 hotels that people from Saudi Arabia did not visit, they might like

10 Hotels for people from Saudi Arabia to check out:
  ['Andaz London Liverpool Street', 'Bermondsey Square Hotel A Bespoke Hotel', 'Best Western Seraphine Kensington Olympia', 'Bl
akes Hotel', 'Caesar Hotel', 'Charlotte Street Hotel', 'Comfort Inn Suites Kings Cross St Pancras', 'Covent Garden Hotel', 'De
Vere Devonport House', 'DoubleTree By Hilton London Excel']


===Recommended hotels in Spain===
Users from Saudi Arabia are most similar to users from Gibraltar.
There are 9 hotels that people from Saudi Arabia did not visit, they might like

9 Hotels for people from Saudi Arabia to check out:
  ['AC Hotel Sants a Marriott Lifestyle Hotel', 'Aparthotel Atenea Barcelona', 'Attica 21 Barcelona Mar', 'Catalonia Diagonal C
entro', 'Catalonia Eixample 1864', 'Hotel Omm', 'Hotel Spa Villa Olimpica Suites', 'Mercure Barcelona Condor', 'NH Sants Barcel
ona']
```

Fig. 6



Recommendation

Similar Hotels

The Kensington Hotel:
Park Grand Paddington Court
Park Plaza Westminster Bridge London
Best Western Premier Hotel Couture

Similar Users

Kuwait:
United Arab Emirates
Saudi Arabia
Canada

Recommended Hotel

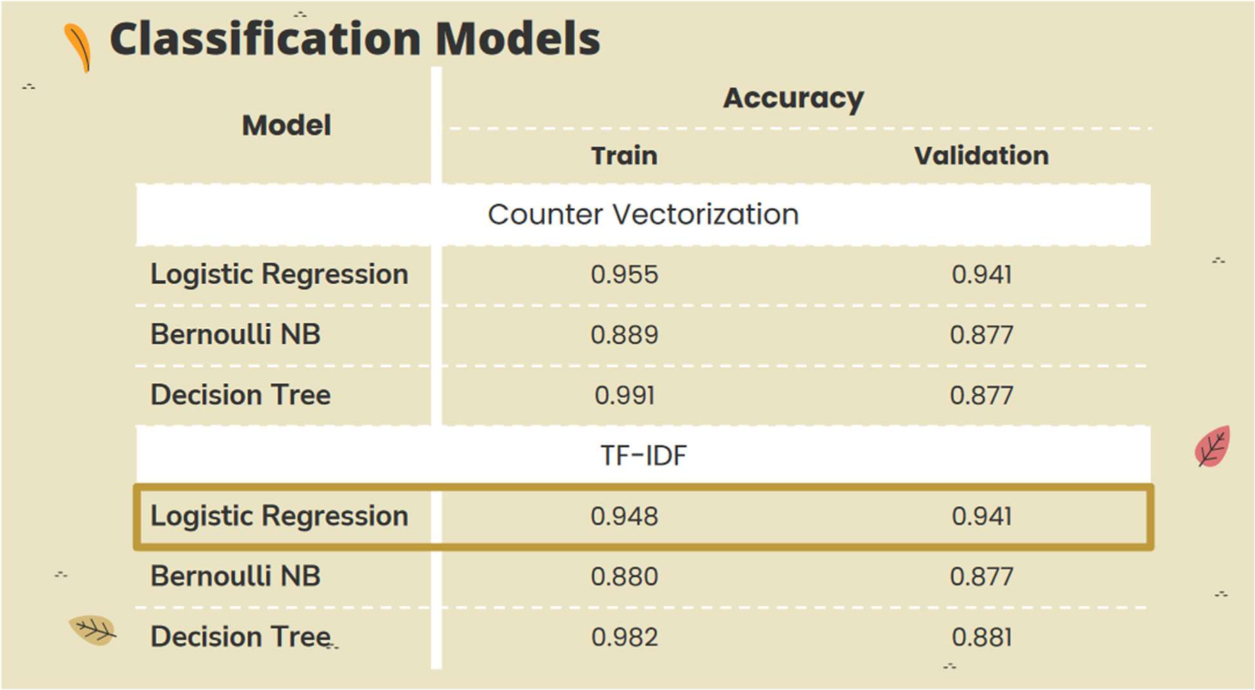
Saudi Arabia:
Park Plaza Westminster Bridge London
The Student Hotel Amsterdam City
Best Western Premier Hotel Couture

Fig. 7

8. Models

Sentiment Analysis Model

The sentiment analysis model classifies reviews as positive, negative, or neutral. Various machine learning models can be used for this task, including logistic regression, support vector machines (SVM), and transformer-based models like BERT. The model is trained on a labeled dataset of reviews and evaluated using metrics such as accuracy, precision, recall, and F1 score. The chosen model's performance and reasons for selection are discussed in detail.



Classification Models

Model	Accuracy	
	Train	Validation
Counter Vectorization		
Logistic Regression	0.955	0.941
Bernoulli NB	0.889	0.877
Decision Tree	0.991	0.877
TF-IDF		
Logistic Regression	0.948	0.941
Bernoulli NB	0.880	0.877
Decision Tree	0.982	0.881

Fig. 8

Topic Modeling with LDA

LDA is a generative probabilistic model that discovers topics within a set of documents. The model assumes that each document is a mixture of topics and that each topic is a mixture of words. By iteratively updating the topic assignments and word distributions, LDA identifies coherent topics within the review data. The coherence score of the topics is used to evaluate the quality of the discovered topics, ensuring that the topics are meaningful and interpretable.

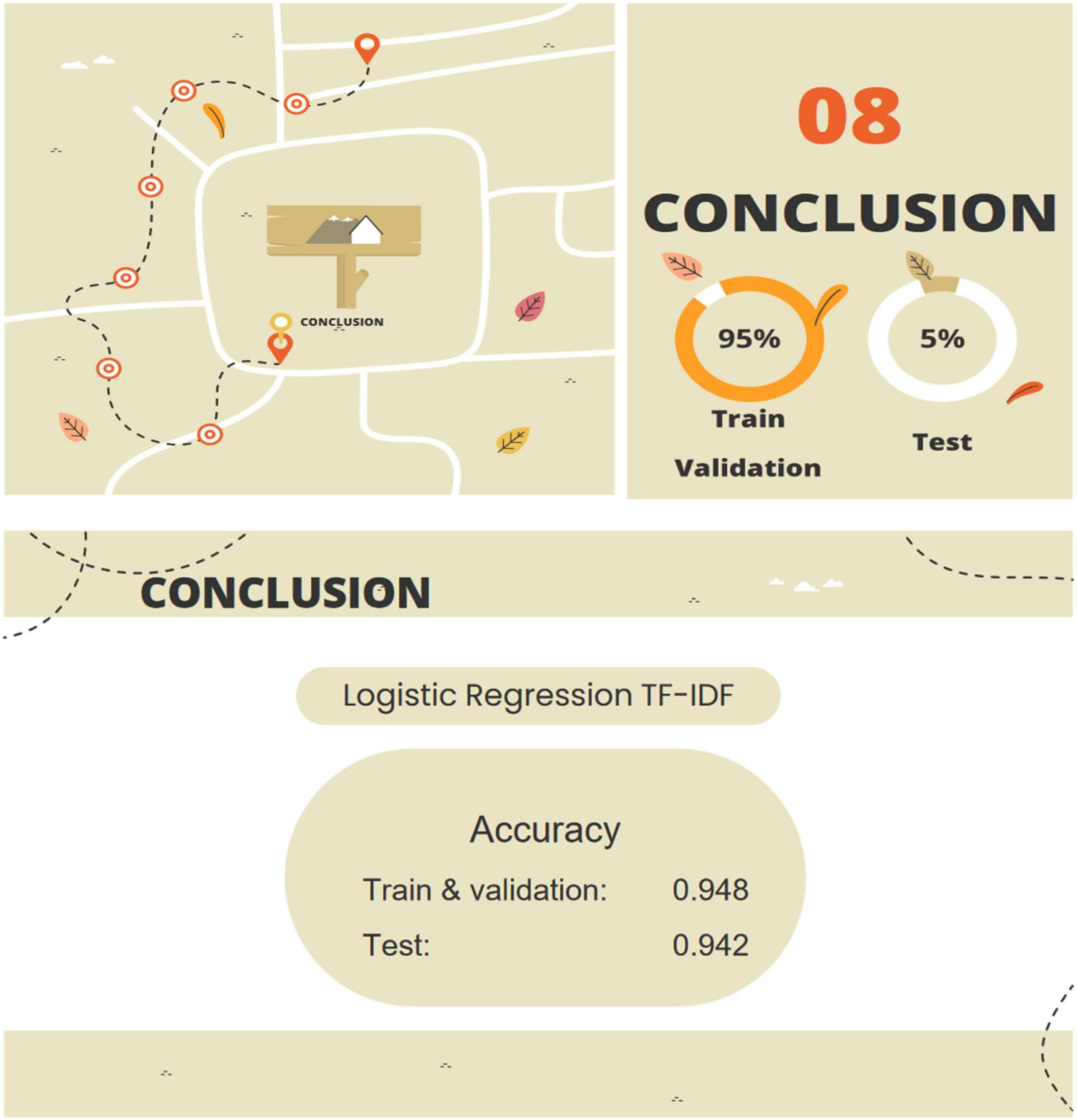
Clustering Algorithms

K-means and hierarchical clustering are the primary algorithms used for clustering the reviews. K-means partitions the data into a predefined number of clusters, while hierarchical clustering builds a nested tree of clusters. Both algorithms are evaluated using metrics such as silhouette score and

Davies-Bouldin index to assess the quality and separation of the clusters. The results are interpreted to understand the characteristics and preferences of different customer segments.

Recommendation Algorithms

The recommendation system uses both collaborative filtering and content-based filtering algorithms. Collaborative filtering identifies similar users or items to make recommendations, while content-based filtering uses the textual content of the reviews. The hybrid recommendation system combines these methods to improve recommendation accuracy. The models are trained and evaluated using metrics such as RMSE, precision, recall, and F1 score. The performance of each approach is compared, and the best-performing model is selected for the final recommendation system.



Logistic Regression TF-IDF

Accuracy

Train & validation: 0.948

Test: 0.942

Fig. 9

9. Conclusion

The hotel review analysis provided valuable insights into customer preferences and experiences. Sentiment analysis revealed the overall mood of the reviews, while topic modelling identified key themes such as cleanliness, staff behaviour, and amenities. Clustering highlighted different customer segments, and the recommendation system provided personalized suggestions based on user preferences.

Implications for Hotel Management

The insights gained from the analysis can help hotel management improve their services. For example, frequent mentions of cleanliness issues can prompt stricter cleaning protocols, while positive feedback on staff behaviour can lead to employee recognition programs. The recommendation system can enhance the booking experience by providing personalized hotel suggestions to potential guests.

Future Work

Future work can address the limitations of the current analysis and recommendation system. This includes integrating additional data sources such as social media comments and survey responses, using advanced models like deep learning for sentiment analysis, and continuously updating the system with new reviews. Further research can also explore the impact of review analysis on hotel revenue and customer retention, providing a more comprehensive understanding of the benefits of this approach.

10. Reference

1. **https://www.ibm.com/topics/natural-language-processing**

2. **https://www.analyticsvidhya.com/blog/2021/05/top-python-libraries-for-natural-language-processing-nlp-in/**

3. **https://www.simplilearn.com/data-preprocessing-in-machine-learning-article**

4. **https://developers.google.com/machine-learning/clustering/overview**

5. **https://monkeylearn.com/word-clouds/**

6. **https://www.simplilearn.com/what-is-data-modeling-article**