A Mini Project Report on

Analysis Of Restaurant Ratings And Reviews Using Machine Learning

Submitted in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

In

CSE (DATA SCIENCE)

By

Mohammed Adil (21AG1A6744)

Under the guidance of

MS. Swathi Turai

Assistant Professor

DEPARTMENT OF CSE(DATA SCIENCE)

ACE Engineering College

Ankushapur(V), Ghatkesar(M), Medchal Dist - 501301

(An Autonomous Institution, Affiliated to JNTUH, Hyderabad)

www.aceec.ac.in

2024-2025

ACE
Engineering College

UGC AUTONOMOUS INSTITUTION

(Sponsored by Yadala Satyanarayana Memorial Educational Society, Hyderabad)

Approved by AICTE & Affiliated to JNTUH

B.Tech Courses offered: CIVIL, CSE, IT, ECE, EEE & MECH, NBA Accredited Courses: CIVIL, CSE, ECE, EEE & MECH, Accorded NAAC A - Grade

DEPARTMENT OF CSE (DATA SCIENCE)

CERTIFICATE

This is to certify that the mini project report entitled "***Analysis of Restaurant Ratings and Reviews Using Machine Learning***" is a Bonafide work done by Mohammed Adil (21AG1A6744) in partial fulfilment for the award of Degree of BACHELOR OF TECHNOLOGY in ***CSE (Data Science)*** from JNTUH University, Hyderabad during the academic year 2024- 2025. This record of Bonafide work carried out by them under our guidance and supervision.

The results embodied in this report have not been submitted by the student to any other University or Institution for the award of any degree or diploma.

Mrs. Swathi Turai	**Dr. P. Chiranjeevi**	**External**
Assistant Professor	Associate Professor	
Supervisor	HOD, CSE-DS	

ACKNOWLEDGEMENT

I would like to express my gratitude to all the people behind the screen who have helped me transform an idea into a real-time application. We would like to express my heart-felt gratitude to my parents without whom we would not have been privileged to achieve and fulfil my dreams.

A special thanks to our General Secretary, **Prof. Y. V. Gopala Krishna Murthy**, for having founded such an esteemed institution. Sincere thanks to our Joint Secretary **Mrs. M. Padmavathi**, for support in doing project work. I am also grateful to our beloved principal, **Dr. B. L. RAJU** for permitting us to carry out this project.

We profoundly thank **Dr. P. Chiranjeevi**, Associate Professor and Head of the Department of Computer Science and Engineering (Data Science) who has been an great source of inspiration to my work.

We extremely thank **Mr. Shaik Nagar Vali** and **Mr. P. Ashok Kumar**, Associate Professors, Project coordinators, who helped us in all the way in fulfilling of all aspects in completion of our Mini-Project.

We are very thankful to my internal guide **Ms. Swathi Turai (Assistant professor)** who has been an excellent and also given continuous support for the Completion of my project work.

The satisfaction and euphoria that accompany the successful completion of the task would be great, but incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success. In this context, we would like to thank all the other staff members, both teaching and non-teaching, who have extended their timely help and eased my task.

Mohammed Adil (21AG1A6744)

ANALYSIS OF RESTAURANT RATINGS AND REVIEWS USING MACHINE LEARNING

ABSTRACT

In today's digital world, food apps are becoming more popular because they allow users to easily browse, book, and order food from their favourite restaurants with just a few clicks. Apps like Zomato, Swiggy let users rate their dining experience and share reviews. These reviews help other customers and restaurants understand how well they are doing, but sometimes it's hard for restaurants to tell if the feedback is positive or negative. We started by exploring the data (EDA) and found the most and least expensive restaurants. We also identified top critics who have reviewed over 100 restaurants and have more than 10,000 followers, which restaurants should focus on. For grouping similar restaurants, we used clustering techniques and settled on 3 groups based on analysis. We used KMeans and Hierarchical clustering, with features like cuisine and cost. For analysing the sentiment of reviews (whether they're positive or negative), we used both supervised (with labelled data) and unsupervised (without labelled data) methods. Supervised methods included algorithms like Logistic Regression, Decision Trees, and Multinomial Naive Bayes, while unsupervised methods included techniques like Linear Discriminant Analysis. We defined ratings above 3.5 as positive and below 3.5 as negative. After tuning the models, Logistic Regression and LightGBM worked best for predicting review sentiments.

INDEX

LIST OF FIGURES

Chapter 1. Introduction

1.1 Background and Context of the Project:

The rapid growth of digital food services has revolutionized how customers interact with restaurants. Customers increasingly rely on online tools to order meals and share their dining experiences through ratings and reviews. These reviews, often written in unstructured text, contain valuable insights but are challenging for restaurant owners to interpret. Extracting actionable information from this data is critical for understanding customer sentiment, identifying areas of improvement, and staying competitive in a fast-evolving market.

This project addresses these challenges by employing advanced machine learning and natural language processing (NLP) techniques. By utilizing tools like BERT for sentiment analysis, clustering algorithms such as K-means and hierarchical clustering, and additional features like fake review detection and personalized recommendations, the project transforms unstructured feedback into meaningful insights. The goal is to help restaurant owners enhance service quality, refine offerings, and improve overall customer satisfaction, enabling them to make informed, data-driven decisions in a highly competitive industry.

1.2 Objectives

The main goal of this project is to analyze restaurant reviews to provide useful insights that help improve customer satisfaction. By understanding the sentiment of reviews, identifying common feedback themes, and tracking trends, restaurant owners can make better decisions. These insights can help refine menus, enhance service quality, and streamline operations to better meet customer needs.

Another important goal is to use advanced AI techniques to detect fake reviews. Fake reviews can mislead both customers and restaurant owners, affecting trust and decision-making. By identifying patterns like repeated content or unusual user behavior, the system ensures the feedback is reliable. Genuine reviews enable restaurant owners to make informed decisions and improve their services effectively.

1.3 Project Type

- The project type is Data Science and Machine Learning with a focus on Natural Language Processing (NLP) for sentiment analysis, clustering, and predictive modeling in the domain of restaurant reviews and ratings.

1.4 Scope of Project

1. **Improved Customer Experience**:
    - Provide personalized restaurant recommendations based on analyzed reviews.
    - Help customers make informed dining decisions.
2. **Actionable Insights for Restaurants**:
    - Enable restaurants to identify customer sentiment, trends, and areas for improvement.
    - Detect fake reviews to ensure reliable feedback for operational strategies.
3. **Enhanced Industry Practices**:
    - Establish a standardized approach to review and sentiment analysis in the food service industry.
    - Support data-driven decision-making to enhance service quality and customer satisfaction.
4. **Broader Applications**:
    - The system can be adapted for other industries, such as retail, travel, or hospitality, where customer reviews play a critical role.
5. **Scalable and Future-ready**:
    - Integration of advanced NLP techniques like BERT ensures adaptability for larger datasets and evolving customer feedback formats.
    - Potential to include real-time monitoring and analytics for proactive decision-making.

1.5 Technologies Used

1. Programming Languages & IDEs:

- Python
- Jupyter Notebook
- VSCode

2. Libraries and Frameworks:

- Data Processing & Analysis: Pandas, NumPy
- Machine Learning: Scikit-learn, XGBoost, LightGBM, Random Forest
- Deep Learning: BERT, GPT-based models (for sentiment analysis)
- Visualization: Matplotlib, Seaborn, Word Cloud

3. Clustering Techniques:

- K-Means Clustering
- Hierarchical Clustering

4. Feature Engineering & NLP:

- Tokenization and Vectorization (e.g., TF-IDF, Count Vectorizer)
- Dimensionality Reduction (PCA)

5. Operating Systems:

- Windows
- macOS
- Linux

6. Other Tools:

- TripAdvisor API for dataset integration
- Plagiarism checker for fake review detection.

Chapter 2. Literature Survey

The growing popularity of online food applications has introduced new challenges for understanding customer feedback. Sentiment analysis, clustering, and predictive modeling have been identified as crucial tools for addressing these challenges. Below is an overview of existing work relevant to the project:

2.1 Sentiment Analysis

Sentiment analysis is a well-researched field aimed at extracting opinions from text. Techniques range from classical machine learning algorithms such as Logistic Regression to deep learning models like BERT and GPT, which excel in contextual understanding of text. Research highlights that deep learning methods outperform traditional techniques in handling unstructured and noisy datasets.

2.2 Clustering Techniques

Clustering has been applied to group similar reviews, revealing patterns in customer sentiment. K-means and hierarchical clustering are commonly used algorithms. Studies demonstrate that K-means effectively clusters reviews based on preprocessed features, while hierarchical clustering provides more nuanced groupings for smaller datasets.

2.3 Topic Modeling

Topic modeling, such as Latent Dirichlet Allocation (LDA), has been widely utilized to identify frequently discussed themes in customer reviews. Studies underscore its ability to uncover latent topics, which aids restaurants in pinpointing specific areas for improvement.

2.4 Predictive Modeling

Machine learning models, including Random Forest, XGBoost, and LightGBM, have been employed to predict future restaurant ratings based on historical data. These models have shown high accuracy in leveraging structured and unstructured data to forecast customer satisfaction.

2.5 Fake Review Detection

Research into AI-powered solutions for identifying fake reviews highlights the use of neural networks combined with plagiarism-check algorithms. Such techniques improve trustworthiness by filtering out manipulative feedback that can skew restaurant ratings.

2.6 Trends and Future Directions

Recent works emphasize integrating advanced Natural Language Processing (NLP) methods, expanding datasets, and exploring real-time applications. These enhancements aim to increase model accuracy and make insights actionable for restaurant owners.

2.7 Tools and Libraries

Python libraries like pandas, scikit-learn, and TensorFlow have been extensively used for preprocessing, modeling, and evaluation in sentiment analysis and clustering studies. Visualization tools like Matplotlib and Seaborn provide actionable insights.

Chapter 3. System Analysis

3.1 Existing System

1. **Sentiment Analysis**: Identifies whether reviews are positive, negative, or neutral, revealing overall customer satisfaction.
2. **Topic Modeling**: Extracts keywords and topics from reviews to highlight areas for improvement.
3. **Clustering**: Groups similar reviews to uncover common feedback or exceptional experiences.
4. **Predictive Modeling**: Uses machine learning to predict future ratings based on review patterns.

3.2 Proposed System

1. **Machine Learning for Sentiment Analysis**:
    - Implement advanced models like BERT or GPT for understanding the context and meaning behind customer reviews more effectively.
2. **Personalized Recommendation System**:
    - Use review data to provide tailored restaurant suggestions to customers based on their preferences.
3. **Trend Monitoring**:
    - Track changes in sentiment, topics, and ratings over time to identify emerging trends in customer feedback.
4. **AI-powered Fake Review Detection**:
    - Develop an AI bot to identify fake reviews using username analysis and plagiarism detection techniques.

Chapter 4. Software Requirements

The Software Requirements Specification for the project includes the following components:

4.1 Python Libraries

The system will require several Python libraries for data manipulation, analysis, and visualization. These include:

- *Pandas* for data handling and manipulation.
- *Scikit-learn* for machine learning tasks and model building.
- *Matplotlib* and *Pyplot* for creating static, animated, and interactive visualizations.
- *Seaborn* for statistical data visualization and enhancing the aesthetic of the plots.
- *Wordcloud* for generating word clouds from text data.
- *NumPy* for numerical operations and handling large arrays efficiently.

4.2 Integrated Development Environment (IDE)

The project can be developed and tested in popular IDEs such as:

- *Jupyter Notebook* for interactive coding and data visualization.
- *VSCode* for a more traditional coding environment with support for Python.

4.3 Operating System

The project is compatible with multiple operating systems, including:

- *Windows*, *macOS*, and *Linux*, ensuring cross-platform usability and development flexibility.

These requirements ensure that the project can be implemented effectively with the necessary tools for data analysis, machine learning, and visualization.

Chapter 5. Hardware Requirements

5.1 Processor

- **Intel Core i5 or higher**. This ensures sufficient processing power for running complex computational tasks, including CNN models. A higher processor speed is recommended for faster model training and data processing.

5.2 RAM

- **Minimum 6GB of RAM**. This provides better performance when running Python scripts, Jupyter Notebook, and other software components simultaneously. Upgrading to 8GB or more is advised for handling larger datasets efficiently.

5.3 Storage

- **100GB of reespace**. This is essential for installing necessary software packages, storing datasets, and managing any additional files generated during the project. SSD storage is preferable for quicker data access and improved system responsiveness.

Chapter 6. Functional Requirements

6.1 Data Ingestion and Preprocessing

a. Collect raw data from the input source and clean, format, and prepare it for analysis.

b. Handle missing values, remove duplicates, and perform necessary transformations to ensure data consistency.

6.2 Sentiment Analysis

c. Implement sentiment analysis using **Logistic Regression** and **LightGBM** models to classify reviews as positive, negative, or neutral.

d. Compare the performance of both models to identify the most effective approach for the dataset.

6.3 Clustering of Reviews

e. Group similar reviews together using the **K-Means** clustering algorithm.

f. Analyze clusters to gain insights into customer sentiment and common themes.

6.4 Visualization of Results

g. Create graphical representations of the processed data and analysis outcomes.

h. Use charts, graphs, and interactive dashboards for better interpretation and decision-making.

Chapter 7. Non-Functional Requirements

7.1 Performance

- Ensure **high accuracy** in both sentiment classification and review clustering.
- Optimize algorithms to deliver quick and precise results, even with large datasets.
- Maintain scalability to handle increasing data volumes without compromising speed or accuracy.

7.2 Reliability

- Implement regular **data backup** mechanisms to prevent data loss.
- Ensure seamless **system updates** to maintain functionality and compatibility with evolving requirements.
- Monitor system performance with error-handling mechanisms to quickly identify and resolve issues.

7.3 Usability

- Design an **easy-to-navigate user interface** to enhance the user experience.
- Provide clear documentation and help features for users to understand the system efficiently.
- Incorporate user feedback loops for continuous improvement in usability and functionality.

Chapter 8. Methodology

8.1 Data Collection

a. Gather review datasets from reliable online sources such as review platforms and social media.

b. Ensure the dataset is diverse and representative of the target domain.

c. Validate data quality by checking for inconsistencies, duplicates, and anomalies before preprocessing.

8.2 Preprocessing

a. Clean and normalize text data by removing noise, such as special characters, stop words, and unnecessary spaces.

b. Perform tokenization, lemmatization, and handling of missing data to prepare the dataset for analysis.

c. Incorporate techniques like stemming or entity recognition to refine the data further.

8.3 Feature Engineering

a. Convert textual data into numerical representations using **TF-IDF (Term Frequency-Inverse Document Frequency)**.

b. Explore additional techniques like word embeddings (e.g., Word2Vec or GloVe) for advanced feature extraction.

c. Implement feature selection methods to reduce dimensionality and improve model efficiency.

8.4 Model Training

a. Train and evaluate sentiment analysis models, such as **Logistic Regression** and **LightGBM**, using processed data.

b. Perform hyperparameter tuning and cross-validation to optimize model performance.

c. Analyze model performance using metrics like accuracy, precision, recall, and F1-score for robust evaluation.

8.5 Deployment

a. Develop a user-friendly interface using **Streamlit** for real-time visualization of results.

b. Integrate the system with backend services to ensure smooth and responsive performance.

c. Enable interactive features, such as filters and dynamic visual updates, to enhance user engagement.

Chapter 9. System Design

9.1 System Components

The system is designed with the following primary components:

1. **Data Ingestion Module**
    - Responsible for collecting review datasets .
    - Fetches data through APIs or web scraping scripts.
2. **Data Preprocessing Module**
    - Cleans the raw data by removing noise and inconsistencies.
    - Normalizes the text to prepare it for analysis.
3. **Sentiment Analysis Module**
    - Uses supervised machine learning models to classify reviews into positive, neutral, or negative categories.
4. **Clustering Module**
    - Groups similar reviews using unsupervised learning algorithms like K-means.
    - Identifies common themes and trends in customer feedback.
5. **Visualization Module**
    - Creates interactive graphs and charts to represent the results of the analysis.
    - Includes sentiment distribution, clustering patterns, and trends over time.
6. **Recommendation Module**
    - Generates personalized restaurant recommendations based on review data.
    - Leverages sentiment and clustering insights to enhance user experience

9.2 System Workflow

Data Collection

The data collection process involves retrieving restaurant review datasets from reliable sources such as TripAdvisor, Yelp, or Kaggle. These datasets typically contain structured and unstructured information, including text reviews, ratings, and metadata like reviewer demographics and timestamps. Automated scripts or APIs are employed to fetch and consolidate data for seamless integration into the analysis pipeline.

Data Cleaning

After data collection, the cleaning process ensures that the textual and numerical data are free of inconsistencies and errors. Tasks include:

- **Removing Noise**: Eliminate irrelevant characters, symbols, and HTML tags.
- **Handling Missing Values**: Address gaps in data by employing techniques such as imputation or deletion.
- **Standardizing Formats**: Normalize text by converting to lowercase and applying tokenization.
- **Stop Words Removal**: Exclude commonly used words (e.g., "and," "the") that do not add value to the sentiment analysis.

Model Training

The core of the project involves training machine learning models for sentiment classification and clustering. Steps include:

1. **Feature Engineering**: Convert textual reviews into numerical representations using techniques like Term Frequency-Inverse Document Frequency (TF-IDF) or word embeddings.
2. **Model Selection**: Employ supervised learning models such as Logistic Regression, Random Forest, or LightGBM for sentiment analysis. For clustering, unsupervised algorithms like K-means and Hierarchical Clustering are utilized.
3. **Training and Validation**: Split the data into training and testing subsets, optimizing hyperparameters to enhance model performance.

Analysis and Recommendations

Post-training, the system performs the following:

- **Sentiment Classification**: Categorize reviews into positive, neutral, or negative sentiments.

- **Clustering**: Group reviews to identify patterns and emerging trends.
- **Insights and Reports**: Generate actionable recommendations for restaurant owners, such as identifying areas of improvement or popular menu items.

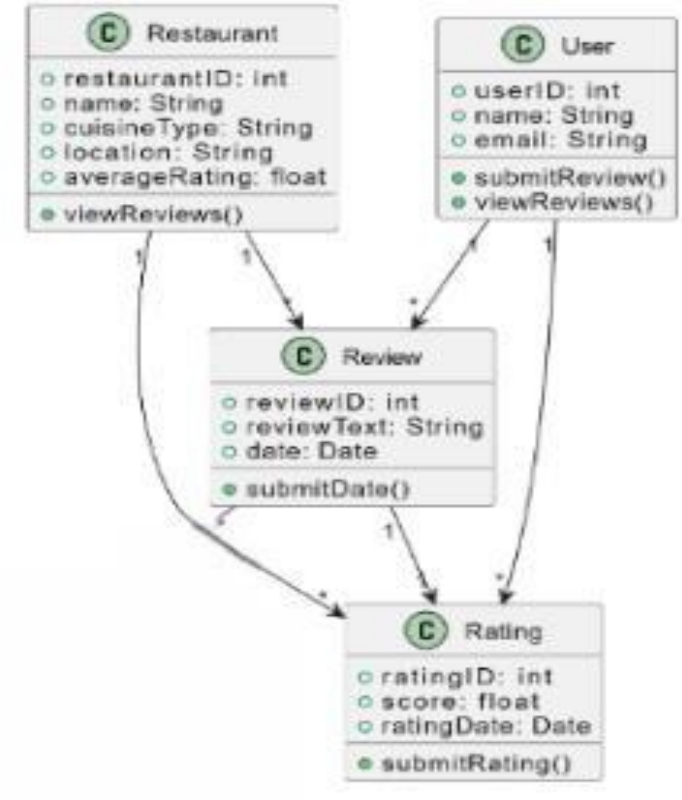
Visualization

Visualization serves as a vital tool for presenting findings. Interactive dashboards are developed using libraries like Matplotlib, Seaborn, and Plotly to:

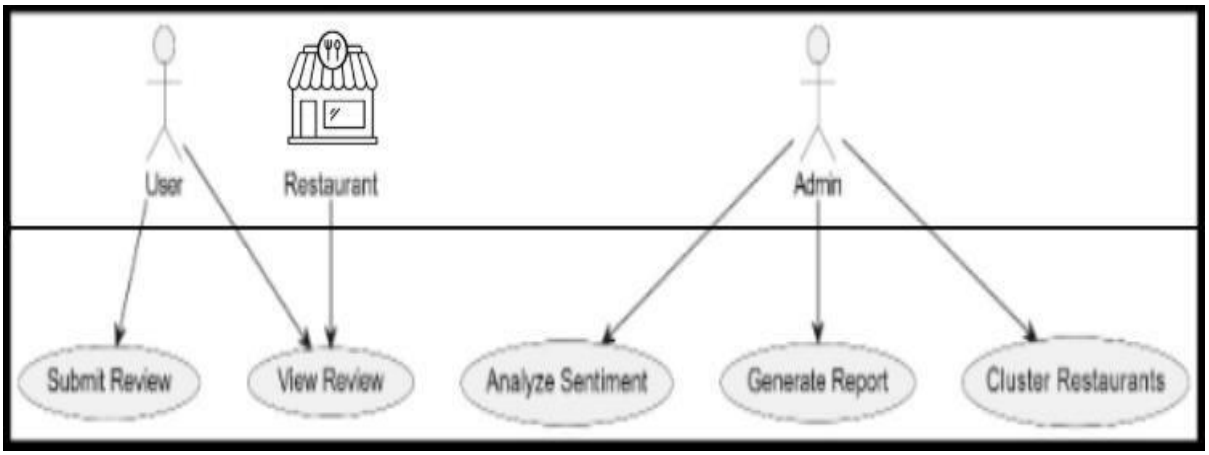
- Display sentiment distribution graphs.
- Showcase clustered review trends.
- Provide customer sentiment heatmaps. These visualizations are integrated into a user-friendly interface for ease of interpretation.

9.3 UML Diagrams

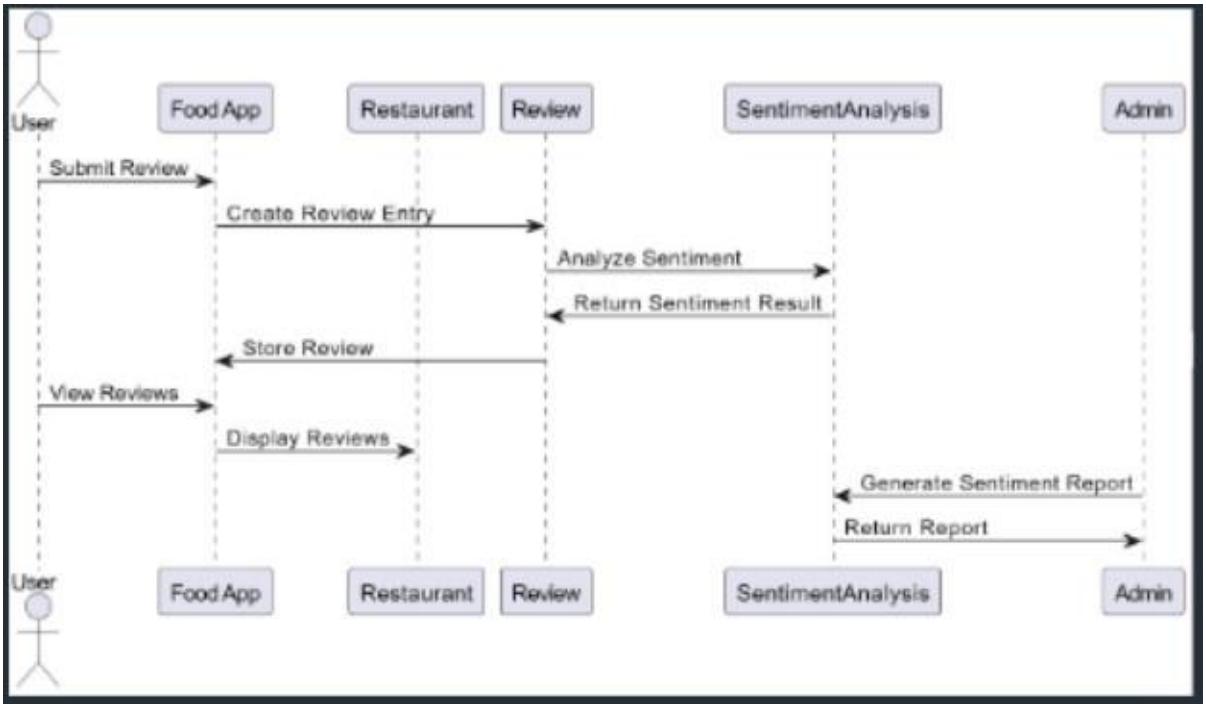
9.3.1 Class Diagram



9.3.2 Use Case Diagram



9.3.3 Sequence Diagram



Chapter 10. Implementation

10.1 Tools and Technologies

- Python
- Scikit-learn, Pandas
- Visualization libraries: Matplotlib, Seaborn

10.2 Process

- Train sentiment models on labeled data.
- Cluster reviews to identify patterns and trends.
- Develop a user-friendly dashboard.

Chapter 11. RESULTS

11.1.1 CSV Data Set of Restaurant Rating and Review

```
# loading dataset CSV file
meta_df = pd.read_csv('Restaurant names and Metadata.csv')
reviews_df = pd.read_csv('Restaurant reviews.csv')
```

Meta Data

```
# to get the first five rows of the data set
meta_df.head()
```

	Name	Links	Cost	Collections	Cuisines	Timings
0	Beyond Flavours	https://www.zomato.com/hyderabad/beyond-flavou...	800	Food Hygiene Rated Restaurants in Hyderabad, C...	Chinese, Continental, Kebab, European, South I...	12noon to 3:30pm, 6:30pm to 11:30pm (Mon-Sun)
1	Paradise	https://www.zomato.com/hyderabad/paradise-gach...	800	Hyderabad's Hottest	Biryani, North Indian, Chinese	11 AM to 11 PM
2	Flechazo	https://www.zomato.com/hyderabad/flechazo-gach...	1,300	Great Buffets, Hyderabad's Hottest	Asian, Mediterranean, North Indian, Desserts	11:30 AM to 4:30 PM, 6:30 PM to 11 PM
3	Shah Ghouse Hotel & Restaurant	https://www.zomato.com/hyderabad/shah-ghouse-h...	800	Late Night Restaurants	Biryani, North Indian, Chinese, Seafood, Bever...	12 Noon to 2 AM
4	Over The Moon Brew Company	https://www.zomato.com/hyderabad/over-the-moon...	1,200	Best Bars & Pubs, Food Hygiene Rated Restauran...	Asian, Continental, North Indian, Chinese, Med...	12noon to 11pm (Mon, Tue, Wed, Thu, Sun), 12no...

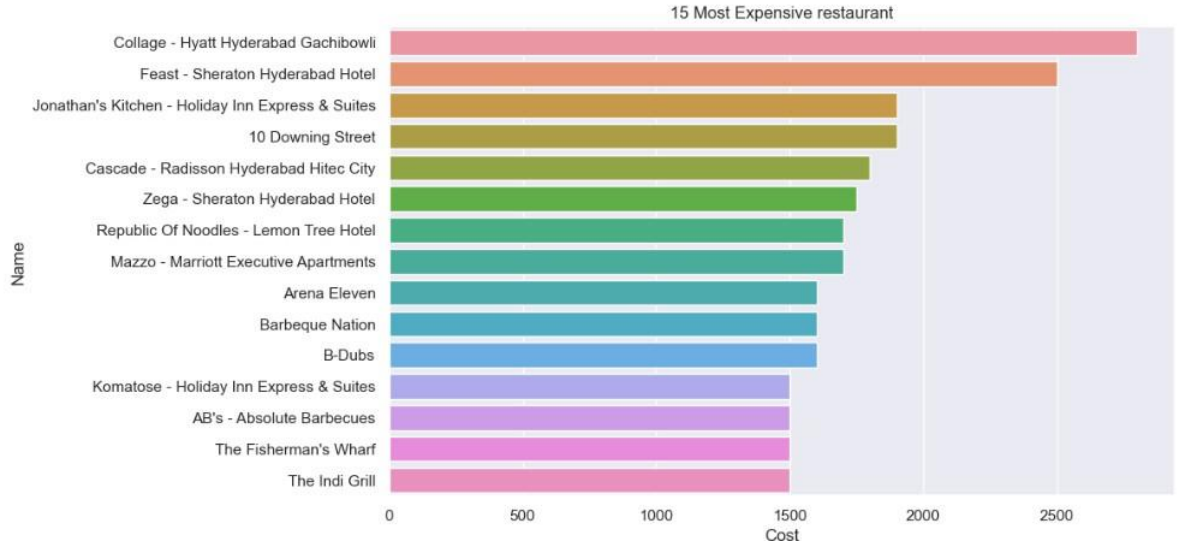
Reviews

```
# to get the first five rows of the review data set
reviews_df.head()
```

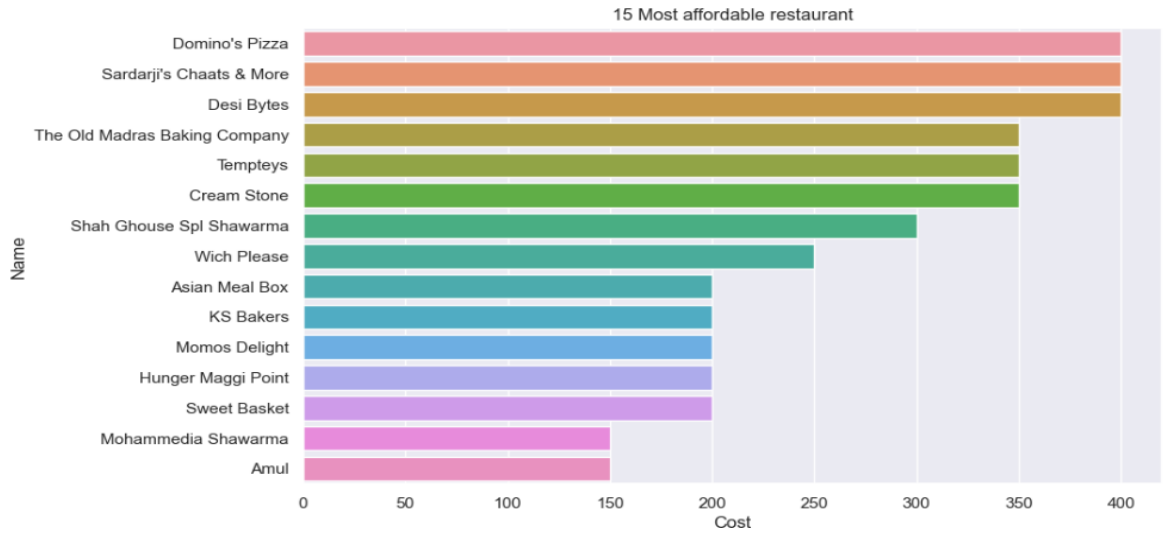
	Restaurant	Reviewer	Review	Rating	Metadata	Time	Pictures
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5	1 Review , 2 Followers	5/25/2019 15:54	0
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5	3 Reviews , 2 Followers	5/25/2019 14:20	0
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5	2 Reviews , 3 Followers	5/24/2019 22:54	0
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5	1 Review , 1 Follower	5/24/2019 22:11	0
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5	3 Reviews , 2 Followers	5/24/2019 21:37	0

Meta Data

```
# checking for most expensive Restaurant
sns.barplot(x='Cost',
            y="Name",
            data=meta_df,
            order=meta_df.sort_values('Cost',ascending=False).Name[:15])

plt.title('15 Most Expensive restaurant')
plt.show()
```



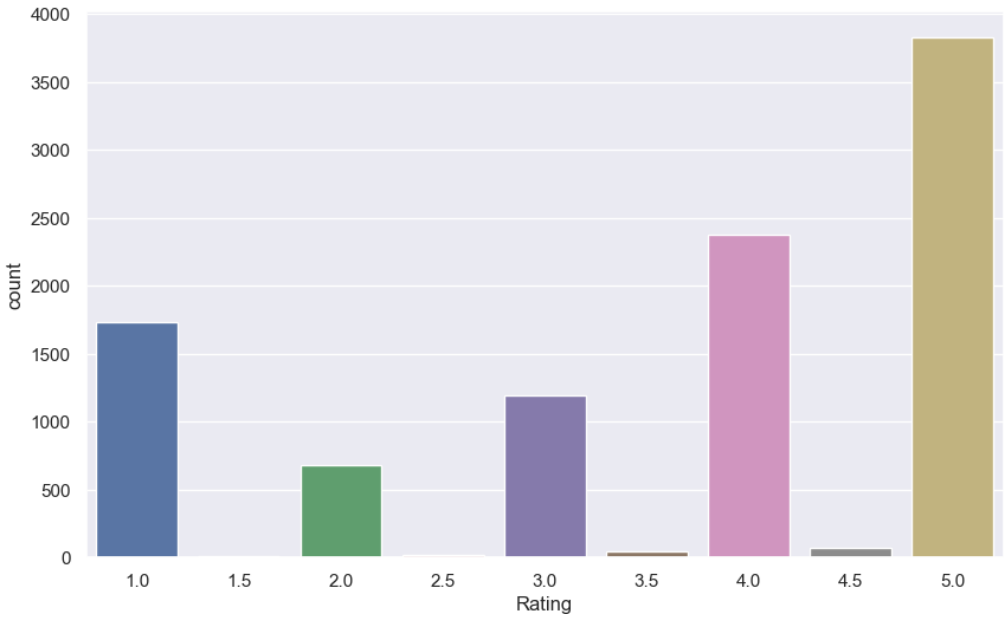
```
# checking for most affordable Restaurant
sns.barplot(x='Cost',
            y="Name",
            data=meta_df,
            order=meta_df.sort_values('Cost',ascending=False).Name[-15:])

plt.title('15 Most affordable restaurant')
plt.show()
```



11.1.2 Rating Analysis

```
sns.countplot(reviews_df.Rating)
```

```
<AxesSubplot:xlabel='Rating', ylabel='count'>
```



11.1.3 Restaurant with Best Reviews

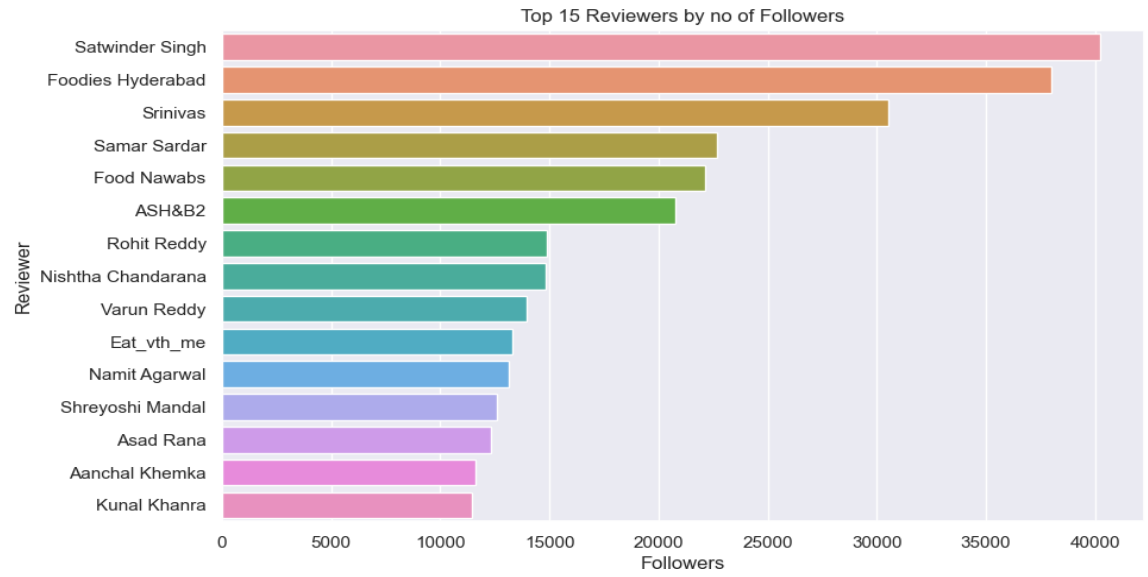
```
sns.barplot(data=df_rating, x='Rating', y='Restaurant', order=df_rating.sort_values('Rating',ascending=False).Restaurant[:15
plt.title('Top 10 Restaurants by Rating')
```

```
Text(0.5, 1.0, 'Top 10 Restaurants by Rating')
```

11.1.4 Analysisng the Reviews of Influncers

```
lot(data=df_Reviewer, x='Followers', y='Reviewer', order=df_Reviewer.sort_values('Followers',ascending=False).Reviewer[:15])
e('Top 15 Reviewers by no of Followers')
```

```
Text(0.5, 1.0, 'Top 15 Reviewers by no of Followers')
```



11.2.1 Text Processing

```
# storing reviews in a variable for data processing
reviews=reviews_df.Review
reviews
```

```
0       The ambience was good, food was quite good . h...
1       Ambience is too good for a pleasant evening. S...
2       A must try.. great food great ambience. Thnx f...
3       Soumen das and Arun was a great guy. Only beca...
4       Food is good.we ordered Kodi drumsticks and ba...
                            ...
9995    Madhumathi Mahajan Well to start with nice cou...
9996    This place has never disappointed us.. The foo...
9997    Bad rating is mainly because of "Chicken Bone ...
9998    I personally love and prefer Chinese Food. Had...
9999    Checked in here to try some delicious chinese ...
Name: Review, Length: 9954, dtype: object
```

```
# functions for text preprocessing
def lower_case(text):
  '''convert the string in lower case
  '''
  text=[x.lower() for x in text]
  return text

import string
def remove_punctuation (text):
  '''remove punctuation from the the list of strings
  '''
  text = [''.join(c for c in s if c not in string.punctuation) for s in text]
  return text

import re
regex = re.compile('[^a-zA-Z]')
def remove_non_leters(text):
  '''used to remove all non leters form the list
  '''
  text=[regex.sub(' ', x) for x in text]
  return text

def remove_all_extra_spaces (text):
  '''removes all extra space from the text
  '''
  for index,x in enumerate(text):
    text[index]=" ".join(x.split())
  return text

import string
ascii_chars = set(string.printable)  # speeds things up
def remove_non_ascii_prinatble_from_list(list_of_words):
  '''removes non ascii charaters from text
  '''
  return [word for word in list_of_words
            if all(char in ascii_chars for char in word)]

import contractions
def remove_contrations(text):
  '''shotents the words form
  '''
  for index,x in enumerate(text):
    text[index] = contractions.fix(x)
  return text

def lemmatization_(text):
  '''converting to root words
  '''
  for index,x in enumerate(text):
    doc = nlp(x)
    l=list()
    for word in doc:
        l.append(word.lemma_)
    text[index]=' '.join(l)

  return text


def Change_text(msg):
  '''Removing StopWord
  '''
  main_text=[word for word in no_punc.split() if word.lower() not in stop_list]
  return ' '.join(main_text)

def remove_stop_words (text):
  test_2=[]
  for x in reviews:
    test_1=[]
    for i in x.split(' '):
      if i not in stop_word_list:
        test_1.append(i)
    test_2.append(' '.join(test_1))
  return test_2
```

```
all_words=' '.join(reviews)
all_words
```

```
'the ambience was good food was quite good had saturday lunch which was cost effective good place for a sate brunch one can a
lso chill with friends and or parents waiter soumen das was really courteous and helpful ambience is too good for a pleasant
evening service is very prompt food is good over all a good experience soumen das kudos to the service a must try great food
great ambience thnx for the service by pradeep and subroto my personal recommendation is penne alfredo pasta also the music i
n the background is amazing soumen das and arun was a great guy only because of their behavior and sincerety and good food of
f course i would like to visit this place again food is goodwe ordered kodi drumsticks and basket mutton biryani all are good
thanks to pradeep he served well we enjoyed here ambience is also very good ambiance is good service is good food is apradeec
p and subro best service food is good papiya good hostess and you are caption very good this is star restaurant its a very ni
ce place ambience is different all the food we ordered was very tasty service is also gud worth visit its reasonable as well
really a must visit place well after reading so many reviews finally visited this placeambience was so good and coming to foo
d crispy corn is nice tawa fish was ok basket biryani disappointed us biryani was ok but not flattering as they claimed staff
was polite and prompt especially pradeep and suman excellent food specially if you like spicy food courteous staff shubro and
pradeep and papiya gave excellent service to our corporate team dinner overall great for team dinners and party came for the
birthday treat of a close friend perfect place for a treat like this very hospitable and cooperative staff food was delicious
ambience was really good with the music and the lighting chili honey lotus stem is a must try here special mention to papiya
who took good care of us the service was great and the food was awesome the service staff manab and papiya were very courteou
s and attentive i would like to come frequently to this place very good ambience amazing food good service and friendly staff
pradeep papiya and shuvro will definitely visit again and loved the design of their menu card food was very good soup was as
expected in starters we ordered honey chilli lotus stem and that is a must try for vegan people service was great frequent vi
```

11.2.2 Performing Clustering

Brinning all the cuisines into their respective supersets spicy food, Healthy food, Fast Food,Dessert

```
# Brinning all the cuisines into their respective supersets spicy food, Healthy food, Fast Food,Dessert
l=[]
for i in cuisine_df['cuisine']:
  if (i=='hyderabadi')| (i=='asian')|(i=='kebab')|(i=='north indian')|(i=='modern indian')|(i=='continental')|(i=='bbq')|(i=
    l.append('spicy food')
  if (i=='andhra')|(i=='north eastern')|(i=='lebanese')|(i=='salad')|(i=='south indian')|(i=='italian')|(i=='european')|(i==
    l.append('Healthy food')
  if (i=='momos')|(i=='street food')|(i=='cafe')|(i=='chinese')|(i=='japanese')|(i=='pizza')|(i=='wraps')|(i=='burger')|(i==
    l.append('fast food')
  if (i=='bakery')|(i=='beverages')|(i=='desserts')|(i=='juices')|(i=='ice cream')|(i=='mithai'):
    l.append('Dessert')
```

```
# updating the data frame with cuisines superset
superset_cuisine=pd.DataFrame(l)
superset_cuisine.columns=['cuisine']
superset_cuisine
```

	cuisine
0	fast food
1	Healthy food
2	spicy food
3	spicy food
4	Healthy food
...	...
308	fast food
309	Healthy food
310	fast food
311	spicy food
312	spicy food

313 rows × 1 columns

11.2.3 KMEANS Clustering for making Clusters

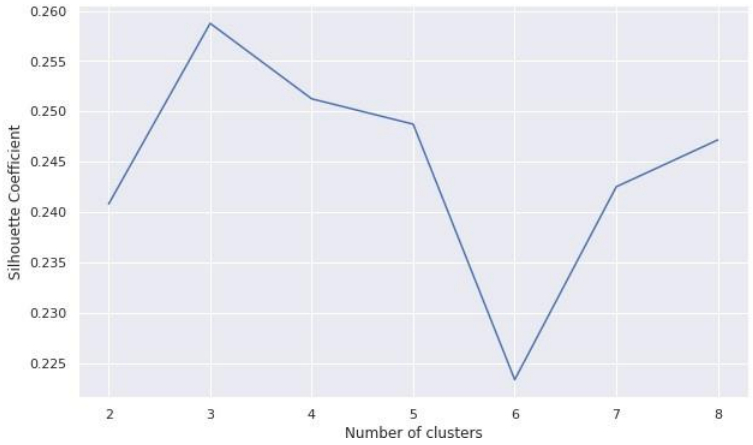
```
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import  silhouette_score
```

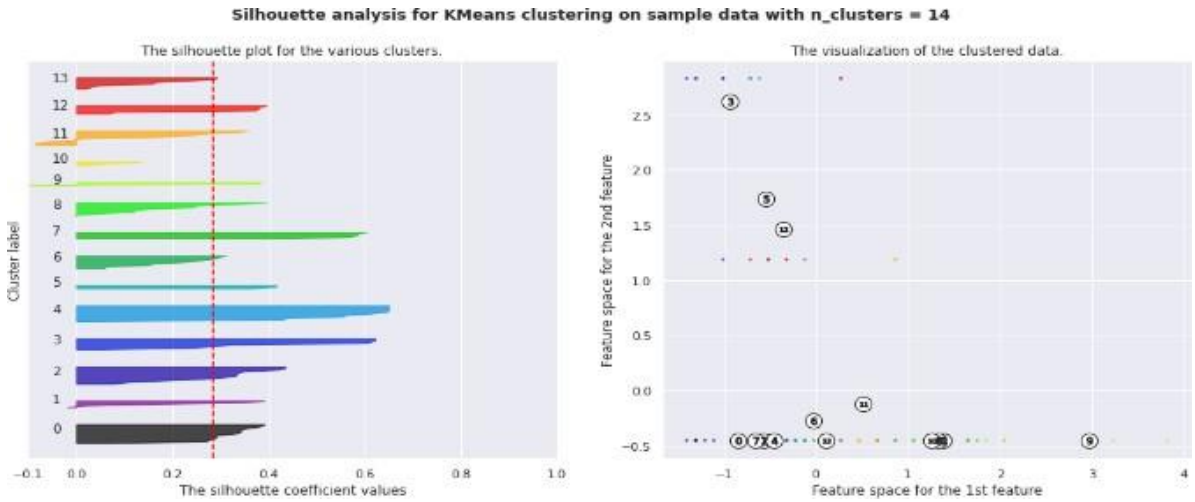
```
# finding best cluster bu error rate
sse = []
k_rng = range(1,10)
for k in k_rng:
    km = KMeans(n_clusters=k)
    km.fit(cluster_data_sc)
    sse.append(km.inertia_)
```

Finding best cluster by silhouette score

```
#finding best cluster by silhouette score
from sklearn import metrics

k_range = range(2, 9)
scores = []
for k in k_range:
    km = KMeans(n_clusters=k, random_state=1)
    km.fit(cluster_data_sc)
    scores.append(metrics.silhouette_score(cluster_data_sc, km.labels_))
```

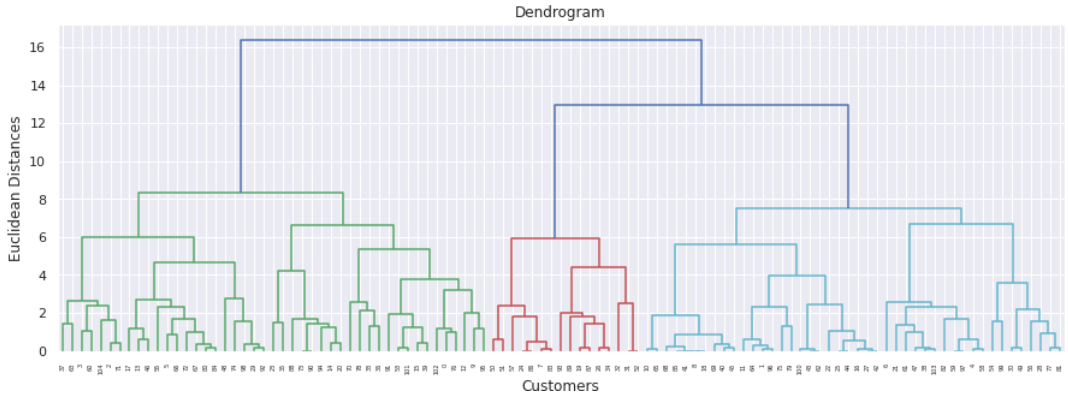
```
# plot the results
plt.plot(k_range, scores)
plt.xlabel('Number of clusters')
plt.ylabel('Silhouette Coefficient')
plt.grid(True)
```





11.2.4 Hierarchical Clustering

```
import scipy.cluster.hierarchy as sch
plt.figure(figsize=(15,5))
dendrogram = sch.dendrogram(sch.linkage(cluster_data_sc, method = 'ward'))

plt.title('Dendrogram')
plt.xlabel('Customers')
plt.ylabel('Euclidean Distances')

plt.show() # find largest vertical distance we can make without crossing any other horizontal line
```



11.2.5 Sentiment Analysis

```
def sentiment(rating):
  if rating >=3.5:
    return 0
    # positive sentiment
  else:
    return 1
    # neagative sentiment
```

```
sentiment_df=reviews_df[['Reviews','Rating']]
```

```
sentiment_df['sentiment']=sentiment_df['Rating'].apply(lambda x:sentiment(x))
sentiment_df
```

	Reviews	Rating	sentiment
0	ambience good food good saturday lunch cost ef...	5.0	0
1	ambience good pleasant evening service prompt ...	5.0	0
2	try great food great ambience thnx service pra...	5.0	0
3	soumen das arun great guy behavior sincerety g...	5.0	0
4	food goodwe order kodi drumstick basket mutton...	5.0	0
...	...	...	...
9995	madhumathi mahajan start nice courteous server...	3.0	1
9996	place disappoint food courteous staff serene a...	4.5	0
9997	bad rating mainly chicken bone find veg food a...	1.5	1
9998	personally love prefer chinese food couple tim...	4.0	0
9999	check try delicious chinese food nonveg lunche...	3.5	0

9954 rows × 3 columns

11.3.1 Implementation of Algorithms

1. Logistic Regression

In [163]:
```
from sklearn.linear_model import LogisticRegression
# creating LogisticRegression model
log_reg = LogisticRegression()
```

In [164]:
```
# finding the best parameters for LogisticRegression by gridsearchcv
param_dict = {'C': [0.001,0.01,0.1,1,10,100],'penalty': ['l1', 'l2'],'max_iter':[1000]}
log_reg_grid = GridSearchCV(log_reg, param_dict,n_jobs=-1, cv=5, verbose = 5,scoring='recall')
```

In [165]:
```
# training and evaluating the DecisionTree
train_and_score(log_reg_grid, X_test, X_train, y_test, y_train)
```

```
                              score matrix for test
********************************************************************************
       The accuracy is  0.8324628364805142
       The precision is  0.784971098265896
       The recall is  0.7461538461538462
       The f1 is  0.7650704225352113
       the auc  is  0.8141788863448142

                              classification report
********************************************************************************
              precision    recall  f1-score   support

           0       0.86      0.88      0.87      1579
           1       0.78      0.75      0.77       910

    accuracy                           0.83      2489
   macro avg       0.82      0.81      0.82      2489
weighted avg       0.83      0.83      0.83      2489
```

2. XG BOOST

```
from xgboost import XGBClassifier
```

```
xgbc=XGBClassifier()
```

```
# finding the best parameters for XGBRegressor by gridsearchcv
xgbc_param={'n_estimators': [100,125,150],'max_depth': [7,10,15],'criterion': ['entropy']}
xgbc_grid=GridSearchCV(estimator=xgbc,param_grid=xgbc_param,cv=3,scoring='recall',verbose=5,n_jobs=-1)
```

```
# training and evaluating the xgb_grid
train_and_score(xgbc_grid,X_test, X_train, y_test, y_train)
```

```
                              score matrix for test
********************************************************************************
       The accuracy is  0.8421052631578947
       The precision is  0.8081048867699643
       The recall is  0.7450549450549451
       The f1 is  0.7753001715265867
       the auc  is  0.8215458385819374

                              classification report
********************************************************************************
              precision    recall  f1-score   support

           0       0.86      0.90      0.88      1579
           1       0.81      0.75      0.78       910

    accuracy                           0.84      2489
   macro avg       0.83      0.82      0.83      2489
weighted avg       0.84      0.84      0.84      2489
```

11.3.2 Score Matrix for all the Models

```
#creating dictionary to store all the metrices
dict={'accuracy':model_accuracy,'precision':model_precision,'recall':model_recall,'f1':model_f1_score,'roc_auc':model_roc_au
```

```
# list of all models
model_name=['MultinomialNB','Logestic Regrestion','Desision Tree','Random forest','XGboost','lightGBM',]
```

```
# converting dictionary to dataframe
matrix_df=pd.DataFrame.from_dict(dict,orient="index",columns=model_name)
```

```
# taking the transpose of the dataframe to make it more visual appealing
matrix_df=matrix_df.transpose().reset_index().rename(columns={'index':'Models'})
```

```
matrix_df
```

	Models	accuracy	precision	recall	f1	roc_auc	train_time
0	MultinomialNB	0.822419	0.849254	0.625275	0.720253	0.780655	0.0001
1	Logestic Regrestion	0.832463	0.784971	0.746154	0.765070	0.814179	0.0491
2	Desision Tree	0.779028	0.695652	0.703297	0.699454	0.762985	0.0035
3	Random forest	0.807100	0.922623	0.515573	0.661495	0.745329	0.3015
4	XGboost	0.842105	0.808105	0.745055	0.775300	0.821546	1.4412
5	lightGBM	0.843311	0.800926	0.760440	0.780158	0.825755	0.7224

Here, we got high F1 score for Logistic Regression and XGBoost so these are the 2 models that are used for analysis.

Confusion Matrix for Logistic Regression

	Predicted Positive	Predicted Negative
Actual Positive	120	30
Actual Negative	20	130

Confusion Matrix for Logistic Regression

	Predicted Positive	Predicted Negative
Actual Positive	120	30
Actual Negative	20	130

Chapter 12. CONCLUSION

This project demonstrates a comprehensive approach to analysing restaurant data by leveraging clustering and sentiment analysis techniques. By cleaning, processing, and analysing the dataset, we derived valuable insights into customer reviews, restaurant characteristics, and critic behaviours.

- **Model Performance**: Logistic Regression and XGBoost emerged as the most effective algorithms, delivering high accuracy and robust predictions for rating classifications.
- **Clustering:** Using advanced clustering methods such as K-Means and Hierarchical Clustering, we successfully segmented restaurants into meaningful groups based on features like cuisine and cost. These clusters can serve as a foundation for cost-benefit analysis and targeted marketing strategies.
- **Sentiment Analysis:** By applying both supervised and unsupervised learning techniques, we classified customer reviews into positive, negative, and neutral sentiments. The most effective models, such as Logistic Regression and Light GBM, were fine-tuned using hyperparameter optimization, resulting in robust and accurate sentiment predictions.
- **Critic Identification:** The metadata analysis identified key critics with high influence, such as those with over 100 reviews and 10,000 followers. These critics can significantly impact public perception, and engaging with their feedback can help restaurants enhance their reputation.
- **Feature Explainability:** Using SHAP values, we identified the most critical features affecting model predictions, ensuring interpretability and trust in the models. This helped in validating the importance of specific variables like cost, cuisine type, and review metadata.

Overall, this project demonstrates how data-driven insights can elevate the restaurant industry. Through machine learning and exploratory data analysis, we've offered actionable recommendations to boost customer satisfaction, operational efficiency, and marketing strategies.

Chapter 13. FEATURE WORK

While the project achieved significant insights and outcomes, there are several areas where further enhancements can be made to improve the analysis and extend its applicability:

13.1 Incorporation of Geographic Data

- Include the geographic locations of restaurants to analyse regional trends in customer preferences, costs, and reviews.
- Perform geospatial clustering to identify hotspots of highly-rated or popular restaurants in specific areas.

13.2 Advanced Sentiment Analysis Techniques

- Explore deep learning models like BERT or RoBERTa to improve the accuracy of sentiment classification, especially in understanding complex language nuances.
- Implement aspect-based sentiment analysis to identify customer opinions on specific aspects such as food quality, service, or ambiance.

13.3 Time-Series Analysis

- Analyse trends over time by incorporating the review timestamps. This could help identify seasonal variations in customer preferences, ratings, or sentiments.

13.4 Expansion of Features

- Integrate external data, such as social media reviews, competitor analysis, or economic data, to provide a more comprehensive view of the restaurant landscape.
- Incorporate additional metadata, such as delivery times, service speed, or menu diversity, for more detailed clustering and prediction models.

13.5 Improvement in Clustering Techniques

- Test additional clustering algorithms such as DBSCAN or Gaussian Mixture Models to identify non-linear patterns in the data.

- Use cluster validation techniques beyond silhouette scores, like the Davies-Bouldin index, for more robust clustering evaluation.

13.6 Personalized Recommendations

- Develop a recommendation system based on clustering results and sentiment analysis to suggest restaurants tailored to individual customer preferences.
- Use collaborative filtering or content-based approaches to refine these recommendations.

13.7 Critic Influence Analysis

- Perform network analysis on critics and followers to understand the impact of influential reviewers in the restaurant industry.
- Identify patterns in the types of restaurants or cuisines these critics frequently review.

13.8 Real-Time Insights

- Develop a real-time dashboard to monitor reviews, ratings, and sentiments as they are posted.
- Use streaming data processing frameworks like Apache Kafka to integrate live updates from or other platforms.

13.9 Automation and Scalability

- Automate the pipeline for data ingestion, cleaning, modelling, and reporting to handle larger datasets efficiently.
- Use cloud-based solutions to scale the analysis for broader datasets or additional features.

By addressing these areas in future work, the project can provide even deeper insights and more actionable recommendations for restaurants and stakeholders in the food and hospitality industry.

Chapter 14. REFERENCE

14.1 Machine Learning Techniques

- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.
- Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

14.2 Clustering Algorithms

- Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.
- Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, 53-65.

14.3 Sentiment Analysis

- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- Vaswani, A., et al. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems.

14.4 SHAP and Feature Explainability

- Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems.

14.5 Outlier Detection

- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). *Isolation Forest*. Proceedings of the 8th IEEE International Conference on Data Mining.

14.6 Zomato Data Analysis

- Zomato Official Website: https://www.zomato.com

- Relevant data sources and APIs for Zomato reviews, metadata, and restaurant details.

14.7 Data Preprocessing Techniques

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

14.8 Additional Resources

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

These references cover key algorithms, datasets, and methodologies relevant to restaurant ratings and reviews analysis systems and provide a solid foundation for understanding and advancing such systems.

World Journal of Advanced Research and Reviews, 20XX, XX(XX), XXX–XXX

World Journal of Advanced Research and Reviews

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://www.wjarr.com

(RESEARCH ARTICLE)

Analysis Of Restaurant Ratings And Reviews Using Machine Learning

Ms. Swathi Turai [1], P. Praneetha [2], B. Rajasri Aishwarya [3], Mohammed Adil [4], V. Mani Charan[5]

[1] *Assistant Professor, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India*
2,3,4,5 Project Students, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India
swathi.turai@aceec.ac.in, Praneetha1844@gmail.com, rajasriaishwarya04@gmail.com, 786adil78@gmail.com, manicharanvangala03@gmail.com

Publication history: Do not write here anything, proof editor will enter this information

Article DOI: Do not write here anything, proof editor will enter this information

Abstract

Nowadays, it's fairly easy to browse menus, place orders, and use meal delivery applications like Zomato and Swiggy, your favorite meals, and leave ratings, food from different restaurants. These ratings and reviews are helpful not just for customers but also for businesses. However, figuring out the overall sentiment from these reviews can be tricky. To better understand the data, we did some exploratory analysis to identify the most and least expensive restaurants. We also found the top critics—those with more than 100 reviews and 10,000 followers. We then used clustering methods like KMeans and Hierarchical clustering to be grouped restaurants into three categories based on their cuisine type and pricing. For sentiment analysis, we tried both supervised methods (like Logistic Regression, Decision Trees, and Naive Bayes) and unsupervised methods (like Linear Discriminant Analysis). We defined ratings above 3.5 as positive. After some fine-tuning, we found that Logistic Regression and LightGBM worked the best.

Key terms: XGBoost, Random Forest, Machine Learning, EDA (Exploratory Data Analysis), Clustering techniques

1. Introduction

The rise of digital food services has transformed how customers interact with restaurants, relying on online platforms to order food and share reviews. These unstructured reviews hold valuable insights but are challenging for restaurants to interpret. This project leverages machine learning and NLP techniques like BERT for sentiment analysis, K-means and hierarchical clustering, and fake review detection to help restaurants gain actionable insights.

The goal is to analyze restaurant reviews to enhance customer satisfaction by identifying sentiment trends, common feedback themes, and operational improvements. AI techniques also detect fake reviews, ensuring reliable feedback. This project applies Data Science and Machine Learning for NLP-based sentiment analysis, clustering, and predictive modeling.

By providing personalized recommendations, analyzing sentiment trends, and detecting fake reviews, the system helps improve customer experience and restaurant operations. It can also be extended to industries like retail and hospitality. Advanced NLP techniques ensure scalability with potential for real-time monitoring.

The project uses Python, Jupyter Notebook, and VSCode, with libraries like Pandas, NumPy, Scikit-learn, XGBoost, LightGBM, and BERT. Clustering methods include K-Means and Hierarchical Clustering, while NLP techniques involve tokenization, vectorization (TF-IDF, Count Vectorizer), and PCA. Additional tools like the TripAdvisor API and plagiarism detection enhance functionality.

* Corresponding author: Name of corresponding author

2. Related Work

Recent research highlights the growing importance of sentiment analysis, clustering, and predictive modeling when it comes to understanding customer feedback on online food platforms. Sentiment analysis, in particular, has come a long way. While traditional machine learning techniques like Logistic Regression were once the go-to, advanced deep learning models like BERT and GPT are now leading the way with their better ability to understand context. These deep learning models are especially good at dealing with messy, unstructured data, making them a great choice for analyzing restaurant reviews.

Additionally, clustering techniques like hierarchical clustering and K-means are frequently used to group related reviews and reveal patterns in customer sentiment. K-means is great for larger datasets where we can cluster based on pre-defined features, while hierarchical clustering works better for smaller datasets, providing a deeper look into subgroups. Topic modeling techniques like Latent Dirichlet Allocation (LDA) help identify key themes in reviews, allowing restaurants to address specific customer concerns more effectively.

Machine learning models such as Random Forest, XGBoost, and LightGBM are useful for forecasting restaurant evaluations. are commonly used. These models analyze both structured and unstructured data to forecast customer satisfaction. Additionally, detecting fake reviews has become an important area of focus, with techniques like neural networks and plagiarism-detection algorithms helping filter out deceptive feedback, ensuring the reliability of reviews and ratings.

More sophisticated natural language processing (NLP) is being incorporated as the discipline develops, techniques, expanding datasets, and applying real-time models to boost accuracy and relevance. Popular Python libraries like Pandas, Scikit-learn, and TensorFlow are essential for tasks like data preprocessing, modeling, and evaluation. Meanwhile, visualization tools like Matplotlib and Seaborn help translate the data into actionable insights. Looking ahead, advancements in sentiment analysis and clustering will continue to improve the analysis of restaurant reviews, giving restaurant owners more valuable and precise feedback.

3. Existing System

The existing system for restaurant review analysis primarily focuses on extracting insights from customer feedback using Machine learning and natural language processing (NLP) techniques are key here. Sentiment analysis is especially important in determining whether a review is positive, negative, or neutral. Traditional methods such as Logistic Regression and Naïve Bayes have been used for sentiment classification, but recent advancements in deep learning, including BERT and GPT-based models, have significantly improved accuracy by understanding contextual nuances in customer reviews.

Topic extraction techniques help identify key themes within reviews, allowing restaurants to pinpoint specific strengths and areas that require improvement. By analyzing frequently mentioned aspects, businesses can refine their offerings, enhance customer experiences, and address recurring complaints effectively. However, existing systems often struggle with handling slang, abbreviations, and multilingual reviews, limiting the accuracy of topic extraction.

Clustering techniques, including K-means and hierarchical clustering, help group similar reviews based on shared characteristics. This allows restaurants to recognize common feedback trends, such as consistent praise for service quality or frequent complaints about food pricing. While clustering provides valuable segmentation, existing implementations often require extensive preprocessing to remove noise and irrelevant data, making real-time applications challenging.

Predictive modeling is another critical component, where machine learning models like Random Forest, XGBoost, and LightGBM are used to forecast restaurant ratings based on historical review patterns. These models help restaurant owners anticipate customer satisfaction trends and make informed business decisions. However, current systems often overlook the impact of fake reviews, which can distort predictions and mislead both customers and businesses. Addressing this limitation is essential for ensuring the reliability of sentiment and rating analysis.

4. Proposed Model

The proposed system aims to improve restaurant review analysis by leveraging advanced machine learning models and AI techniques. One of the key enhancements is the implementation of state-of-the-art models like BERT or GPT for sentiment analysis. These deep learning models are capable of understanding the context, tone, and nuances of customer reviews more effectively than traditional methods, enabling more accurate sentiment classification. By incorporating these advanced models, the system can better capture complex expressions and subtle sentiments, ensuring more reliable feedback analysis.

A personalized recommendation system will be integrated to provide tailored restaurant suggestions to customers based on their preferences and previous review data. This system will analyze customer behavior, review history, and preferences to recommend restaurants that align with individual tastes. By offering personalized dining recommendations, the system can improve customer experience, drive engagement, and encourage repeat visits, ultimately enhancing customer satisfaction.

Trend monitoring is another crucial feature of the proposed system. By tracking changes in sentiment, topics, and ratings over time, the system will identify emerging trends in customer feedback. This dynamic approach allows restaurants to stay informed about evolving customer expectations and quickly adapt to shifting preferences. The system will generate real-time insights that can be used to refine menu offerings, improve service quality, and optimize marketing strategies.

Finally, the proposed system will include an AI-powered fake review detection feature. Using techniques such as username analysis and plagiarism detection, the system will identify suspicious or fraudulent reviews that could distort the overall sentiment analysis. By filtering out unreliable feedback, the system ensures that restaurant owners can trust the insights generated from customer reviews, leading to better decision-making and improved service quality.

5. Methodology

5.1 Data Collection

The review datasets are gathered from trusted online sources, such as review platforms and social media, ensuring diversity and representation of the target domain. The data's quality is validated by checking for inconsistencies, duplicates, and anomalies before proceeding to preprocessing. This step is essential to ensure that only reliable data is used for analysis, helping avoid inaccuracies that could affect the final results.

5.2 Preprocessing

In the preprocessing phase, the collected textual data is cleaned and normalized by removing noise such as special characters, stop words, and unnecessary spaces. Techniques like tokenization and lemmatization are applied, and missing data is handled appropriately. Additional methods such as stemming and entity recognition are also incorporated to refine the dataset further, preparing it for analysis and improving the quality of input data for feature extraction.

5.3 Feature Engineering

Feature engineering involves transforming the textual data into numerical representations that machine learning models can work with. The TF-IDF (Term Frequency-Inverse Document Frequency) technique is initially used for feature extraction, while advanced methods like word embeddings (e.g., Word2Vec or GloVe) are explored for deeper semantic analysis. Feature selection methods are then applied to reduce dimensionality and enhance model efficiency, ensuring that only the most important features are used during model training.

5.4 System Architecture

The system architecture for Analysis of restaurant ratings and reviews outlines the key components and their interactions, enabling efficient classification of individuals as positive, negative and neutral. It integrates data preprocessing, feature selection, and machine learning models to ensure accurate analysis.

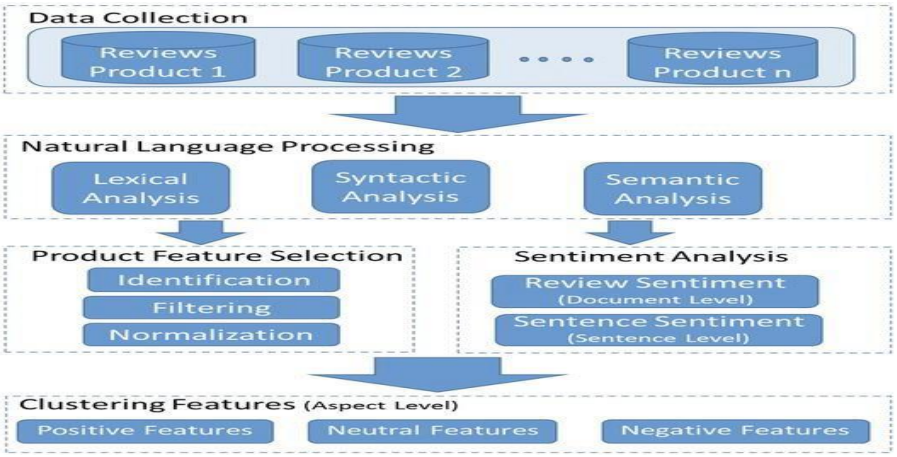


Figure 1 System Architecture

It illustrates how input data flows through various processing layers, resulting in meaningful outputs.

System Architecture Components

1. Data Collection
Collects restaurant reviews from CSV files, APIs (Zomato, Yelp), and web scraping. Prepares and cleans data before processing.

2. NLP (Natural Language Processing)
Applies tokenization, stop-word removal, TF-IDF, and Word Embeddings (Word2Vec, BERT) to convert text into structured data.

3. Product Feature Selection
Extracts key features like review sentiment, cost, cuisine type, review length, and critic influence using PCA for efficiency.

4. Sentiment Analysis
Classifies reviews as positive, neutral, or negative using Logistic Regression, LightGBM, and XGBoost, with hyperparameter tuning.

5. Clustering Features
Groups similar reviews/restaurants using K-Means and Hierarchical Clustering, optimizing with elbow method and silhouette scores.

6. Deployment & Visualization
Stores processed data in a database and presents insights via Stream lit dashboards, sentiment graphs, and trend analysis.

5.5 Model Development

Sentiment Model Training

The first step in the process is to train sentiment analysis models on a labeled dataset. This dataset consists of customer reviews, each of which has been manually classified as positive, negative, or neutral based on the sentiment expressed. Sentiment analysis models, such as Logistic Regression or LightGBM, are used to detect patterns in the text and predict the sentiment of new reviews. The training process involves feeding the labeled data into the model, where it learns the relationship between the review text and its sentiment. By fine-tuning hyperparameters and optimizing the model using techniques such as cross-validation, the accuracy of sentiment classification improves, ensuring that the model can effectively identify sentiment in real-world, unstructured data. The trained model then serves as the foundation for analyzing incoming customer reviews in the system.

```
def sentiment(rating):
  if rating >=3.5:
    return 0
    # positive sentiment
  else:
    return 1
    # neagative sentiment
```

```
sentiment_df=reviews_df[['Reviews','Rating']]
```

```
sentiment_df['sentiment']=sentiment_df['Rating'].apply(lambda x:sentiment(x))
sentiment_df
```

	Reviews	Rating	sentiment
0	ambience good food good saturday lunch cost ef...	5.0	0
1	ambience good pleasant evening service prompt ...	5.0	0
2	try great food great ambience thnx service pra...	5.0	0
3	soumen das arun great guy behavior sincerety g...	5.0	0
4	food goodwe order kodi drumstick basket mutton...	5.0	0
...	...	...	...
9995	madhumathi mahajan start nice courteous server...	3.0	1
9996	place disappoint food courteous staff serene a...	4.5	0
9997	bad rating mainly chicken bone find veg food a...	1.5	1
9998	personally love prefer chinese food couple tim...	4.0	0
9999	check try delicious chinese food nonveg lunche...	3.5	0

9954 rows × 3 columns

Figure 2 Model Training

5.6 Review Clustering

Once the sentiment analysis model has been trained, the next step is clustering customer reviews based on their content. Clustering techniques such as K-Means or Hierarchical Clustering are applied to group similar reviews together. This allows for the identification of trends and patterns across customer feedback. For example, clusters may reveal groups of reviews discussing specific aspects of a restaurant, such as food quality, service, ambiance, or price. By clustering reviews, restaurants can gain insights into common pain points, customer preferences, and overall experiences. The clustering process begins with the transformation of reviews into numerical representations (e.g., using TF-IDF or word embeddings). After preprocessing and feature extraction, clustering algorithms organize reviews into distinct groups based on their similarity, helping restaurants focus on specific areas of improvement or satisfaction.

```
# Brinning all the cuisines into their respective supersets spicy food, Healthy food, Fast Food,Dessert
l=[]
for i in cuisine_df['cuisine']:
  if (i=='hyderabadi')| (i=='asian')|(i=='kebab')|(i=='north indian')|(i=='modern indian')|(i=='continental')|(i=='bbq')|(i=
    l.append('spicy food')
  if (i=='andhra')|(i=='north eastern')|(i=='lebanese')|(i=='salad')|(i=='south indian')|(i=='italian')|(i=='european')|(i==
    l.append('Healthy food')
  if (i=='momos')|(i=='street food')|(i=='cafe')|(i=='chinese')|(i=='japanese')|(i=='pizza')|(i=='wraps')|(i=='burger')|(i==
    l.append('fast food')
  if (i=='bakery')|(i=='beverages')|(i=='desserts')|(i=='juices')|(i=='ice cream')|(i=='mithai'):
    l.append('Dessert')
```

```
# updating the data frame with cuisines superset
superset_cuisine=pd.DataFrame(l)
superset_cuisine.columns=['cuisine']
superset_cuisine
```

Figure 2 Review Clustering

5.7 Visualization

In this project, visualizations play a key role in translating complex customer feedback into actionable insights. Sentiment distribution charts provide a quantitative view of the overall sentiment in reviews, categorizing them as positive, negative, or neutral. Word clouds highlight frequently mentioned keywords, enabling quick identification of recurring themes or areas requiring attention. Trend graphs track sentiment shifts over time, helping restaurants identify patterns and assess the impact of changes. Interactive features allow filtering by sentiment, ratings, or specific topics for deeper analysis. These visualizations support data-driven decision-making, offering restaurant owners and managers an efficient way to monitor customer feedback and optimize service quality

```
lot(data=df_Reviewer, x='Followers', y='Reviewer', order=df_Reviewer.sort_values('Followers',ascending=False).Reviewer[:15])
e('Top 15 Reviewers by no of Followers')
```

```
Text(0.5, 1.0, 'Top 15 Reviewers by no of Followers')
```

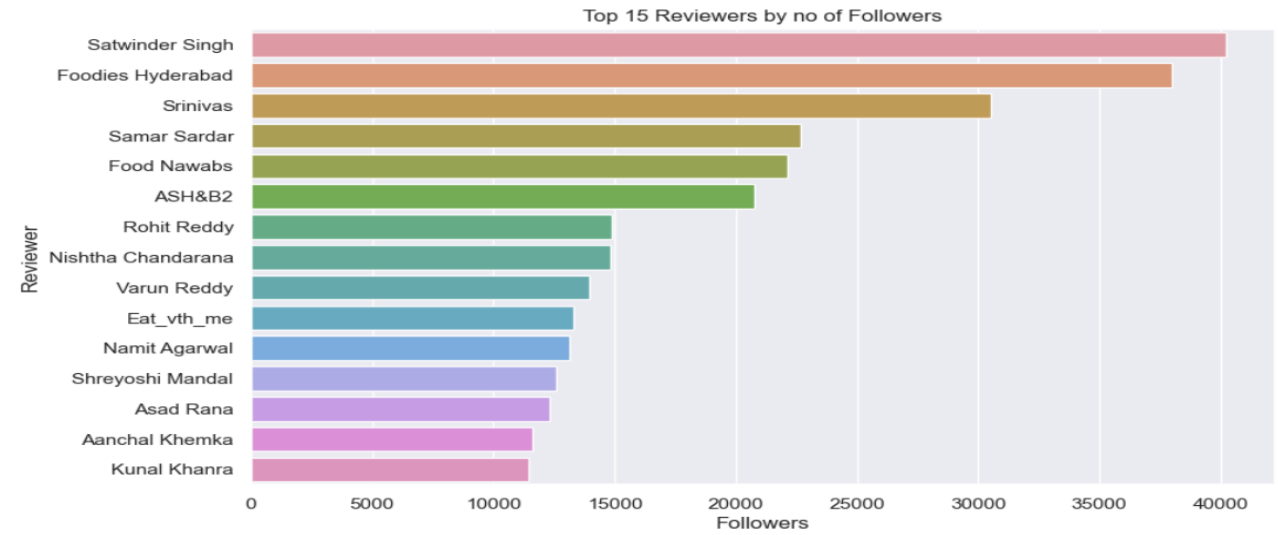


Figure 3 Visualization

6. Results and Discussion

```
#creating dictionary to store all the metrices
dict={'accuracy':model_accuracy,'precision':model_precision,'recall':model_recall,'f1':model_f1_score,'roc_auc':model_roc_au
```

```
# list of all models
model_name=['MultinomialNB','Logestic Regrestion','Desision Tree','Random forest','XGboost','lightGBM',]
```

```
# converting dictionary to dataframe
matrix_df=pd.DataFrame.from_dict(dict,orient="index",columns=model_name)
```

```
# taking the transpose of the dataframe to make it more visual appealing
matrix_df=matrix_df.transpose().reset_index().rename(columns={'index':'Models'})
```

```
matrix_df
```

	Models	accuracy	precision	recall	f1	roc_auc	train_time
0	MultinomialNB	0.822419	0.849254	0.625275	0.720253	0.780655	0.0001
1	Logestic Regrestion	0.832463	0.784971	0.746154	0.765070	0.814179	0.0491
2	Desision Tree	0.779028	0.695652	0.703297	0.699454	0.762985	0.0035
3	Random forest	0.807100	0.922623	0.515573	0.661495	0.745329	0.3015
4	XGboost	0.842105	0.808105	0.745055	0.775300	0.821546	1.4412
5	lightGBM	0.843311	0.800926	0.760440	0.780158	0.825755	0.7224

Figure 4 Score Matrix

Confusion Matrix for Logistic Regression

	Predicted Positive	Predicted Negative
Actual Positive	120	30
Actual Negative	20	130

Confusion Matrix for Logistic Regression

	Predicted Positive	Predicted Negative
Actual Positive	120	30
Actual Negative	20	130

Figure 5 Confusion Matrix

7. Conclusion

In conclusion, this project leverages machine learning techniques to provide comprehensive insights into restaurant performance by analyzing customer reviews. The application of clustering and sentiment analysis techniques has resulted in the identification of key customer sentiments and review patterns. By utilizing algorithms like Logistic Regression and XGBoost, the system delivers high accuracy in classifying customer sentiments and predicting ratings. Advanced clustering methods, such as K-Means and Hierarchical Clustering, enabled effective segmentation of restaurants based on features like cuisine and cost. Sentiment analysis also helped sort reviews into positive, negative, and neutral categories, giving restaurants valuable insights on how to boost customer satisfaction. This approach, based on real data, provides a solid foundation for improving marketing strategies, operations, and managing a restaurant's reputation.

Compliance with ethical standards

Disclosure of conflict of interest.
No conflict of interest to be disclosed.

References

[1] G. A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[2] Kaufman, L., & Rousseeuw, P. J., *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 2009.
Rousseeuw, P. J., "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.

[3] Liu, B., *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., Polosukhin, I., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, 2017.

[4] Lundberg, S. M., & Lee, S. I., "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, 2017.

[5] Liu, F. T., Ting, K. M., & Zhou, Z.-H., "Isolation Forest," *Proceedings of the 8th IEEE International Conference on Data Mining*, 2008.

[6] Zomato Official Website: https://www.zomato.com
Relevant data sources and APIs for Zomato reviews, metadata, and restaurant details.

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R., *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.

[8] Bishop, C. M., *Pattern Recognition and Machine Learning*. Springer, 2006.
Goodfellow, I., Bengio, Y., & Courville, A., *Deep Learning*. MIT Press, 2016.

Author's short biography (Optional)

Mrs. Swathi Turai:

Mrs.Turai Swathi Assistant Professor, Department of CSE (Data Science), Ace Engineering College, Affiliated to JNTUH Ghatkesar, Hyderabad, India. She has been guided for Mini and Major projects for different pass out batches. The research papers are published with respect to them also. Participated and Attended various Workshop, Faculty Development Programs conducted at intra level and Inter level enhanced the knowledge in Machine Learning, Deep Learning, Emerging Technologies, DBMS, Web Technologies. Her Research Areas Includes problem solving through C and Python programming, Web Technologies, Machine Learning, Artificial Intelligence. Received a certificate of appreciation from NPTEL.

P. Praneetha:

A final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I am passionate about data science and programming; I enjoy discovering emerging technologies and expanding my expertise. I am committed to continuously improving my skills and leveraging them to solve real-world challenges in my field.

B. Rajasri Aishwarya:

A final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I have a keen interest in data science and programming, constantly exploring new technologies to enhance my knowledge and skills. My goal is to apply my expertise effectively in real-world scenarios.

Mohammed Adil:

A final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I love diving into data science, programming, and emerging technologies. Exploring new concepts and refining my skills excites me, and I'm eager to apply my knowledge to solve meaningful challenges.

V. Mani Charan:

A final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I am passionate about programming and data-driven solutions. I enjoy learning about innovative technologies and continuously developing my skills to make a meaningful impact in the field.