SENTIMNET ANALYSIS OF HOTEL REVIEWS

SUMMER INTERNSHIP PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS TO

RGUKT- SRIKAKULAM

FOR THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY IN
COMPUTER SCIENCE AND ENGINEERING

Submitted By:

B.SREEHITHA S190605

Under the Esteemed Guidance of

Mr.Aswini kumar, M.Tech
Tautor at EDUNET foundation (APSSDC)

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES
SRIKAKULAM

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
RGUKT-SRIKAKULAM, ETCHERLA- June 2024
RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

CERTIFICATE

This is to certify that the summer Internship project report titled “SENTIMENT ANALYSIS OF HOTEL REVIEWS” was successfully completed by BASIREDDY SREEHITHA (S190605) under the guidance of Mr. Aswini Kumar Tautor at Edunet Foundation. In partial fulfillment of the requirements for the Summer Internship Project in Computer Science and Engineering of Rajiv Gandhi University of Knowledge Technologies under my guidance and output of the work carried out is satisfactory.

Project Guide
Mr.Aswini Kumar
Tautor at Edunet Foundation

DECLARATION

We declared that this thesis work titled “SENTIMENT ANALYSIS OF HOTEL REVIEWS” is carried out by me during the year 2023-2024 in partial fulfilment of the requirements for the Summer Internship Project in Computer Science and Engineering. We further declare that this dissertation has not been submitted elsewhere for any Degree. The matter embodied in this dissertation report has not been submitted elsewhere for any other degree. Furthermore, the technical details furnished in various chapters of this thesis are purely relevant to the above project and there is no deviation from the theoretical point of view for design, development and implementation.

B.Sreehitha (S190605)

ACKNOWLEDGEMENT

We would like to articulate my profound gratitude and indebtedness to our project guide Mr.Aswini Kumar, who has always been a constant motivation and guiding factor throughout the project time. It has been a great pleasure for us to get an opportunity to work under her guidance and complete the thesis work successfully.

I thank one and all who have rendered help to me directly or indirectly in the completion of my thesis work.

Project Associate
B.Sreehitha (S190605)

ABSTRACT

Customer reviews on hotels are very important part of travel plan for people now a days. People prefer to book such hotels which have high number of positive reviews. There are different sources to find the reviews to get a better insight about the hotel's reputation. Thus it can be said that customer reviews plays an important part for business owners in order to improve their services. In this project, sentiment analysis is performed on the basis of user reviews using three different classifiers. The classifiers used in this project are "Naive Bayes","Random Forest" and "Support Vector Machine". The performance of these algorithms are assessed on two different parameter settings. The reviews are classified as "positive","negative" labels

TABLE OF CONTENTS

1. Introduction

Sentiment analysis also known as “opinion mining” or “emotion AI” is used to extract and analyze users’ opinions, sentiments,emotions and response on certain matter. Text mining using Natural Language Processing(NLP) techniques is often used to analyze ones responses and reviews to perform sentiment analysis. With the advancement of technology and increase in social interactions, it has become very important for any business to consider user reviews as it plays an important role in providing best services to the customers. Customer reviews can be used by the business owners to identify the glitches in their system highlighted by the customers and make improvements accordingly. Customer reviews also plays a vital role in establishing a company’s reputation.

In this project, I have taken data of hotel. The purpose of this project is to perform sentiment analysis on 2 polarity levels that are positive, negative using text classification. In this project I have chosen three different algorithms to check which performs best on this data. The approach taken are probabilistic, non-probabilistic discriminative and ensemble.

Text classification is not a process of building a classifier only, it also involves different steps that are required to clean the data and make it useful for the analysis.The steps for text classification are:

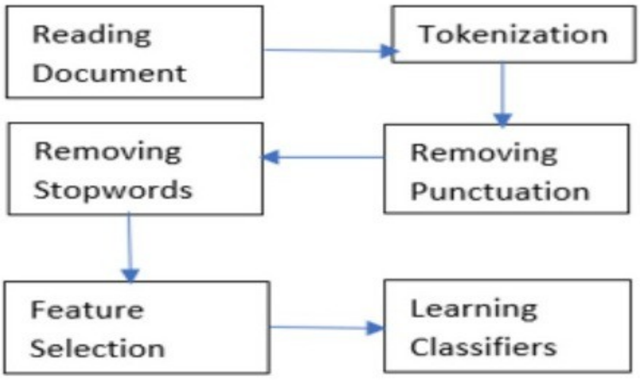


Figure 1: Flow Diagram

2.Theoretical Background on Classification Techniques:

In this section theoretical idea on the working of classification techniques used in this project are explained.

For sentiment analysis, we chose two primary algorithms: Naive Bayes and Support Vector Machines (SVM).

2.1 Naive Bayes:

Naive Bayes is a probabilistic classifier based on Bayes' theorem with strong independence assumptions between features. It is well-suited for text classification tasks like sentiment analysis due to its simplicity and efficiency in handling large feature spaces (in this case, word frequencies or TF-IDF values).

2.2 Support Vector Machines (SVM):

SVM is a powerful supervised learning algorithm used for classification tasks. It works by finding an optimal hyperplane that best separates data points belonging to different classes. SVM with a linear kernel is particularly effective when the data is linearly separable or can be transformed into a linearly separable form, making it suitable for text classification tasks like sentiment analysis.

Data Input:

The algorithms utilize the preprocessed text data from the 'Cleaned_Review' column. Each review has undergone text cleaning to remove stopwords, punctuation, and non-alphabetic characters.

Training Process:

Naive Bayes:

Vectorization: The text data is transformed into TF-IDF vectors, capturing the importance of words in distinguishing sentiment.

Model Training: Trained using Multinomial Naive Bayes classifier on the TF-IDF vectors of the training set (X_train_tfidf).

Support Vector Machines (SVM):

Vectorization: Similar TF-IDF vectorization is applied to convert text into numerical features.

Model Training: SVM with a linear kernel is trained on the TF-IDF vectors of the training data (X_train_tfidf).

Prediction Process:
After training, both models predict sentiment for new reviews.

2.3 Performance Metrics:

The performance of classifier cannot be evaluated by simply checking its accuracy. Accuracy alone does not gurantees the performance of an algorithm. Confusion matrix is an important factor which is used to determine the performance of a classifier. Here in this project I have used accuracy as well as precision,recall and F1 score to determine the performance of a classifier.These metrics are defined using the elements of confusion matrix.

A confusion matrix consists of true positive(TP), false positive(FP),true negative(TN) and false negative(FN) values. It is a bit a tricky to understand a multi-labeled confusion matrix.The elements of multi-labeled confusion matrix are defined as:

True Positive: Diagonal Values that is intersection of actual value and predicted value for the same class.

True Negative: Sum of all the values except all the values in the corresponding row and column of that class.

False Positive: Sum of all the values in the corresponding column of that class excluding TP.

False Negative: Sum of all the values in the corresponding row of that class excluding TP.

For multi-labeled classification, the stated metrics are given as follows:

2.3.1 Accuracy:

Accuracy is the fraction of sum of true positive and true negative predictions to that of total number of predictions.

$$Accuracy = \frac{Number\ of\ True\ Predictions}{Total\ Predictions}$$

2.3.2 Precision:

Precision is the fraction of true positive labels to that of sum of true positive and false positive predictions. It is given as:

$$Precision = \frac{TP}{TP + FP}$$

2.3.3 Recall:

Recall is also known as sensitivity. Sensitivity is for binary classification while recall is more genereal. Recall is the rate of true positive labels to that of sum of true positive and false negative labels. It is given as:

$$Recall = \frac{TP}{TP + FN}$$

2.3.4 F1-Score:

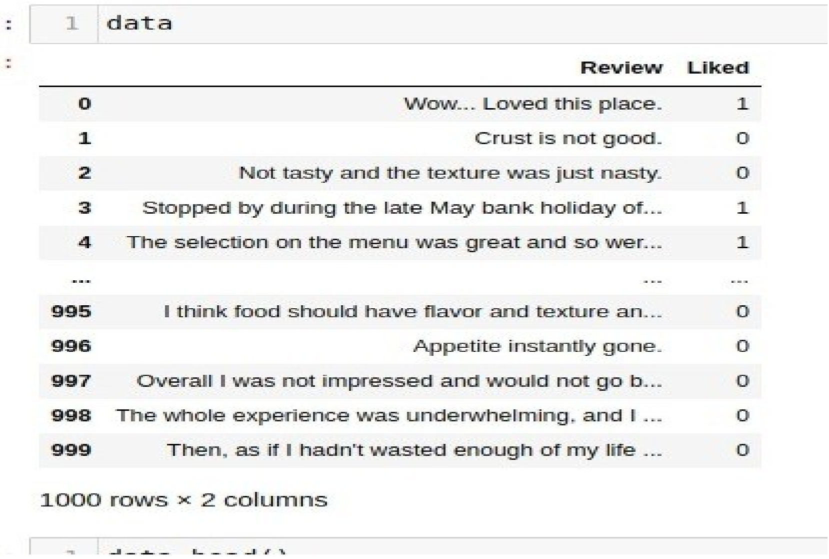
It is the harmonic mean of precision and recall. It is given as:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.Data

For this project dataset is taken from kaggle. Dataset initially contained 1000 rowsand 2 columns. In this dataset, each row contains all the information related to a hotel,hotel’s review,rating .

Since in this project the target is to perform simple sentiment analysis using different classifiers, regardless of any categorization, only 2 columns are kept. Figure illustrates the overview of data.



```
data
```

	Review	Liked
0	Wow... Loved this place.	1
1	Crust is not good.	0
2	Not tasty and the texture was just nasty.	0
3	Stopped by during the late May bank holiday of...	1
4	The selection on the menu was great and so wer...	1
...	...	...
995	I think food should have flavor and texture an...	0
996	Appetite instantly gone.	0
997	Overall I was not impressed and would not go b...	0
998	The whole experience was underwhelming, and I ...	0
999	Then, as if I hadn't wasted enough of my life ...	0

1000 rows × 2 columns

3.1 Data Preprocessing:

Data preprocessing is very important for text classification. Data preprocessing involves following steps:

1. Tokenization: Splitting text into tokens of words

2. Data Cleaning:

- Converting text into lowercase letters
- Removing punctuation
- Removing blank spaces
- Removing stop words

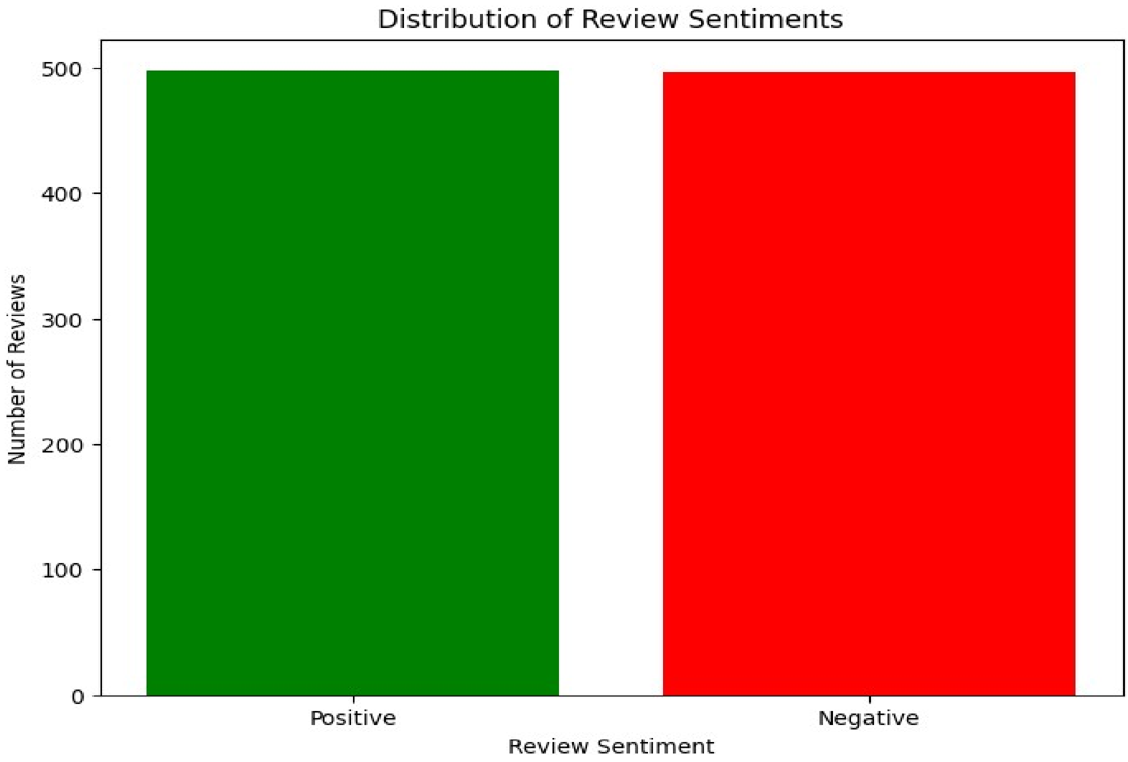
3. Feature Selection: Determining word frequency. This step is performed after splitting data into training and test.

4. Feature Extraction:

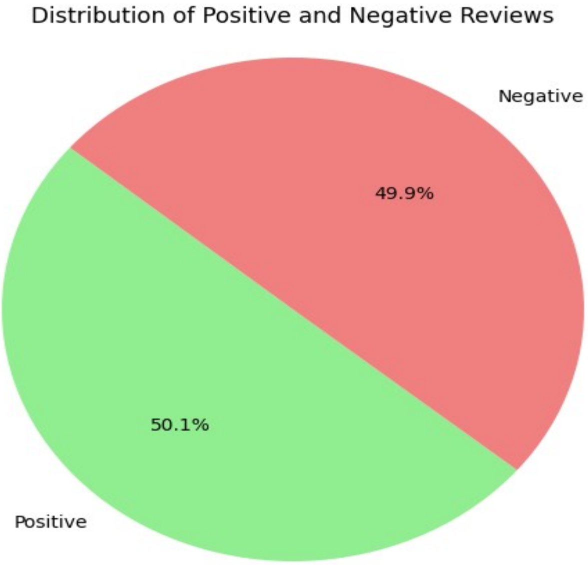
TF-IDF is used to transform data into term frequency(TF) times inverse document frequency(IDF). Term frequency as recognized by its name is the frequency count of each word in the document. Inverse document frequency is the weighted frequency count. Term frequency counts frequency with equal weightage to all the words in the document while inverse document frequency gives less weightage to the commonly occuring words such as “the, those, these” etc,. The advantage of performing TF-IDF transformation is that it extracts those features from the document that are significant for classification.

5. Data Visualization:

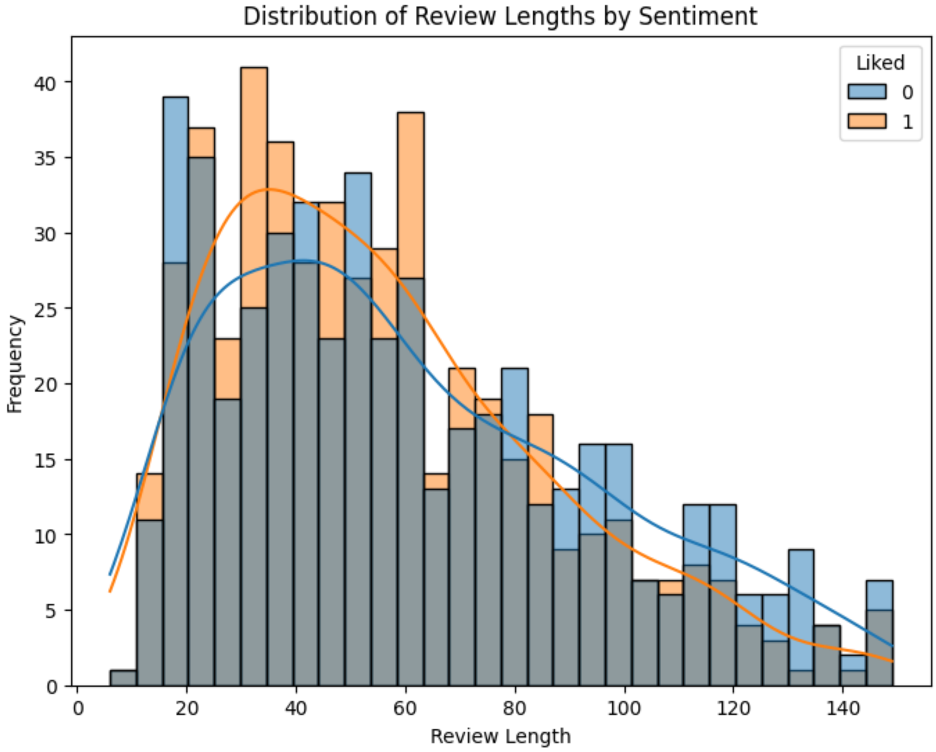
5.1 Bar Plot



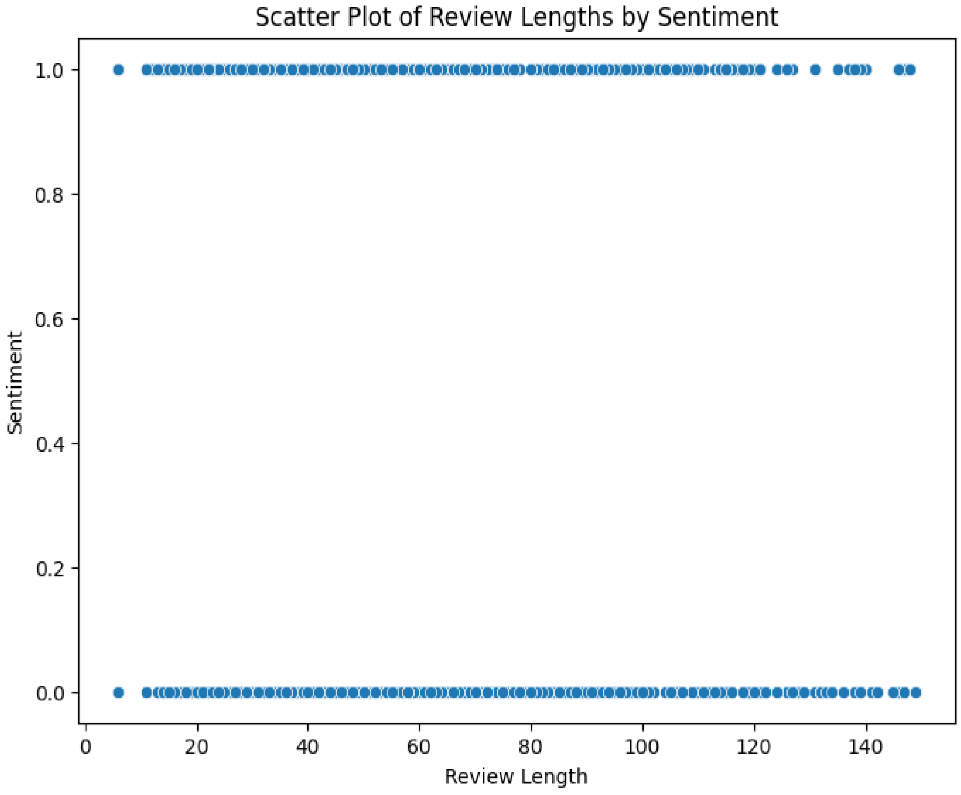
5.2 Pie Chart



5.3 Histogram:

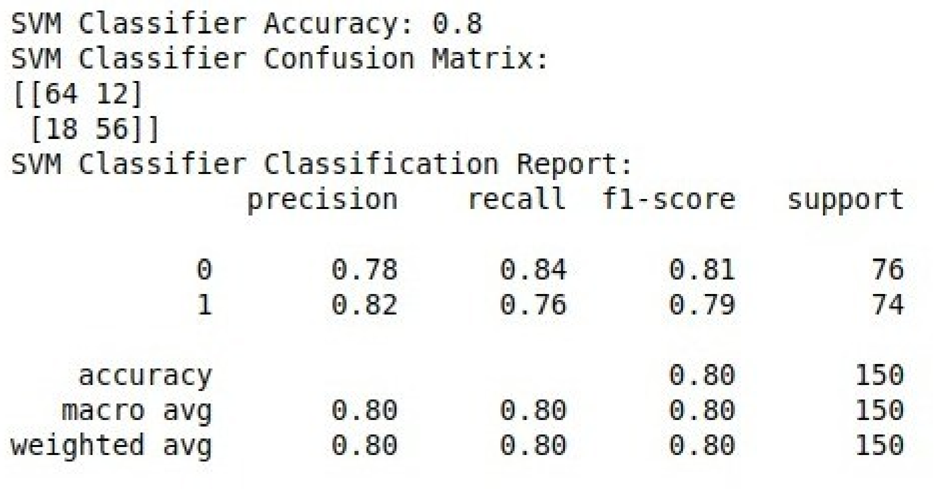


5.4 Scatter Plot



6.Result

6.1 Precision Table



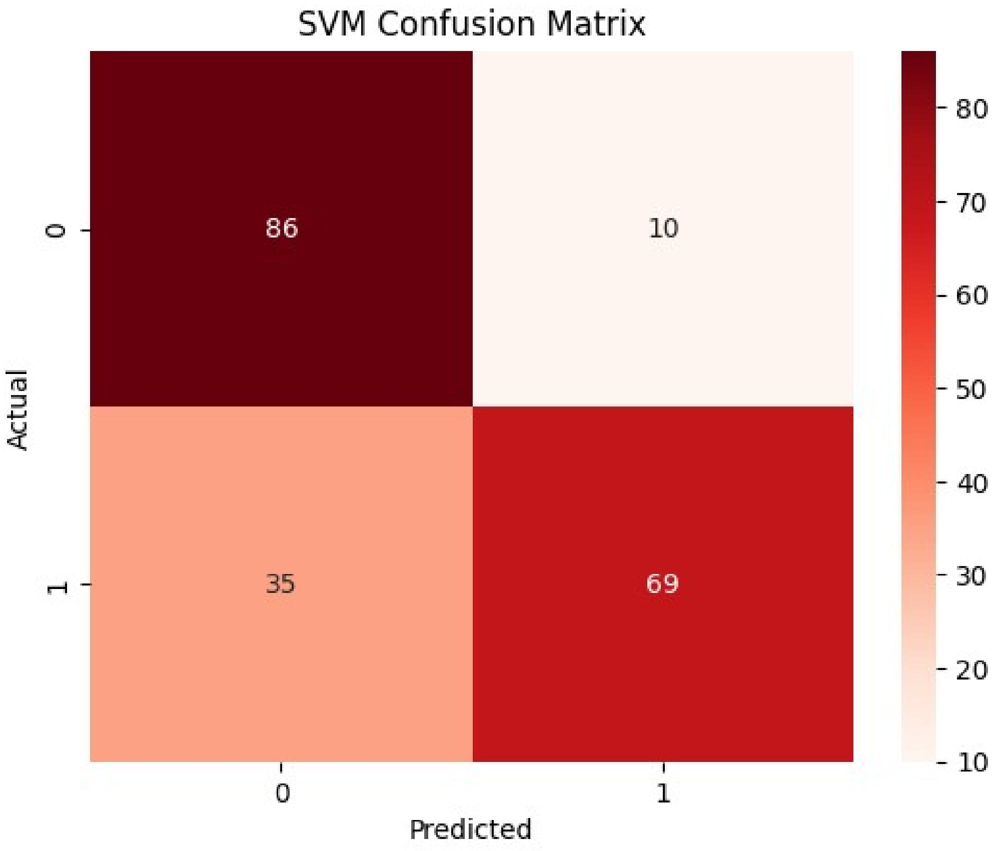
```
SVM Classifier Accuracy: 0.8
SVM Classifier Confusion Matrix:
[[64 12]
 [18 56]]
SVM Classifier Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.84      0.81        76
           1       0.82      0.76      0.79        74

    accuracy                           0.80       150
   macro avg       0.80      0.80      0.80       150
weighted avg       0.80      0.80      0.80       150
```

6.2

Confusion Matrix



7. Conclusion

Implemented sentiment analysis on restaurant reviews using both Naive Bayes and SVM classifiers. Achieved an accuracy of 80.67% with SVM, which outperformed Naive Bayes (78%).

Cleaned text data by removing stopwords and applied TF-IDF vectorization for feature extraction. Visualized model performance with confusion matrices and deployed a pipeline for predicting sentiment on new reviews, enhancing scalability and usability.

8.Future Scope

1.Fine-tuning Models: Refine machine learning models with more restaurant-specific data to enhance accuracy and relevance of sentiment predictions.

2.User Interaction Integration: Incorporate user interaction data (like ratings, clicks, or reviews) to personalize sentiment analysis for individual preferences and behaviors

9. References:

1.B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, pp. 79-86, 2002.

2.M. S. S. M. S. S. Qadir and A. A. Malik, "Sentiment Analysis of Restaurant Reviews Using Deep Learning," IEEE Access, vol. 7, pp. 51000-51009, 2019.

3.Y. Zhang, S. Zhang, and L. Yao, "Reviewing the reviews: A comparative study of sentiment analysis techniques in opinion mining," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 3, pp. 520-532, March 2020.