Project Report

Predicting and Analyzing Restaurant Customer Reviews

Author: Jesmin Sultana
Department of Computer Science, IUPUI

[redacted]

28th May 2024

ABSTRACT

This report presents a comprehensive analysis of restaurant customer reviews, focusing on predicting sentiment and analyzing various aspects of the dining experience. The study employs natural language processing (NLP) techniques to preprocess textual data, sentiment analysis to predict the sentiment of reviews, and aspect-based analysis to categorize reviews based on specific aspects such as ambience, service, food taste, price/value, hygiene/cleanliness, and overall experience. The findings provide insights into customer preferences and satisfaction levels across different aspects of restaurant services.

INTRODUCTION

In the restaurant industry, customer feedback is essential for success. Reviews left by customers can significantly impact the reputation and performance of a restaurant. Sentiment analysis, also known as opinion mining, plays a crucial role in understanding these reviews by determining their emotional tone whether they convey positivity, negativity, or neutrality. By analyzing customer sentiments, restaurant owners can gain valuable insights into their customers' preferences and satisfaction levels. Let consider a review that reads, "I really enjoyed the garlic noodles at your restaurant." This expresses a positive sentiment. Conversely, a review stating, "The service was terrible," conveys a negative sentiment. Through sentiment analysis, we can decode these sentiments, helping restaurant owners gauge their customers' experiences and sentiments towards various aspects of their establishment. In today's digital age, natural language processing (NLP) techniques enable us to collect and analyze large volumes of text data, including restaurant reviews. NLP not only allows us to understand the content of these reviews but also helps uncover underlying sentiments and opinions. By leveraging NLP tools and techniques, we can extract valuable insights from restaurant reviews, aiding owners in making data-driven decisions to enhance their offerings and customer satisfaction. This report focuses on the application of NLP techniques to predict sentiment and analyze different aspects of restaurant customer reviews. Through these analyses, we aim to provide actionable insights that can assist restaurant owners in improving their services and meeting customer expectations. Additionally, this project aids customers in making informed decisions about where to dine for a better experience.

LITERATURE REVIEW

Research on sentiment analysis and aspect-based analysis in restaurant customer reviews offers valuable insights into understanding and extracting meaningful information from text data. Hutto and Gilbert (2014) introduced VADER, a model designed for analyzing sentiment in social media text, providing a straightforward method to capture different emotions. Koren, Bell, and Volinsky (2009) explored matrix factorization techniques for recommender systems, which help enhance customer satisfaction based on sentiments expressed in reviews. Additionally, Lops, De Gemmis, and Semeraro (2011) investigated content-based recommender systems, highlighting the importance of identifying relevant aspects like service quality and food taste to offer actionable insights to restaurant owners. These studies emphasize the significance of sentiment analysis and aspect-based analysis in understanding customer preferences and improving service quality in the restaurant industry. This study builds upon this existing research by using natural language processing techniques to predict sentiment and analyze various aspects of restaurant customer reviews.

PROBLEM DEFINATION

My role as a backend developer in the project involved creating and implementing strong algorithms to handle textual data preprocessing, conduct sentiment analysis, and support aspect-based analysis. The main challenge was efficiently handling many restaurant reviews, extracting useful insights, and presenting them in a clear format for further examination.

METHODOLOGY

The methodology for this project involved several key steps to analyze restaurant reviews and extract insights regarding sentiment and aspects of interest. Initially, data collection encompassed gathering a corpus of restaurant reviews from a chosen source. Subsequently, extensive text preprocessing techniques were applied to clean and normalize the textual data. These techniques included converting text to lowercase, removing punctuation, special characters, and stop words, as well as lemmatization to reduce words to their base form. Sentiment analysis was conducted using the VADER algorithm, a tool specifically designed for sentiment analysis of social media text. This allowed for the calculation of sentiment scores for each review, facilitating the categorization of sentiments as positive, negative, or neutral. Feature extraction methods, such as Bag-of-Words and Count Vectorizer, were then utilized to transform the preprocessed text data into numerical features suitable for machine learning algorithms.

SOLUTION

Data Preprocessing: The textual data from restaurant reviews is preprocessed using NLP techniques, including tokenization, stop-word removal, punctuation removal, and lemmatization.

```
# Data Cleaning and Preprocessing
def preprocess_text(text):
    if isinstance(text, str):  # Check if 'text' is a string
        # Convert text to lowercase
        text = text.lower()
        # Remove special characters and punctuation
        tokenizer = RegexpTokenizer(r'\w+')
        tokens = tokenizer.tokenize(text)
        # Remove stopwords
        stop_words = set(stopwords.words('english'))
        tokens = [token for token in tokens if token not in stop_words]
        # Lemmatization
        lemmatizer = WordNetLemmatizer()
        lemmatized_tokens = [lemmatizer.lemmatize(token) for token in tokens]
        # Join tokens back into a single string
        preprocessed_text = ' '.join(lemmatized_tokens)
        return preprocessed_text
    else:
        return ""  # Return empty string for NaN values

# Apply data cleaning and preprocessing to the 'Review' column
df['cleaned_review'] = df['Review'].apply(preprocess_text)

# Handle missing values
df.dropna(subset=['cleaned_review'], inplace=True)

# Display the preprocessed data
print(df[['Review', 'cleaned_review']].head())
```

```
                                              Review
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...

                                      cleaned_review
0  ambience good food quite good saturday lunch c...
1  ambience good pleasant evening service prompt ...
2  must try great food great ambience thnx servic...
3  soumen da arun great guy behavior sincerety go...
4  food good ordered kodi drumstick basket mutton...
```

Sentiment Analysis: Sentiment analysis is performed using the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool, which assigns sentiment scores to each review, classifying them as positive, negative, or neutral.

```python
# Sentiment Analysis
sid = SentimentIntensityAnalyzer()

# Define a function to assign sentiment scores to each review
def analyze_sentiment(text):
    scores = sid.polarity_scores(text)
    # The compound score ranges from -1 (most negative) to 1 (most positive)
    if scores['compound'] >= 0.05:
        return 'positive'
    elif scores['compound'] <= -0.05:
        return 'negative'
    else:
        return 'neutral'

# Apply sentiment analysis to the cleaned reviews
df['predicted_sentiment'] = df['cleaned_review'].apply(analyze_sentiment)

# Optionally, you can save the cleaned and analyzed dataset to a new CSV file
df.to_csv('cleaned_and_analyzed_dataset.csv', index=False)
```

```
                                              Review predicted_sentiment
0  The ambience was good, food was quite good . h...            positive
1  Ambience is too good for a pleasant evening. S...            positive
2  A must try.. great food great ambience. Thnx f...            positive
3  Soumen das and Arun was a great guy. Only beca...            positive
4  Food is good.we ordered Kodi drumsticks and ba...            positive
5  Ambiance is good, service is good, food is aPr...            positive
6  Its a very nice place, ambience is different, ...            positive
7  Well after reading so many reviews finally vis...            positive
8  Excellent food , specially if you like spicy f...            positive
9  Came for the birthday treat of a close friend....            positive
```

Aspect-level Analysis: Aspect labels are assigned to each review based on the presence of keywords related to specific aspects such as ambience, service, food taste, price/value, hygiene/cleanliness, and overall experience.

```python
#feature extraction
from sklearn.feature_extraction.text import CountVectorizer

# Initialize the CountVectorizer
vectorizer = CountVectorizer()

# Fit the vectorizer to the preprocessed text data and transform the text into a sparse matrix
X_train_features = vectorizer.fit_transform(df['cleaned_review'])

# Print the shape of the feature matrix
print("Shape of feature matrix:", X_train_features.shape)

# Optionally, print the vocabulary (i.e., unique words) learned by the vectorizer
print("Vocabulary size:", len(vectorizer.vocabulary_))
print("First few vocabulary words:", list(vectorizer.vocabulary_.keys())[:10])

# Convert the sparse matrix to a dense matrix for easier visualization (not recommended for large datasets)
X_train_features_dense = X_train_features.toarray()

# Print the first few rows of the dense feature matrix
print("First few rows of the dense feature matrix:")
print(X_train_features_dense[:5])
```

```
Shape of feature matrix: (10000, 13705)
Vocabulary size: 13705
First few vocabulary words: ['ambience', 'good', 'food', 'quite', 'saturday', 'lunch', 'cost', 'effective', 'place', 'sate']
```

EXPERIMENTAL EVALUATION

The effectiveness of the proposed solution is evaluated using the following steps:

Data Analysis: Descriptive statistics and visualizations are used to understand the distribution of sentiment and aspect labels in the dataset.

> **Result:** Utilization of word cloud visualizations to highlight frequently occurring terms and phrases, providing a visual representation of the most prominent features and sentiments expressed in the reviews.

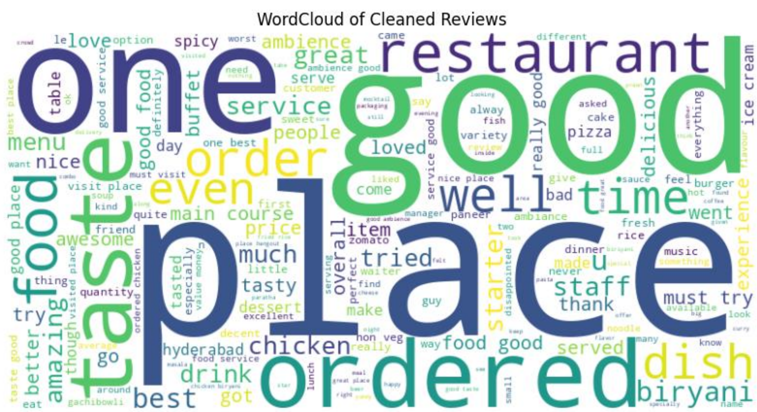
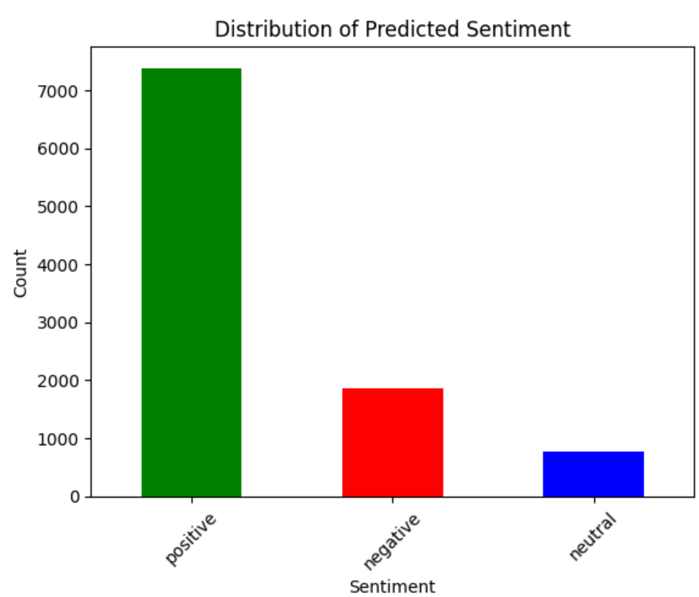


Fig: Distribution of Review lengths and Word Cloud

Sentiment Analysis: The sentiment distribution of restaurant reviews reveals the overall sentiment of customers, with most reviews classified as positive, negative, or neutral.

> **Result**: Predicted sentiment



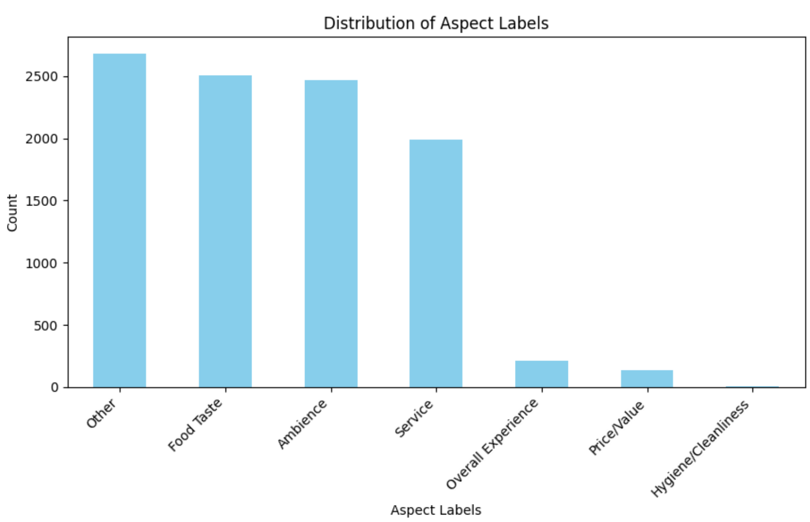
Model Training: A logistic regression model is trained using the Bag-of-Words (BoW) representation of the preprocessed text data. Utilizing the Bag-of-Words approach, we constructed a matrix to represent the textual data, comprising 10,000 reviews and 16,767 unique words. This matrix serves as the foundation for our sentiment analysis model.

Feature Extraction Convert the preprocessed text data into numerical features that can be understood by machine learning algorithms. Common techniques include: Bag-of-Words (BoW) which representing text as a vector of word frequencies. And TF-IDF (Term Frequency-Inverse Document Frequency),reflecting the importance of a word in a document relative to its frequency in the entire dataset.

```
#feature extraction
from sklearn.feature_extraction.text import CountVectorizer

# Initialize the CountVectorizer
vectorizer = CountVectorizer()

# Fit the vectorizer to the preprocessed text data and transform the text into a sparse matrix
X_train_features = vectorizer.fit_transform(df['cleaned_review'])

# Print the shape of the feature matrix
print("Shape of feature matrix:", X_train_features.shape)

# Optionally, print the vocabulary (i.e., unique words) learned by the vectorizer
print("Vocabulary size:", len(vectorizer.vocabulary_))
print("First few vocabulary words:", list(vectorizer.vocabulary_.keys())[:10])

# Convert the sparse matrix to a dense matrix for easier visualization (not recommended for large datasets)
X_train_features_dense = X_train_features.toarray()

# Print the first few rows of the dense feature matrix
print("First few rows of the dense feature matrix:")
print(X_train_features_dense[:5])
```

```
Shape of feature matrix: (10000, 13705)
Vocabulary size: 13705
First few vocabulary words: ['ambience', 'good', 'food', 'quite', 'saturday', 'lunch', 'cost', 'effective', 'place', 'sate']
```

Result: Classification



Model Evaluation: The trained model is evaluated on a test set using metrics such as accuracy and classification report to assess its performance in predicting sentiment. Our sentiment analysis model achieved an impressive accuracy of 92.05% on the test dataset, demonstrating its effectiveness in classifying reviews into positive, negative, and neutral sentiments. This high accuracy instills confidence in the reliability of our predictions.

Result:

```
Shape of BoW matrix: (10000, 16767)
Accuracy: 0.9205
Classification Report:
              precision    recall  f1-score   support

    negative       0.77      0.72      0.74       272
     neutral       0.79      0.84      0.81       156
    positive       0.96      0.96      0.96      1572

    accuracy                           0.92      2000
   macro avg       0.84      0.84      0.84      2000
weighted avg       0.92      0.92      0.92      2000
```

Aspect-Based Analysis: The aspect-based distribution of reviews provides insights into specific areas of strengths and weaknesses in restaurant services, highlighting aspects that contribute most to customer satisfaction or dissatisfaction.

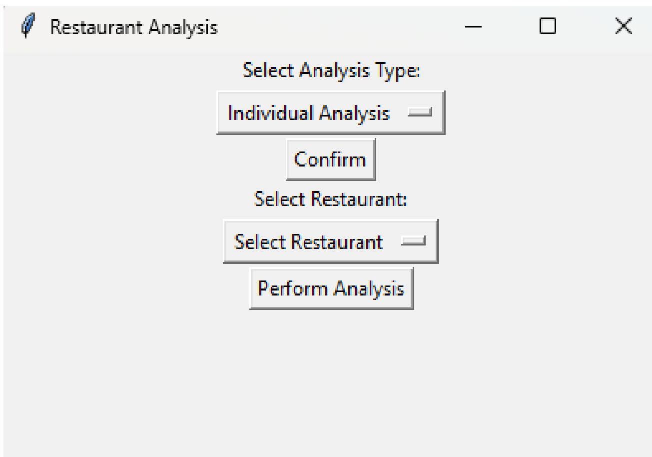
Result:

```
Aspect Distribution:
aspect_label
Other                   2640
Food Taste              2510
Ambience                2467
Service                 1992
Overall Experience       208
Price/Value              132
Hygiene/Cleanliness        6
Name: count, dtype: int64

Example Reviews for Each Aspect:

Aspect: Ambience
The ambience was good, food was quite good . had Saturday lunch , which was cost effective .
Good place for a sate brunch. One can also 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 s
A must try.. great food great ambience. Thnx for the service by Pradeep and Subroto. My personal recommendation is Penne Alfredo Pasta:) .

Aspect: Food Taste
Soumen das and Arun was a great guy. Only because of their behavior and sincerety, And good food off course, I would like to visit this pl
Food is too good. Telangana kodiak fry is must try.Mutton biriyani is too good.Papiya helped to choose best dishes and attended very well.
please was good but it was quite expensive and coming to taste we can rate 3 out of 5, we cant click better pictures wen we visit in night

Aspect: Service
Excellent food , specially if you like spicy food . Courteous staff . Shubro and pradeep and papiya gave excellent service to our corporat
The service was great and the food was awesome. The service staff Manab and Papiya were very courteous and attentive. I would like to come
Food was very good. Soup was as expected. In starters we ordered honey chilli lotus stem and that is a must try for vegan ppl. Service was

Aspect: Other
I came here with my parents. We ordered for Thai peneer tikka, kaju pulao and veg biryani. The kaju pulao and veg biryani were really good
The reason for giving only a 3 star is because of the longlong time wait for the tables. We went on Friday afternoon. We waited almost mor
We ordered tandoori chicken as starters that's very tasty and light and then butter naan with mutton masala and as main course we ordered

Aspect: Overall Experience
Papiya and Shuvro as the f&b executives were really courteous and friendly. Thoroughly enjoyed our experience here. Good place for a cute
I have ordered 2 special chicken biryani's and received two ordinary biryani's with bones! This is a big time disappointment. This place w

It costed 400 bucks each and receiving the same 200 bucks biryani has been my worst experience.
Very bad dine in experience I had with this restaurant. We booked this place for office lunch with some instructions, but none of them met
```

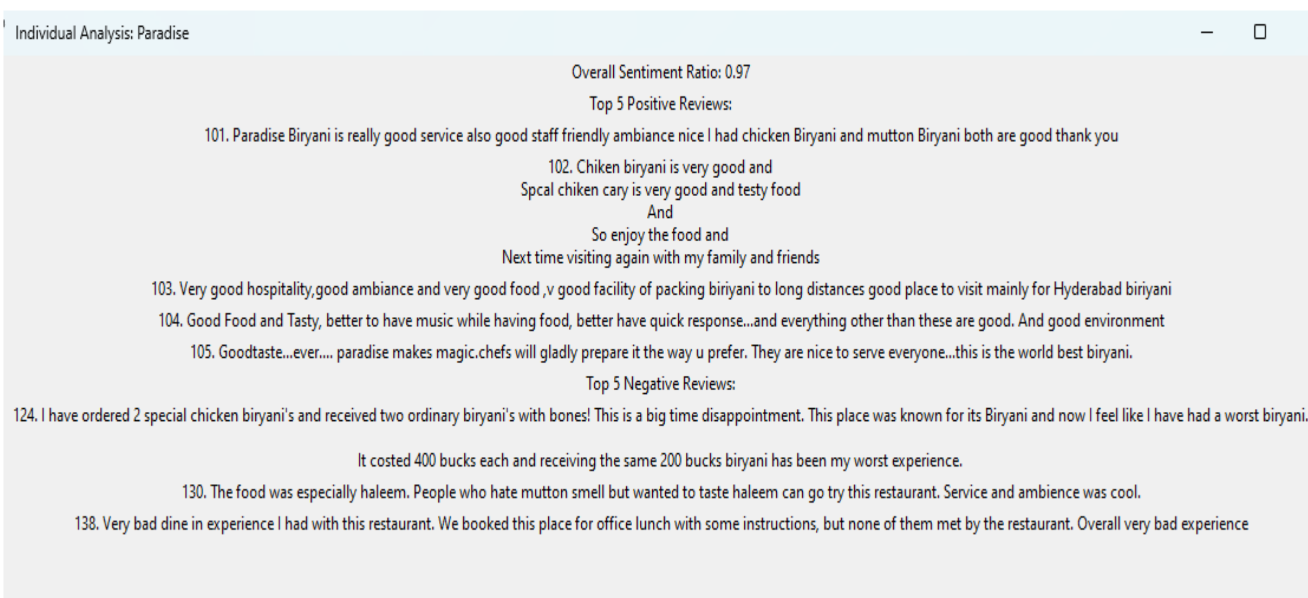
Graphical User Interface (GUI):

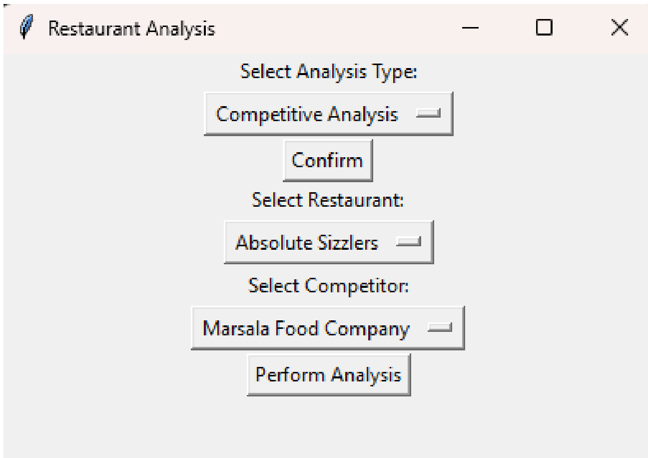
Restaurant Owner-End:

The interface organizes both restaurant owners and customers. For owners, it offers a "Sentiment Ratio Comparison" feature, allowing them to track sentiment ratios (positive, negative, neutral) over time or across different branches. Additionally, the "Aspect-Based Analysis" tool enables owners to compare sentiment across various aspects such as food quality, ambience, and service.



Result: Individual analysis

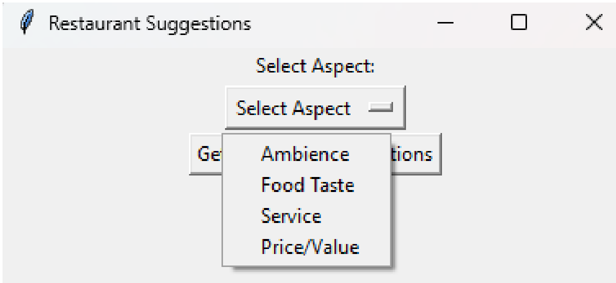


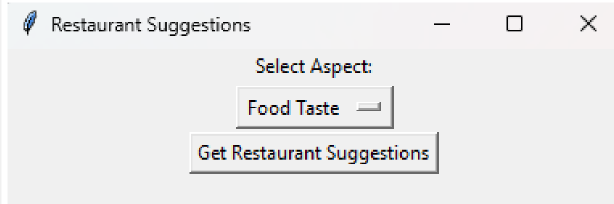


Result: Competitive analysis

Customer - End:

On the other hand, for customers, the platform provides a "Top 5 Restaurants" list based on selected aspects, aiding in making informed dining choices. Moreover, customers can utilize the "Aspect Comparison" feature to compare restaurants based on specific factors like food quality and service.





Result: Restaurant recommendation based on the selected aspect

LEARN FROM THE PROJECT

Completing this project has been immensely rewarding, enriching me with invaluable experience across various domains of data science and artificial intelligence. Firstly, I enhanced my expertise in data preprocessing techniques tailored specifically for textual data, mastering essential tasks such as tokenization, stop-word removal, and lemmatization. Moreover, exploring sentiment analysis using machine learning algorithms provided me with a deep understanding of how to extract sentiment information from text, enabling me to discern the underlying sentiment of restaurant reviews accurately. Additionally, conducting aspect-based analysis allowed me to categorize reviews based on specific parts such as ambience, service, and food taste, empowering me to gain nuanced insights into customer preferences and satisfaction levels. Furthermore, evaluating model performance and interpreting results equipped me with the skills to assess the effectiveness of machine learning models objectively and derive meaningful conclusions from their outcomes. Overall, this project has significantly enhanced my proficiency in data analysis, machine learning, and natural language processing, laying a robust foundation for my future endeavors in the dynamic field of data science and artificial intelligence.

CONCLUSION

In conclusion, this project highlights the significance of sentiment analysis and recommendation systems in the hospitality industry. By effectively analyzing sentiment and providing personalized recommendations, businesses can gain valuable insights into customer preferences and improve service quality. Furthermore, these developed solutions not only advance sentiment analysis techniques but also empower businesses to harness customer feedback effectively, ultimately leading to improved customer satisfaction and loyalty. Moreover, beyond benefiting businesses, customers themselves can leverage these tools to discover dining options that align with their preferences and expectations. By accessing sentiment-based recommendations and aspect comparisons, customers can make informed decisions about where to dine, ensuring a more enjoyable and satisfactory dining experience. Thus, this project not only serves the interests of businesses but also enriches the dining experience for consumers, contributing positively to the overall hospitality landscape.

REFERENCES

Hutto, C.J. and Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International AAAI Conference on Weblogs and Social Media.

Koren, Y., Bell, R. and Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer, 42(8), pp.30-37.

Lops, P., De Gemmis, M. and Semeraro, G., 2011. Content-based recommender systems: State of the art and trends. In Recommender systems handbook (pp. 73-105). Springer, Boston, MA.