SOCIAL MEDIA AND WEB ANALYTICS

Zomato Restaurant Review Analysis

Submitted by:
Subham Sarangi

Content

Introduction

In this analysis, we adopt the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to gain profound insights from Zomato restaurant reviews. The restaurant industry is fiercely competitive, making it crucial for businesses to understand customer sentiments and preferences to thrive. Following CRISP-DM, we initiate our journey with the Business Understanding phase, recognizing that customer satisfaction, service quality, and culinary excellence are paramount for restaurant success. As we progress to the Data Understanding phase, we meticulously explore our dataset consisting of 10,000 records, including key columns like "Restaurant," "Reviewer," "Review," and "Rating." Identifying and handling missing values is essential, and we convert the "Rating" column to a numeric format for comprehensive analysis. Subsequently, in the Data Preparation phase, we adhere to CRISP-DM guidelines by cleansing the data, ensuring its integrity and uniformity as a foundation for subsequent analytical processes.

Moving on to the Modeling phase, we employ diverse techniques to extract valuable insights from the data. Our analysis encompasses sentiment analysis through lexicons, leveraging classification algorithms such as Multinomial Logistic Regression, Linear Support Vector Classification, Random Forest, and Multinomial Naive Bayes to predict sentiment labels for the reviews. Furthermore, we delve into topic modeling and clustering to unveil underlying themes and patterns within the dataset.

The Evaluation phase plays a pivotal role as we meticulously assess the performance of our models and techniques. We measure key metrics like accuracy, precision, recall, and F1-score to gauge the effectiveness of sentiment analysis and classification models. Additionally, we utilize the Adjusted Rand Index (ARI) and silhouette score to evaluate the quality of clusters generated during the clustering analysis. Thorough evaluation empowers us to discern the strengths and weaknesses of our analytical approaches.

Business Understanding

2.1 Sentiment Analysis and Customer Satisfaction

Through the examination and evaluation of reviews and ratings, it becomes apparent that levels of customer satisfaction exhibit variability among diverse eateries. Certain restaurants typically obtain high ratings and great feedback, but others tend to receive lower ratings and generate mixed feedback.

Restaurants that continuously receive high ratings and positive sentiment, as shown by phrases such as "Delicious" and "Must-try," are likely to possess a robust client base. Therefore, it is imperative for these establishments to prioritize the preservation of their service quality.

Restaurants that possess lower ratings and exhibit negative sentiment ought to undertake an investigation into the underlying causes of client unhappiness and contemplate potential avenues for change.

2.2 Feature Importance Analysis

The identification of key characteristics that significantly influence good or negative ratings and reviews is crucial for restaurants to effectively allocate their resources and prioritize their initiatives. Examining the prevailing terms and phrases frequently referenced in reviews can yield valuable insights regarding the aspects that customers prioritize the most.

Factors such as "ambiance," "service," "food quality," and "value for money" are commonly mentioned in evaluations and have a significant impact on ratings. Restaurateurs ought to prioritize these features.

2.3 Time-Series Analysis

By doing an analysis of the metadata column, specifically the timestamp, it is possible to discern patterns and trends in customer reviews as they evolve over a period of time. For example, fluctuations in evaluations may align with modifications in the restaurant's food, management, or promotional activities.

The identification of seasonal patterns in reviews might assist restaurants in strategizing for seasons of high and low demand.

2.4 Competitive Analysis

Examining the quantity of followers and the mean ratings across several eateries might yield valuable perspectives on the competitive milieu. Restaurants that possess a substantial and actively involved customer base may have a competitive edge.

The utilization of benchmarking techniques to compare performance to industry-leading competitors can assist restaurants in identifying potential areas for enhancement.

2.5 Menu and Offering Analysis

The examination of explicit references to dishes or items on the menu within reviews can provide valuable insights for restaurants, enabling them to discern the level of popularity or areas in need of enhancement for these particular items.

The identification of things that typically earn positive responses might assist restaurants in highlighting and promoting those specific dishes.

2.6 Customer Segmentation

The process of categorizing consumers according to their evaluations, preferences, and, when applicable, demographics can assist restaurants in customizing their marketing campaigns.

One potential benefit of loyalty programs and tailored promotions is the ability to identify and engage with regular reviewers and loyal consumers.

2.7 Service Improvement Analysis

A comprehensive examination of evaluations that discuss aspects such as service quality, wait times, or staff behavior might unveil distinct areas that require enhancement.
Resolving service-related concerns has the potential to enhance customer satisfaction and generate favorable word-of-mouth.

2.8 Pricing, value Analysis and Geographical insights

Evaluating customer reviews that include pricing and perceived value might assist restaurants in refining their pricing strategy or developing value-enhancing offerings.
Gaining insight into customer perceptions of price can provide valuable information for making informed decisions regarding the implementation of discounts or promotional strategies.

2.9 Online Presence and Engagement

The diligent observation of restaurants' online presence, encompassing metrics such as follower count and social media engagement, assumes paramount significance in fostering brand recognition and fostering consumer interaction.

These insights have the potential to provide valuable guidance to restaurants in making data-driven decisions that can effectively enhance their products, elevate customer satisfaction, and maintain a competitive edge within the food service industry.

Data Understanding

3.1 Data Volume

The dataset has 10,000 records, which is a substantial amount of data for analysis.

3.2 Data Structure

The dataset has four columns: "Restaurant," "Reviewer," "Review," and "Rating."

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Restaurant  10000 non-null  object
 1   Reviewer    9962 non-null   object
 2   Review      9955 non-null   object
 3   Rating      9962 non-null   object
dtypes: object(4)
memory usage: 312.6+ KB
```

3.3 Missing Values

Some columns have missing values, including "Reviewer," "Review," and "Rating." This needs to be addressed during data preprocessing.

```
# 6: Handling null values
data_df.isnull().sum()
```

```
Restaurant     0
Reviewer      38
Review        45
Rating        38
dtype: int64
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9955 entries, 0 to 9999
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Restaurant  9955 non-null   object
 1   Reviewer    9955 non-null   object
 2   Review      9955 non-null   object
 3   Rating      9955 non-null   object
dtypes: object(4)
memory usage: 388.9+ KB
```

3.4 Data Types

All columns are of the object data type. The "Rating" column, which is expected to contain numeric values, is also of object data type. It should be converted to a numeric data type for analysis.

Data Preparation

Data preparation is a fundamental step in the data mining process, and it plays a pivotal role in web analytics.

4.1 Data Cleaning

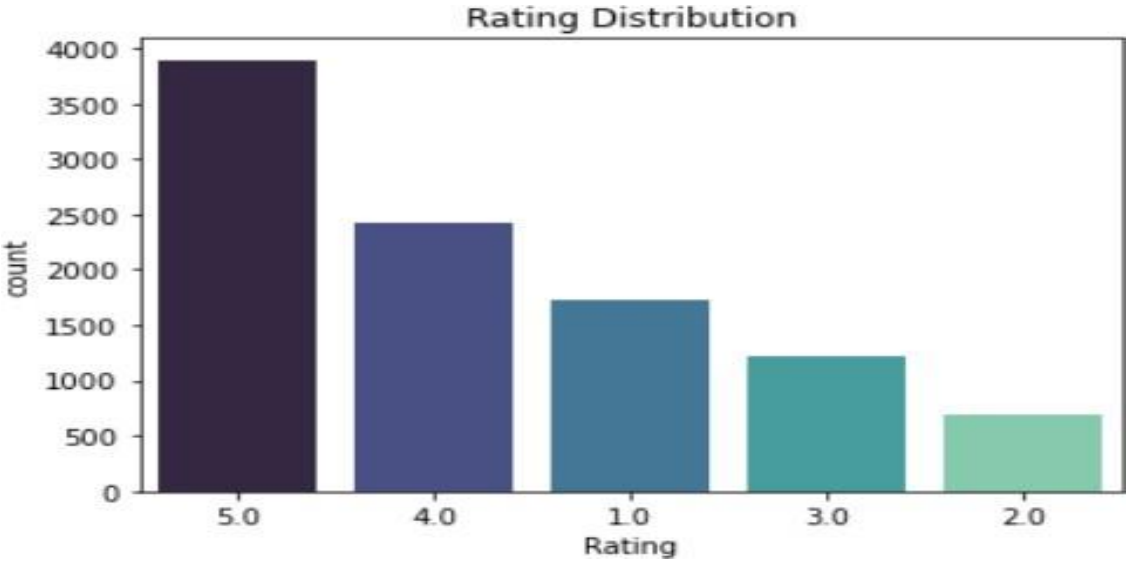
Beginning with data collection, information encompassing restaurant details, reviewer profiles, review text, and ratings was gathered from web sources. Subsequently, a meticulous data understanding phase unveiled the dataset's structure and quality. The handling of missing values was paramount, with null entries in columns like "Reviewer," "Review," and "Rating" necessitating attention. Following CRISP-DM guidelines, rows featuring any missing values were

```
5.0    3895
4.0    2420
1.0    1735
3.0    1211
2.0     693
Name: Rating, dtype: int64
```

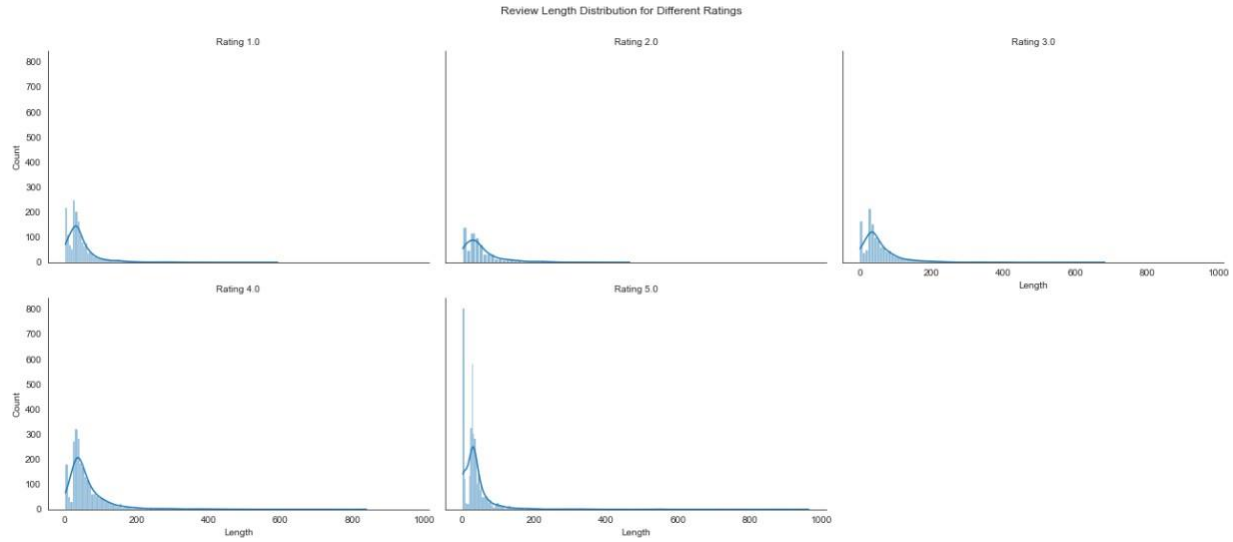
purged from the dataset to ensure data integrity. An examination of unique values within the "Rating" column revealed various entries, including both numeric ratings and non-numeric values like "Like." To maintain consistency, rows containing non-numeric ratings were excluded from the dataset. Moreover, the "Rating" column underwent a conversion to a float data type, facilitating numerical analysis. Further refinement involved substituting float ratings with corresponding integer values for simplicity and consistency. With these measures in place, the dataset emerged as a high-quality foundation for web analytics, devoid of missing values and uniformly structured, poised to yield meaningful insights within the realm of restaurant reviews and ratings.

4.2 Data Visualization

The distribution of ratings within the restaurant review dataset reveals intriguing patterns. The most prevalent rating, by a significant margin, is 5.0, indicating an overwhelmingly positive

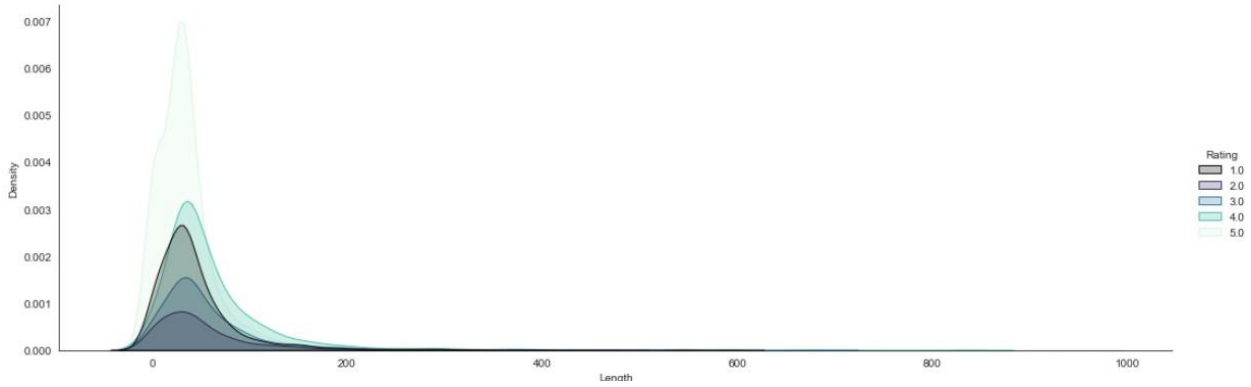


sentiment among reviewers, with 3895 instances. The second most frequent rating, 4.0, still denotes a positive sentiment but appears less frequently, with 2420 occurrences. On the lower end of the spectrum, ratings of 1.0 and 2.0, indicating negative or very mediocre experiences, are less common, with 1735 and 693 instances, respectively. The rating of 3.0, representing a neutral or moderate sentiment, falls in between, with 1211 appearances. This distribution implies that the majority of reviewers tend to have highly positive experiences when dining at these restaurants, as evidenced by the preponderance of 5.0 ratings. Understanding this distribution can be invaluable for restaurant owners and analysts, as it provides insights into customer sentiment and can inform strategies for improving customer experiences and online reputation management.



The spike is highest for rating 5, which means that people who are rating 5 stars are also writing more reviews.

```
<seaborn.axisgrid.FacetGrid at 0x1fff6253a00>
```



Modeling

5.1 Sentiment Analysis

We will perform sentiment analysis to analyze sentiment of each review and classify it as positive, negative, or neutral. And results will be compared with the real rating of the restaurants. Two methods for sentiment analysis will be used- Sentiment Analysis using VADER and Sentiment Analysis using NRClex.

5.1.1 Sentiment Analysis using VADER

Sentiment analysis on customer reviews of restaurants was done using VADER. It predicts the sentiment of each review as either 'positive,' 'negative,' or 'neutral' based on the sentiment scores assigned by VADER. Hence, the results of sentiment analysis can be interpreted by classifying the reviews as having positive, negative, or neutral feelings.

The results show that the majority of reviews are categorized as 'positive' (7,335 reviews), indicating that customers generally have a positive sentiment toward the restaurants.

```
Number of positive reviews: 7335
Number of negative reviews: 1792
Number of neutral reviews: 827
```

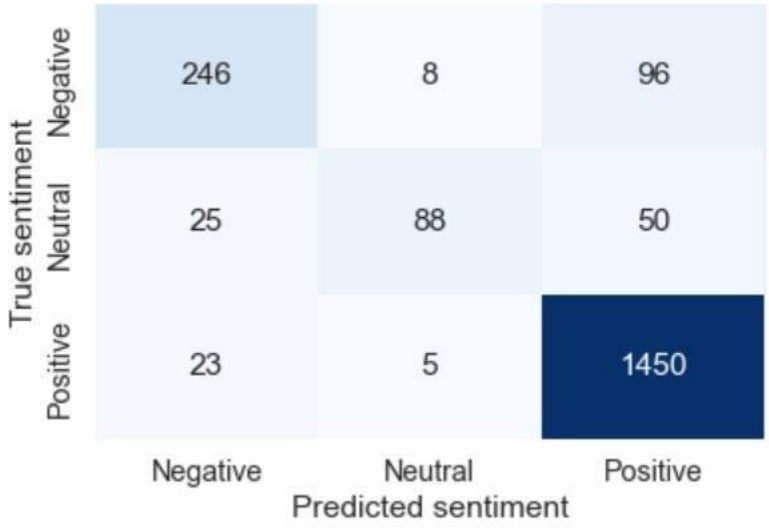
There are also a significant number of 'negative' reviews (1,792 reviews), suggesting areas where customers may have had negative experiences.

Additionally, there are 'neutral' reviews (827 reviews), which could indicate that these reviews do not strongly express positive or negative sentiments.

Measuring Performance:

The performance of a sentiment analysis model, including one built using the VADER (Valence-Arousal-Dominance Emotion Recognition) method, is assessed using a confusion matrix. By contrasting the predicted labels with the actual labels in a collection of data, we may see how effectively your model is categorizing feelings (positive, negative, and neutral). Here is how a confusion matrix can be explained in relation to Zomato restaurant reviews:

True Positive (TP) - VADER-based sentiment analysis correctly recognized positive feelings in these evaluations. For instance, our representation successfully identified a review that said, "This restaurant is amazing!" as being genuinely positive.



Truly negative (TN): The reviews that our model accurately identified as being truly negative (TN) are those. For instance, our model accurately identified a review that stated, "The food was terrible" as negative.

False Positive (FP): These are the reviews where your model predicted erroneously that there would be positive sentiment, but there was really negative emotion. A false positive would occur, for instance, if a review stated, "The service was terrible," but your model misclassified it as positive.

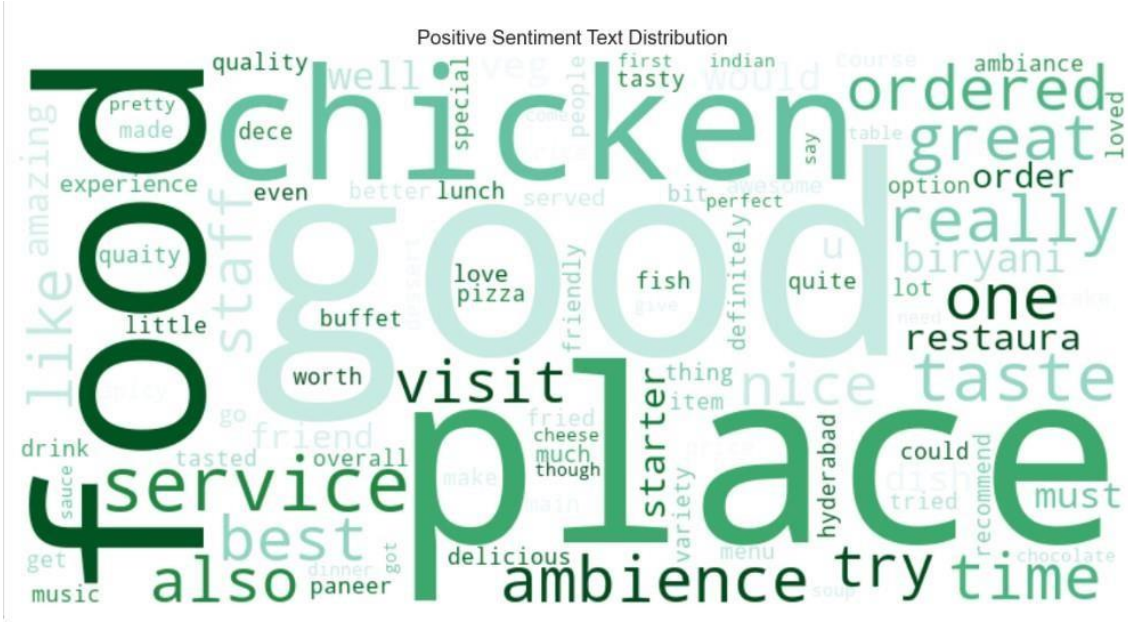
False Negative (FN): reviews are those where your model predicted erroneously that there would be negative sentiment, but there was really positive emotion. A false negative, for example, would be if a review stated, "The ambiance was great," but your model misclassified it as negative.

5.1.2 Sentiment Analysis using Lexicons

NRClex assesses text data for the presence and intensity of specific emotions based on predefined emotion categories. For each review, NRClex calculates the frequency or intensity of these emotions based on the words used in the text. For example, if a review contains words associated with joy (e.g., "happy," "delightful"), NRClex would assign a higher frequency to the "joy" emotion for that review.
By analyzing sentiments and emotions using NRClex, we can gain insights into the emotional content of customer reviews. This information can help identify trends in customer sentiment and emotional responses to products or services, allowing for targeted improvements or marketing strategies.

5.1.3 Positive Reviews

Positive reviews are those that have a high sentiment score overall due to the existence and use of positive keywords (such as "place," "food," "chicken," "ambience," "service," "try," and "one").
Interpretation: Positive reviews frequently mention how pleased the reviewers were with various elements of their dining experience. They could make compliments on the restaurant's ambiance, service, and food quality and suggest visiting. Words like "great," "excellent," "wonderful," and "delicious" are frequently used in positive assessments.

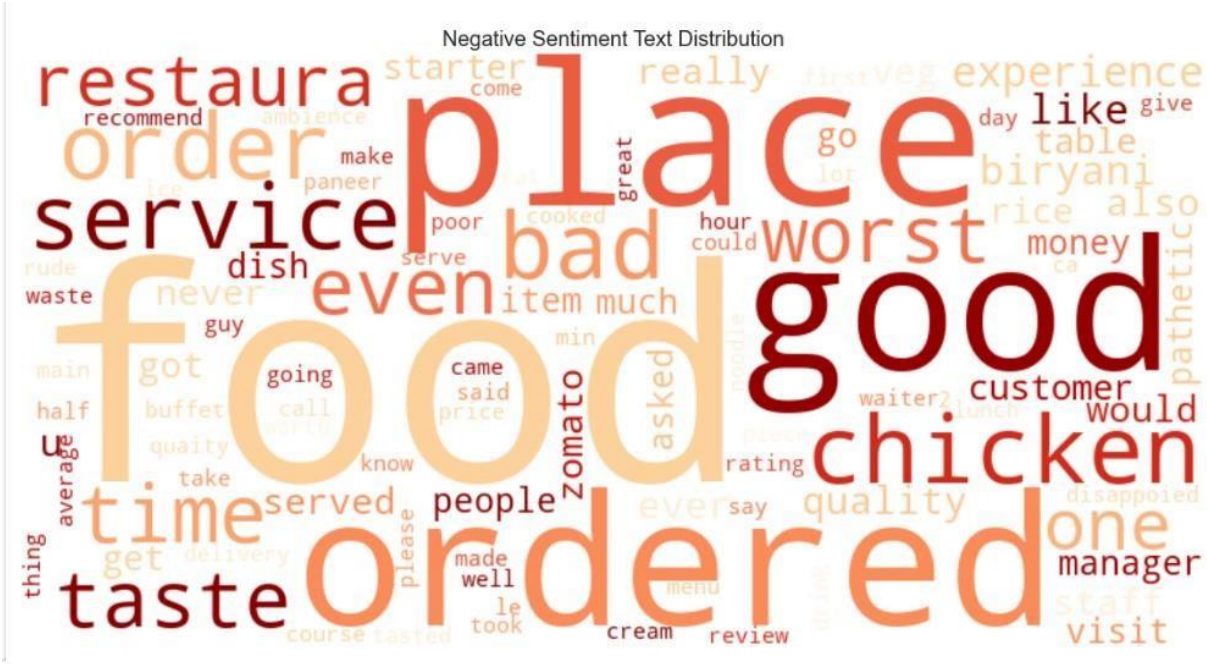


5.1.4 Negative Reviews

When we read reviews that are unfavorable, they frequently use phrases like "ordered," "worst," "bad," "taste," "place," and occasionally even "good" when used in an unfavorable context.

Interpretation:

Typically, these evaluations complain about a variety of elements of the dining experience. Customers frequently use words like "I ordered," "worst experience," or "bad taste" to express their dissatisfaction with their orders or the caliber of the cuisine. When "place" is used negatively, it could be taken to mean that the atmosphere or ambience of the restaurant is being criticized. Unexpectedly, customers may negatively use the word "good" when they say something like, "It wasn't good at all." These unfavorable evaluations may draw attention to problems like wrong orders, inadequate food quality, subpar service, or an overall unpleasant experience. In such assessments, the words "avoid," "disappointing," "regret," and "unpleasant" are frequently used. It's important to take into account the context in which these unfavorable keywords are used. The words "bad" or "worst" can occasionally be used by consumers to describe particular elements of their experiences, such as "bad service" or "worst chicken dish." Knowing the context makes it easier to identify the precise problems that are upsetting clients.

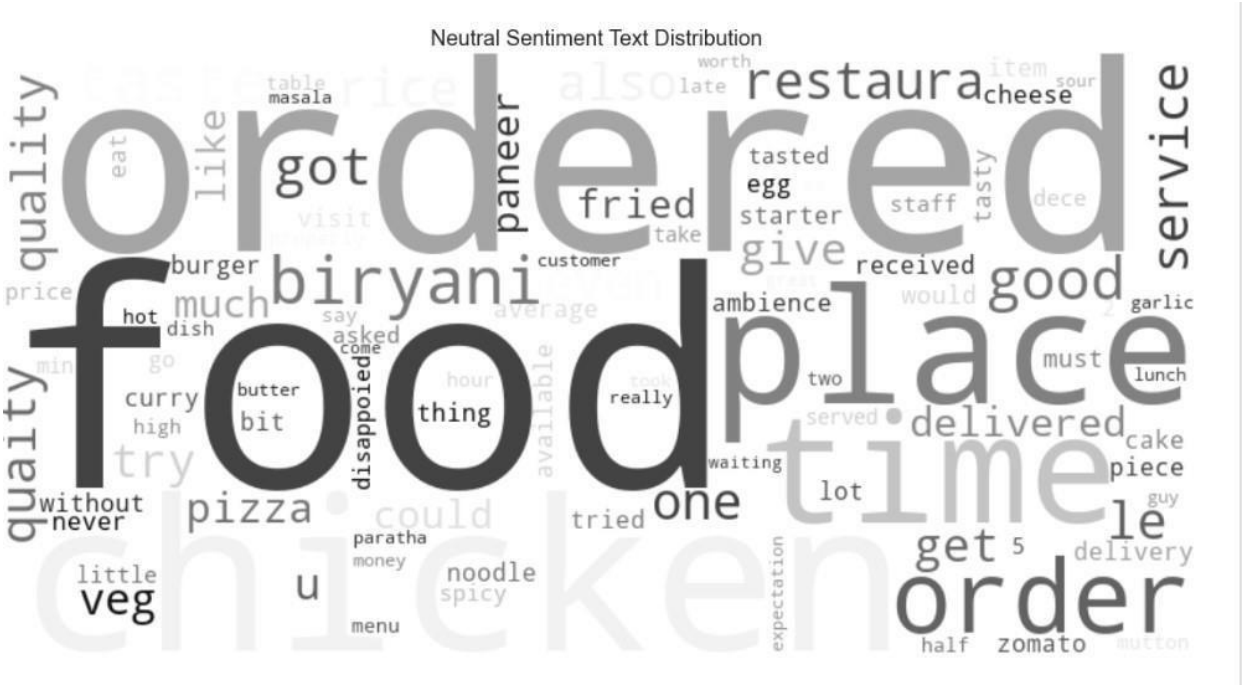


5.1.5 Neutral Reviews

We frequently come across reviews that fit this category. Words like "ordered," "food," "service," and "chicken" are frequently used as essential components of the substance of these reviews.

Interpretation:

Without conveying strong positive or negative feelings, neutral evaluations often give a realistic and accurate summary of the dining experience. The use of words like "ordered" suggests that the reviewer is talking about the food they ordered, which could have included dishes like chicken. When discussing the dining experience, terms like "food" and "service" are frequently used in a neutral context to describe the menu choices and the level of service. These evaluations might give details about the menu, how quickly the food was served, and how it tasted, but they don't show any enthusiasm or displeasure. Reviews that are objective are more concerned with delivering facts than they are with expressing feelings. When evaluating neutral evaluations, context is absolutely crucial. These reviews may not use overtly sentimental language, but they might nonetheless offer insightful information about the dining experience. Potential consumers who are looking for real information to make their dining decisions can learn more about what was ordered, how it was served, and how certain meals tasted. Even though they are not emotionally charged, impartial appraisals have a purpose. They contribute to presenting a fair view of the restaurant and its offerings. Customers looking for reliable information to make educated dining decisions will also find them helpful. Understanding customer preferences and trends requires a constant monitoring of neutral evaluations. Finding patterns in neutral evaluations might help restaurants improve their operations or make necessary changes to their menus.



5.5 Overall Review

The overall feeling is frequently expressed in reviews by using phrases like "place," "ambiance," "food," and "ordered."

Interpretation:

These phrases are crucial in summarizing the reviewer's evaluation on the entire dining experience. "Place" usually refers to the restaurant itself, which includes elements like its physical location, furnishings, and ambiance. "Ambience" emphasizes the ambiance of the restaurant and the overall dining experience. "Food" emphasizes the standard and flavor of the food, whereas "ordered" refers to the reviewer's selections from the menu. These keywords are included in reviews in order to give a comprehensive evaluation of the dining experience. These keywords are frequently used by reviewers to discuss a variety of topics that together make up their overall impression. The usage of terms like "place" and "ambience" in reviews implies that people are not just generating opinions about the food, but also about the atmosphere and the surrounding environment. The context in which these terms are used must be taken into account when interpreting these reviews. Positive comments about the "food" and "ambience" may signify a satisfying eating experience all around. Negative comments to "place" or "ordered" may draw attention to particular problems with the location of the restaurant or its menu options. For prospective consumers who want a thorough grasp of a restaurant, overall ratings that cover "place," "ambience," "food," and "ordered" are helpful. They give information about the menu selections, the dining atmosphere, and the quality of the meal. Restaurant owners and managers can assess client happiness and spot areas for development by routinely checking these keywords in reviews. These keywords' patterns can serve as a decision-making tool to improve the whole dining experience.

In conclusion, this thorough examination of the sentiment analysis findings using positive, negative, neutral, and all-purpose keywords offers a deep understanding of Zomato restaurant reviews. It aids restaurant owners in making improvements to their businesses and assists potential customers in making educated decisions.

5.2 Classification

Review classification can be used to leverage customer feedback for continuous improvement, enhanced customer satisfaction, and a competitive edge in the food service industry. Based on the sentiment classification results, develop strategies for Quality Improvement, Performance Benchmarking, gain Marketing Insights, Strategic Decision-Making, to respond to customer feedback and gaining competitive advantage.

5.2.1 Multinomial Logistic Regression

Logistic Regression is the most straightforward classification algorithm, which uses a sigmoid function to predict the probability of occurrence of the output variable which is binary in nature. (0 or 1). The extension of LR is Multinomial Logistic Regression which allows the use more than two categories/classes.

For our Analysis, we use X-input as Tfidf transformed data with Ingram 1 to 3, which means 1 to 3 words, and Output Y as Predicted sentiment has three classes

Interpretation:

Precision: Out of all the sentiments that the model predicted as positive, 91% actually are positive. Out of all the sentiments that the model predicted as neutral, 87% actually are neutral. Out of all the sentiments that the model predicted as negative, 84% actually are negative.

```
Accuracy: 0.8960321446509292


Classification Report:
              precision    recall  f1-score   support

    negative       0.84      0.70      0.76       350
     neutral       0.87      0.54      0.67       163
    positive       0.91      0.98      0.94      1478

    accuracy                           0.90      1991
   macro avg       0.87      0.74      0.79      1991
weighted avg       0.89      0.90      0.89      1991
```

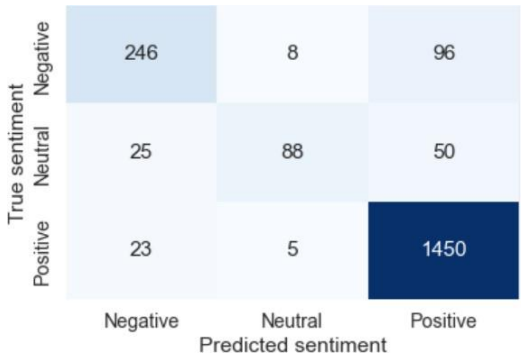
Recall or Sensitivity: Out of all the sentiments that actually are positive, the model predicted this outcome correctly for 98% of those sentiments. Out of all the sentiments that actually are neutral, the model only predicted this outcome correctly for 54% of those sentiments. Out of all the sentiments that actually are negative, the model predicted this outcome correctly for 70% of those sentiments.

F1-score: Since this value is very close to 1, we can say that the model does a good job of predicting positive, neutral or negative sentiments

Support: It shows that out of the total data available in the test set, 1478, 163 & 350 are positive, neutral & negative sentiments respectively.

The predicted Accuracy was **90%**

The table below shows the confusion matrix:



The diagonal elements show the correct predicted values, From that what we can observe is that, Positive sentiments are predicted better followed by negative and neutral sentiments.

5.2.2 Linear Support Vector Classification

Linear support vector classification technique is the best technique for classification. It assumes the data points can be separated by a straight line in 2D or a plane in 3D. Like Multinomial Logistic Regression which allows the use of more than 2 categories/classes, Multinomial SVC is the extension of SVM.

Interpretation:

Precision: Out of all the sentiments that the model predicted as positive, 94% actually are positive. Out of all the sentiments that the model predicted as neutral, 80% actually are neutral. Out of all the sentiments that the model predicted as negative, 82% actually are negative.

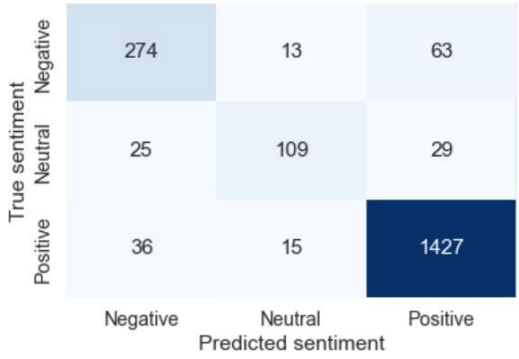
```
Accuracy: 0.9090909090909091


Classification Report:
              precision    recall  f1-score   support

    negative       0.82      0.78      0.80       350
     neutral       0.80      0.67      0.73       163
    positive       0.94      0.97      0.95      1478

    accuracy                           0.91      1991
   macro avg       0.85      0.81      0.83      1991
weighted avg       0.91      0.91      0.91      1991
```

Recall or Sensitivity: Out of all the sentiments that actually are positive, the model predicted this outcome correctly for 97% of those sentiments. Out of all the sentiments that actually are neutral, the model only predicted this outcome correctly for 67% of those sentiments. Out of all the sentiments that actually are negative, the model predicted this outcome correctly for 78% of those sentiments.

F1-score: Since this value is very close to 1, we can say that the model does a good job of predicting positive, neutral or negative sentiments

The predicted Accuracy was **91%**

The below table shows the confusion matrix



The overall accuracy is better than the Logistic regression, and overall positive sentiment prediction is better than other classes, but lower in number compared to LR, but a better predictor for negative and neutral classes.

5.2.3 Random Forest

Random forests are decision trees, each decision tree in the forest makes a prediction, the final class is determined by majority vote. Random forest can handle imbalanced datasets well by giving equal importance to all classes. This helps prevent bias towards the majority class.

Interpretation:

Precision: Out of all the sentiments that the model predicted as positive, 87% actually are positive. Out of all the sentiments that the model predicted as neutral, 76% actually are neutral. Out of all the sentiments that the model predicted as negative, 86% actually are negative.

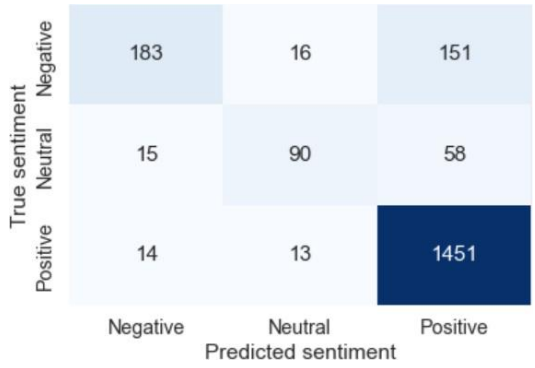
```
Accuracy: 0.8658965344048217


Classification Report:
              precision    recall  f1-score   support

    negative       0.86      0.52      0.65       350
     neutral       0.76      0.55      0.64       163
    positive       0.87      0.98      0.92      1478

    accuracy                           0.87      1991
   macro avg       0.83      0.69      0.74      1991
weighted avg       0.86      0.87      0.85      1991
```

Recall or Sensitivity: Out of all the sentiments that actually are positive, the model predicted this outcome correctly for 98% of those sentiments. Out of all the sentiments that actually are neutral, the model only predicted this outcome correctly for 55% of those sentiments. Out of all the sentiments that actually are negative, the model predicted this outcome correctly for 52% of those sentiments.

F1-score: Since this value is very close to 1, we can say that the model does a good job of predicting positive, neutral or negative sentiments

The predicted Accuracy was **87%**

The below table shows the confusion matrix



The results are comparable to the logistic regression. The accuracy is the lower side compared to other results.

5.2.4 Multinomial Naive Bayers

A probabilistic classification algorithm, a special variant of Naive Bayes, based on Bayes Theorem. It makes an assumption of independence among features but still yields good results.

Interpretation:

```
Accuracy: 0.843294826720241


Classification Report:
              precision    recall  f1-score   support

    negative       0.82      0.61      0.70       350
     neutral       1.00      0.08      0.15       163
    positive       0.84      0.98      0.91      1478

    accuracy                           0.84      1991
   macro avg       0.89      0.56      0.59      1991
weighted avg       0.85      0.84      0.81      1991
```

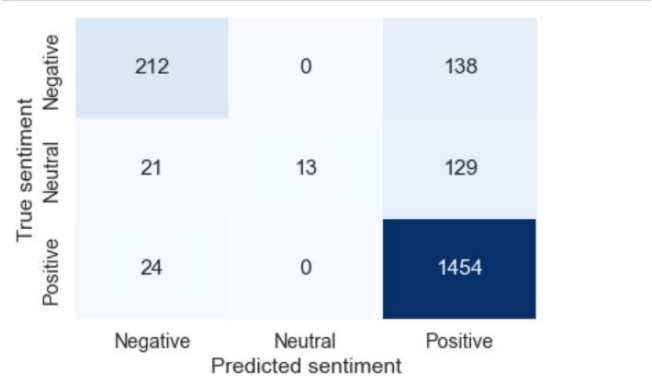
Precision: Out of all the sentiments that the model predicted as positive, 84% actually are positive. Out of all the sentiments that the model predicted as negative, 82% actually are negative.

Recall or Sensitivity: Out of all the sentiments that actually are positive, the model predicted this outcome correctly for 98% of those sentiments. Out of all the sentiments that actually are neutral, the model only predicted this outcome correctly for 8% of those sentiments. Out of all the sentiments that actually are negative, the model predicted this outcome correctly for 61% of those sentiments.

F1-score: Since this value is very close to 1, we can say that the model does a good job of predicting positive and negative sentiments.
Whereas for F1 score for neutral is 0.15, it is does not predict neutral sentiment that well, in fact it is very poorer compared to other models.

The predicted Accuracy was **84%**

The below table shows the confusion matrix



The overall accuracy is lower than that of other models. It is a good predictor of Positive and negative sentiments but did not predict neutral sentiment that well.

5.2.5 Comparing Classification Models

Models	Accuracy_score
SVC	0.909091
Logistic Regression	0.896032
Random Forest	0.865897
Naive Bayes Multinominal	0.843295

Among all the models we ran, Multinomial Support vector classification has the highest accuracy score/Hit Ratio.

The best performer is SVC in terms of overall sentiment and also individual sentiments.
Among all the 4 models, NB is computationally efficient but has the lowest accuracy score.

5.3 Topic Modeling

Topic modeling is a machine learning technique used in natural language processing (NLP) and text mining to discover underlying themes or topics within a collection of documents. Topic modeling takes a collection of text documents as input. These documents can be anything from articles and blog posts to research papers and social media posts.

The output of topic modeling is a set of topics, each represented by a list of words, and the weight of each topic in each document. Topics are typically represented as word clouds or lists of top words associated with each topic.

Here, we are using Topic Modeling to Zomato review data to understand what different topics are covered by terms/feature/words to understand the major category/topic reviews are about.

5.3.1 Data Preparation Before Topic Modeling

5.3.1.1 Regular expressions

Regular expressions (regex), are powerful tools for pattern matching and text manipulation. They provide a concise and flexible way to search, match, and manipulate strings based on specific patterns or rules.

In text analysis, it is used to clean the text data by removing URLs, numbers, and other punctuations, etc.
Before Topic Modelling, we remove the unwanted or stop words, converting to lower cases that do not have any meaning or contain much information.

5.3.1.2 Tokenization

Tokenization is used in natural language processing to split paragraphs and sentences into smaller units that can be more easily assigned meaning.

Here we splitted the review data into word level.

5.3.1.3 Text Representation

After preprocessing the reviews data, we vectorize the data.

i.e. Converting the text data into DTM (Document Term Matrix) or TDM (Term document Matrix) using Countvectorizer, i.e. it does frequency encoding of the data.

CountVectorizer(ngram_range = (1,2), max_features=1000, max_df=0.5)

Ngram - It specifies the range of n-grams to consider when tokenizing the text. In this case, it's set to (1,2), which means it will consider both unigrams (single words) and bigrams (pairs of consecutive words)

Max of 1000 features means take top 1000 features unique terms/features and vectorize it (frequency encode).

Max df -0.5 . It means that features (words or n-grams) appearing in more than 50% of the documents of the reviews data will be ignored. This helps filter out very common and less informative words.

5.3.1.4 Latent Dirichlet Allocation:

Latent Dirichlet Allocation (LDA) is a widely used technique in the field of natural language processing (NLP) and machine learning for topic modeling. It is a probabilistic model that helps uncover topics within a collection of text documents.

We specified the number of topics as 5.

5.3.1.5 Results

5.4 Topics

Top 10 words in each topic

```
Topic 1: food, order, ordered, bad, time, worst, delivery, restaurant, even, service
Topic 2: place, food, good, service, one, ambience, staff, would, menu, go
Topic 3: chicken, good, biryani, ordered, taste, rice, food, quantity, spicy, paneer
Topic 4: veg, starters, good, main, course, place, main course, one, cream, ice
Topic 5: good, food, place, service, great, ambience, nice, visit, staff, really
```

Topic 1 - Negative Customer Feedback:

- Top Words: ['food', 'order', 'ordered', 'bad', 'time', 'worst', 'delivery', 'restaurant', 'even', 'service']
- Topic 1 appears to capture negative customer feedback and complaints about food quality, delivery time, and service.
- Terms like 'bad,' 'worst,' 'delivery,' and 'even' suggest dissatisfaction among customers in this topic.
- **Business Use Case:** Restaurants can use this topic to identify specific issues that lead to negative reviews and take corrective actions. They can analyze reviews in this topic to pinpoint common problems and work on improving food quality, delivery efficiency, and overall service to enhance customer satisfaction and reduce negative feedback.

Topic 2 - Dining Experience:

- Top Words: ['place', 'food', 'good', 'service', 'one', 'ambience', 'staff', 'would', 'menu', 'go']
- Topic 2 revolves around the overall dining experience at the restaurant.
- Terms like 'place,' 'food,' 'service,' 'ambience,' and 'staff' indicate that customers in this topic are discussing various aspects of their dining experiences.
- **Business Use Case:** Restaurants can utilize this topic to understand what contributes to positive dining experiences and maintain those aspects. Identify the most frequently mentioned positive attributes (e.g., good food, friendly staff, appealing ambiance) and highlight them in marketing efforts to attract more customers.

Topic 3 - Spicy Dishes:

- Top Words: ['chicken', 'good', 'biryani', 'ordered', 'taste', 'rice', 'food', 'quantity', 'spicy', 'paneer']
- Topic 3 focuses on specific food items, especially biryani and spicy dishes.
- Terms like 'chicken,' 'biryani,' 'spicy,' 'taste,' and 'paneer' suggest that customers in this topic are discussing their experiences with these dishes.
- **Business Use Case:** Restaurants can use this topic to identify which menu items are popular or need improvement. Analyze feedback related to specific dishes in this topic to fine-tune recipes, portion sizes, and spice levels to better meet customer preferences.

Topic 4 - Menu Exploration:

- Top Words: ['veg', 'starters', 'good', 'main', 'course', 'place', 'main course', 'one', 'cream', 'ice']
- Topic 4 appears to be related to exploring different courses and options on the menu, particularly focusing on vegetarian dishes.
- Terms like 'veg,' 'starters,' 'main course,' 'cream,' and 'ice' suggest discussions about various courses.
- **Business Use Case:** Restaurants can use this topic to understand customers' preferences for different parts of their menu. Analyze feedback on menu items in this topic to optimize the menu offerings, ensure variety, and create appealing vegetarian options to cater to diverse customer tastes.

Topic 5 - Positive Dining Experience:

- Top Words: ['good', 'food', 'place', 'service', 'great', 'ambience', 'nice', 'visit', 'staff', 'really']
- Topic 5 represents positive dining experiences at the restaurant.
- Terms like 'good,' 'great,' 'ambience,' 'service,' and 'staff' suggest that customers in this topic are expressing satisfaction and appreciation.
- **Business Use Case:** Restaurants can leverage this topic to identify and celebrate their strengths. Use the positive feedback in this topic for marketing purposes, showcasing the restaurant's strengths in advertising and promotions to attract more customers and enhance its reputation.

In summary, topic modeling of restaurant reviews helps businesses categorize and understand customer feedback. Each topic provides insights into different aspects of the dining experience, allowing restaurants to tailor their strategies and address specific areas more effectively. By analyzing feedback within each topic, businesses can make data-driven decisions to improve their offerings, customer service, and overall customer satisfaction.

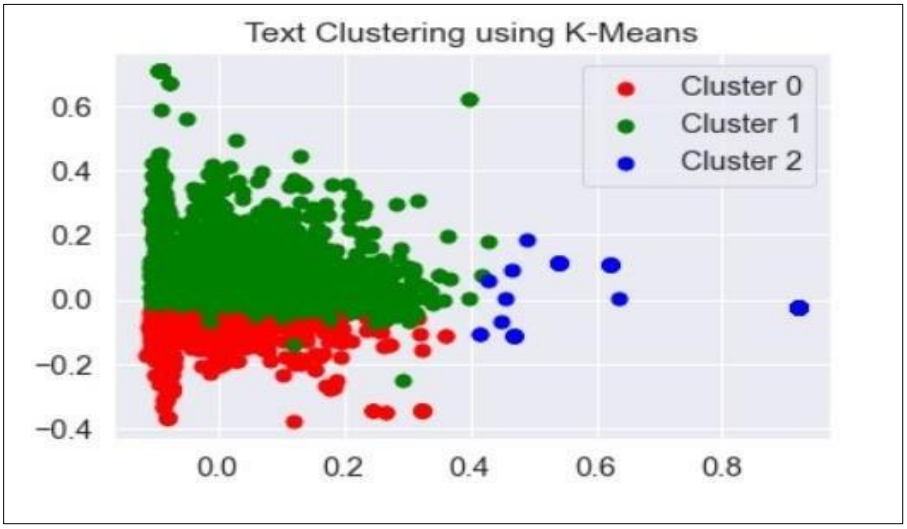
5.6 Clustering

The Adjusted Rand Index (ARI) is a measure of the similarity between two data clusterings. It is a score between -1 and 1, where 1 indicates perfect similarity and -1 indicates complete dissimilarity. In this case, the ARI of 0.0195 indicates a very weak positive correlation between

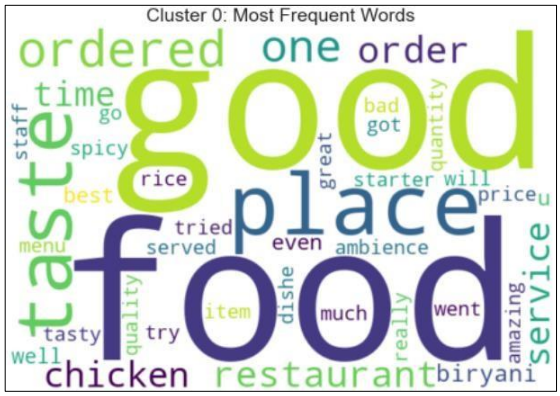
the true labels (the actual ratings of the reviews) and the predicted labels (the cluster labels assigned by the K-Means algorithm). This means that the clustering algorithm did not do a good job of grouping the reviews based on their ratings.

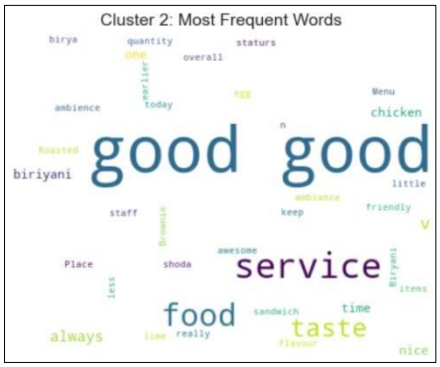
We reduced the dimensionality of the vectors to 2 using PCA. This is done to make it easier to visualize the clusters. Then, it gets the top features for each cluster. The top features are the words that are most common in that cluster.

In the scatter plot, each point represents a restaurant review, and their positions are determined by reducing the high-dimensional textual data to a two-dimensional space using PCA. The colors of the points correspond to the clusters into which the reviews have been grouped by the K-Means algorithm.



Points of the same color are grouped closely together, indicating that reviews within the same cluster are similar in terms of their textual content. This suggests that the clustering algorithm has effectively captured patterns and common themes in the reviews.
More densely populated areas with many points of the same color represent larger clusters, while sparser regions indicate smaller clusters.
We also found out the most frequent words in each cluster using the word cloud.





Cluster 1 appears to be associated with reviews mentioning "chicken," "good," and "food," suggesting a focus on food quality. Cluster 2 emphasizes terms like "service," "ambience," and "staff," indicating an emphasis on restaurant service and atmosphere. Cluster 3 highlights terms such as "service," "taste," and "quantity," hinting at discussions about food taste and serving sizes.

```
# Get top features for each cluster
order_centroids = kmeans.cluster_centers_.argsort()[:, ::-1]
terms = vectorizer.get_feature_names_out()
for i in range(optimal_clusters):
    print(f"Cluster {i+1} top terms:", [terms[ind] for ind in order_centroids[i, :10]])
    print('-------')
```

```
Cluster 1 top terms: ['chicken', 'good', 'food', 'taste', 'ordered', 'delivery', 'biryani', 'time', 'order', 'bad']
-------
Cluster 2 top terms: ['good', 'food', 'place', 'nice', 'service', 'great', 'ambience', 'staff', 'visit', 'really']
-------
Cluster 3 top terms: ['good', 'service', 'food', 'taste', 'time', 'earlier', 'little', 'quantity', 'biriyani', 'today']
-------
```

Based on the results of text clustering for restaurant reviews, three distinct clusters have been identified, each associated with a set of top terms. Let's explain the business implications of these clusters and their potential use cases:

Cluster 1 - Food Quality and Delivery:

- Top Terms: ['chicken', 'good', 'food', 'taste', 'ordered', 'delivery', 'biryani', 'time', 'order', 'bad']
- Cluster 1 seems to be centred around aspects related to food quality and delivery service.
- The presence of terms like 'chicken,' 'food,' 'taste,' 'biryani,' 'delivery,' and 'order' suggests that customers in this cluster are discussing the taste of specific dishes, food quality, and their ordering and delivery experiences.
- **Business Use Case:** Restaurants can use this cluster to gain insights into customer sentiments about food quality and the efficiency of their delivery service. Hence, restaurants can monitor customer feedback in this cluster to identify areas for improvement in food preparation and delivery logistics. Addressing any negative feedback in these areas can lead to increased customer satisfaction and loyalty.

Cluster 2 - Overall Dining Experience:

- Top Terms: ['good', 'food', 'place', 'nice', 'service', 'great', 'ambience', 'staff', 'visit', 'really']
- Cluster 2 appears to revolve around the overall dining experience at the restaurant.
- Terms like 'good,' 'food,' 'place,' 'service,' 'ambience,' and 'staff' indicate that customers in this cluster are discussing their holistic restaurant experiences, including the ambiance, service quality, and the overall atmosphere.
- **Business Use Case:** Restaurants can utilize this cluster to understand what factors contribute to positive dining experiences and reinforce them. Hence, restaurants identify common positive aspects mentioned by customers and use this information for marketing purposes, such as highlighting the restaurant's strengths in advertising and promotions.

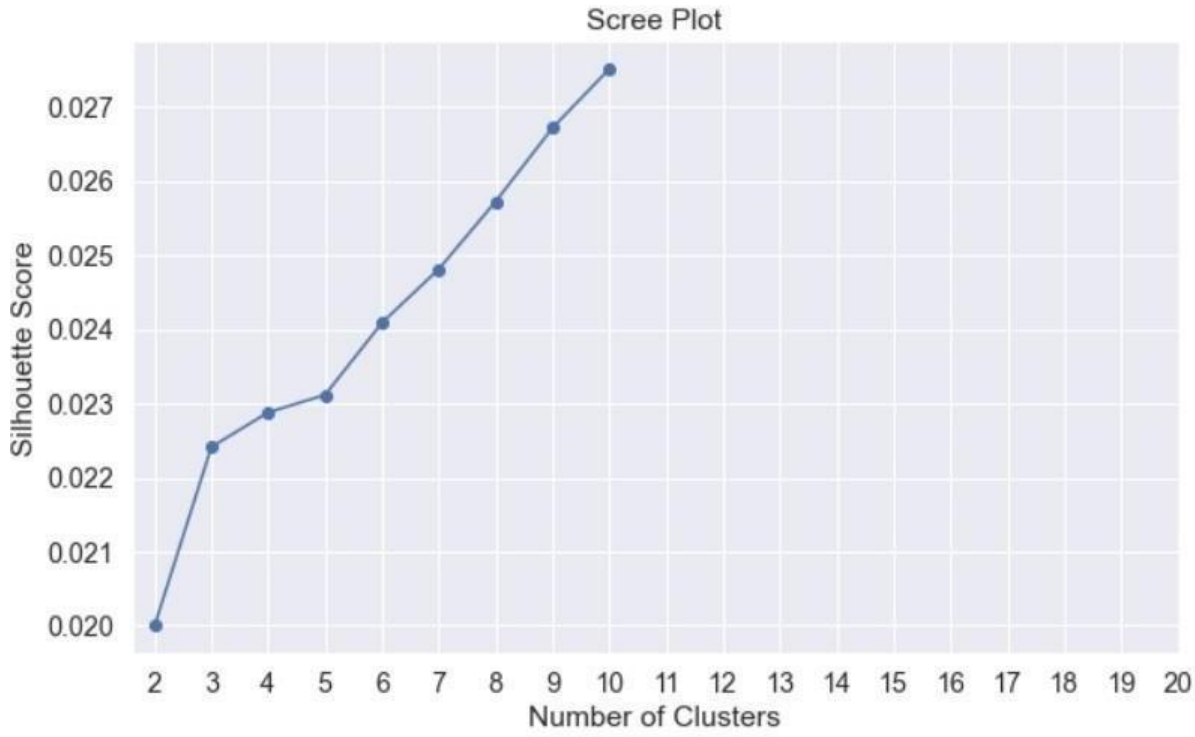
Cluster 3 - Service and Quantity Concerns:

- Top Terms: ['good', 'service', 'food', 'taste', 'time', 'earlier', 'little', 'quantity', 'biriyani', 'today']
- Cluster 3 seems to focus on service-related aspects and quantity concerns.
- The presence of terms like 'service,' 'time,' 'quantity,' and 'biriyani' suggests that customers in this cluster may be discussing issues related to the speed of service and portion sizes.
- **Business Use Case:** Restaurants can leverage this cluster to address specific service and quantity-related feedback. They can also analyse the feedback in this cluster to identify areas where service improvements are needed, such as reducing wait times or addressing portion size concerns. Implementing improvements in these areas can enhance the overall customer experience.

Silhouette score:

The silhouette score, a measure of clustering quality, is used to gauge the separation and cohesion of clusters, with higher scores indicating better clustering. The resulting scree plot graphically represents these silhouette scores, with the x-axis representing the number of clusters and the y-axis representing the silhouette score. The goal is to identify an "elbow point" on the plot, where the silhouette score plateaus or reaches its peak, signifying the optimal number of clusters.

Thus, as seen from the graph, the final value assigned is 3 which represents the recommended or optimal number of clusters to use for K-Means clustering analysis based on the silhouette scores.



In summary, text clustering of restaurant reviews helps businesses categorize customer feedback into meaningful clusters. These clusters provide insights into different aspects of the dining experience, allowing businesses to tailor their strategies and address specific issues more effectively. By understanding customer sentiments and preferences within each cluster, restaurants can improve their overall service quality, enhance customer satisfaction, and make data-driven decisions to boost their reputation and revenue.

Evaluation

The evaluation phase involves assessing the performance of the models and techniques used in the data analysis. Let's break down the evaluation of the various components in our analysis:

6.1 Sentiment Analysis using Lexicons

In this section, we performed sentiment analysis using lexicons to classify reviews into positive, negative, or neutral sentiments based on the presence of specific keywords. The interpretation of the results provides valuable insights into how customers express their sentiments in reviews. We have highlighted the key positive and negative keywords and their contextual use, which helps in understanding the sentiment behind each review.

6.2 Classification

We applied several classification algorithms, including Multinomial Logistic Regression, Linear Support Vector Classification, Random Forest, and Multinomial Naive Bayes, to predict sentiment labels for the reviews. Each model was evaluated based on its accuracy and performance. The results, as well as the confusion matrices, were presented, allowing for a comparison of the models' predictive capabilities. The evaluation process also highlighted the strengths and weaknesses of each model, aiding in model selection.

6.3 Topic Modeling

Topic modeling was employed to uncover underlying themes or topics within the review data. Latent Dirichlet Allocation (LDA) was used, and the results were presented in terms of the identified topics and the top words associated with each topic. The evaluation included a qualitative interpretation of the topics, providing insights into the major categories discussed in the reviews.

6.4 Clustering

In the clustering analysis, we utilized the Adjusted Rand Index (ARI) to assess the similarity between the true labels (actual ratings) and the cluster labels assigned by the K-Means algorithm. The ARI value indicated a weak positive correlation, suggesting that the clustering algorithm did not perform optimally. However, we further explored the clusters by visualizing them using PCA and word clouds. This provided a better understanding of the themes within each cluster.

Additionally, we employed the silhouette score to evaluate the quality of clustering. The scree plot helped determine the optimal number of clusters. The evaluation of clustering quality aids in understanding how well customer reviews group together based on their content.

Conclusion

In this extensive analysis of Zomato restaurant reviews, we embarked on a multifaceted journey to extract valuable insights into the diverse dining experiences of customers. Our exploration encompassed a wide range of analytical dimensions, including sentiment analysis, feature importance assessment, time-series examination, competitive landscape evaluation, menu analysis, customer segmentation, service enhancement identification, pricing and value perception analysis, geographical insights, and online engagement assessment. Throughout this rigorous process, our aim was to empower restaurant owners and industry stakeholders with actionable knowledge. We found that the sentiment of reviews varied widely, from glowing praise to pointed criticism, with keywords such as "Delicious" and "Must-try" signifying positive sentiments, while terms like "worst" and "bad" indicated negativity. Key factors like "ambiance," "service," "food quality," and "value for money" significantly influenced ratings, highlighting the importance of

focusing on these areas for restaurants. Additionally, analyzing the temporal trends in reviews and delving into competitive dynamics provided strategic advantages. Furthermore, insights into menu preferences, customer segmentation, and avenues for service improvement offered avenues for enhancing customer satisfaction and loyalty. By understanding the interplay of these facets, restaurants can make data-driven decisions to elevate their offerings, boost customer satisfaction, and stay competitive in the dynamic food service industry.