Zomato Restaurant Clustering and Sentiment Analysis

Vinayak Marathe, Riya Patel

Abstract - The Zomato Restaurant Clustering and Sentiment Analysis ML unsupervised learning project aims to analyze the restaurant data for each city in India, focusing on customers and the company. The project uses data visualization to cluster restaurants into different segments, providing valuable insights for customers to find the best restaurants in their locality and for the company to identify areas for growth. The sentiment analysis of customer reviews helps to identify the strengths and weaknesses of restaurants, while metadata analysis helps to identify industry critics. The project leverages unsupervised learning techniques to gain insights into the Indian food industry, highlighting the diversity of multi-cuisine available in a large number of restaurants and hotel resorts, and the growing popularity of restaurant food in India.

Keywords: Zomato, Restaurant data analysis, Sentiment analysis, Clustering, Unsupervised learning, Indian food industry, Multi-cuisine, Data visualization, Customer insights, Business growth, Cost vs. benefit analysis, Metadata analysis

I. INTRODUCTION

Zomato is an Indian restaurant aggregator and food delivery start-up founded by Deepinder Goyal and Pankaj Chaddah in 2008. Zomato provides information, menus and user-reviews of restaurants, and also has food delivery options from partner restaurants in select cities. India is quite famous for its diverse multi cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity. Restaurant business in India is always evolving. More Indians are warming up to the idea of eating restaurant food whether by dining outside or getting food delivered. The growing number of restaurants in every state of India has been a motivation to inspect the data to get some insights, interesting facts and figures about the Indian food industry in each city.

The Project focuses on Customers and Company, you have to analyze the sentiments of the reviews given by the customer in the data and make some useful conclusions in the form of Visualizations. Also, cluster the zomato restaurants into different segments: The data is visualized as it becomes easy to analyze data at instant The Analysis also solves some of the business cases that can directly help the customers finding the best restaurant in their locality and for the company to grow up and work on the fields, they are currently lagging in. This could help in clustering the restaurants into segments. Also, the data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis Data could be used for sentiment analysis. Also, the metadata of reviewers can be used for identifying the critics in the industry.

II. PROBLEM STATEMENT

The problem statement of the Zomato Restaurant Clustering and Sentiment Analysis project is to analyze the Zomato restaurant data for each city in India, with a focus on customers and the company. The aim is to cluster restaurants into different segments using unsupervised learning techniques and analyze the sentiments of the customer reviews to provide valuable insights for customers to find the best restaurants in their locality and for the company to identify areas for growth. The project also aims to perform a cost vs. benefit analysis to gain insights. The problem statement addresses the need for data-driven decision making in the restaurant industry in India, where the number of restaurants is growing rapidly, and the demand for restaurant food is increasing among customers.

III. METHODOLOGY

The proposed methodology's implementation begins with downloading the dataset. Then data wrangling and feature manipulation is executed as a step of pre-processing of data. After this, the data is analyzed and a different model is executed. Then we have done the hypothesis testing with 3 different hypothetical statements. At last, all the business insights carried out in this project.

A. Project Datasets

The project begins by understanding the variables in the restaurant and review data frames. The restaurant data frame has 105 rows and 6 columns, with information about restaurant names, links, cost, collections, cuisines, and timings. The review data frame has 10000 rows and 7 columns, with information about the restaurant reviews, such as restaurant name, reviewer name, review, rating, metadata, time, and pictures. The data wrangling process is then performed to find meaningful insights from the data:

Restaurant Dataset:

- **Name**: Name of Restaurants
- **Links**: URL Links of Restaurants
- **Cost**: Per person estimated cost of dining
- **Collections**: Tagging of Restaurants with respect to Zomato categories
- **Cuisines**: Cuisines served by restaurants
- **Timings**: Restaurant timings

Reviews Dataset:

- **Reviewer**: Name of the reviewer
- **Review**: Review text
- **Rating**: Rating provided
- **Metadata**: Reviewer Metadata- No of reviews and followers
- **Time**: Date and Time of Review
- **Pictures**: Number of pictures posted with Review

B. Data Wrangling & Visualization

The quality of data performs a fundamental role, and the most cautiously depicted trouble to be. For this research, as a part of data pre-processing, we have identified the irrelevant or redundant data: such as duplicate observations, entries with missing or invalid values, or data that is outside of the scope of the project and worked on it pretty well by ensuring that all data is in the correct format. Data visualization is used to present the findings in the form of ten different charts, including univariate, bivariate, and multivariate visualizations

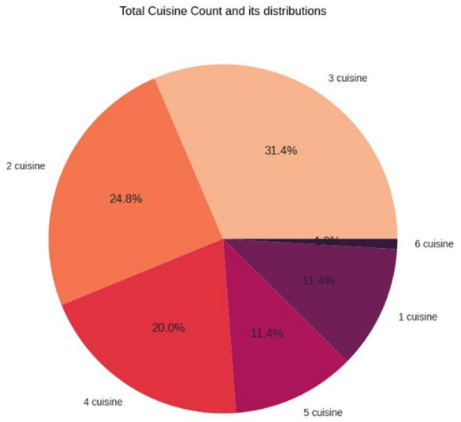

Fig.1 Restaurant Segmentation Using Cuisine Count

Above visualizations shows us

33 restaurants with 3 different cuisines
26 restaurants with 2 different cuisines
21 restaurants with 4 different cuisines
12 restaurants with 5 different cuisines
12 restaurants with only 1 cuisine
Only 1 restaurant with 6 different cuisines

Word clouds can be an important tool for gaining insights into customer reviews.

Fig.2 Most Frequently occurring words in Zomato restaurant Review

Fig.2 shows sentiment analysis on customer reviews. By analyzing the most frequently occurring positive and negative words, businesses can get a sense of the overall sentiment of their reviews

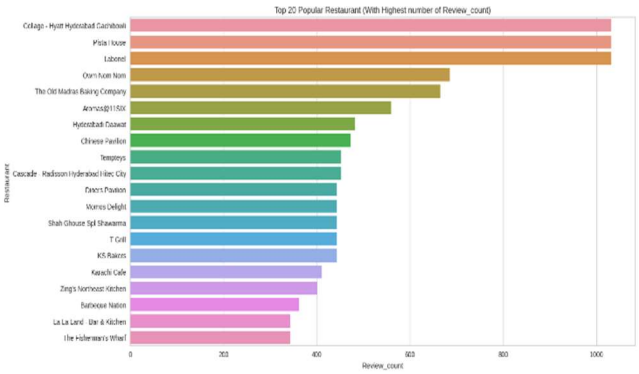

Fig.3 Top 20 Restaurant with highest number of review counts

Collage - Hyatt Hyderabad Gachibowli, Pista House and Labonel is the most reviewed restaurants having **1031** reviews

C. Hypothesis Testing

We have defined three hypothetical statements from the dataset. In the next three questions, perform hypothesis testing to obtain a final conclusion about the statements through our code and statistical testing.

1. More number of cuisines does not affect the ratings or reviews of the restaurant.

- **Null hypothesis**: The number of cuisines does not affect the ratings or reviews of a restaurant.
- **Alternative hypothesis:** The number of cuisines affects the ratings or reviews of a restaurant.

Result: The one-way ANOVA test results show an F-value of 9.82 & a very small p-value of 5.51e-05. This means that there is evidence to reject the null hypothesis that the number of cuisines does not affect the ratings or reviews of a restaurant, & support for the alternative hypothesis that the number of cuisines affects the ratings or reviews of a restaurant.

2. The Standard Deviation of Followers is equal to 620

- **Null hypothesis:** The standard Deviation of Followers is $\sigma = 620$
- **Alternative hypothesis**: The standard Deviation of Followers is not equal to 620

Result: The calculated p-value is 0.5052, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis that the standard deviation of followers is equal to 620. This means that there is not enough evidence to suggest that the standard deviation of followers is significantly different from 620.

3. There is a significant positive correlation between the length of the review and the rating provided by the customer

- **Null hypothesis:** There is no significant correlation between the length of the review and the rating provided by the customer.
- **Alternative hypothesis:** There is a significant positive correlation between the length of the review and the rating provided by the customer.

Result: The Pearson correlation coefficient measures the strength and direction of the linear relationship between two variables. The value of the correlation coefficient ranges from -1 to +1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and +1 indicates a perfect positive correlation. Pearson correlation coefficient is -0.0307, which is a small negative correlation. This suggests that there is a weak, negative relationship between the length of a review and the rating of the restaurant. That is, as the length of the review increases, the rating of the restaurant tends to decrease slightly.

D. Feature Engineering

In feature engineering, we have done some manipulation with our dataset. Following are the steps we have performed on our dataset to get the proper insights:

- **Handling Missing Values -** In Zomato dataset Restaurant_name, Links, and Cost have no missing values. However, the Collections column has 5003 missing values, making it unclear what this column represents. So, we are filling the null value in the collection column with 'not available' tag. Cuisines also have no missing values. Restaurant_timings has 100 missing values, which means that some restaurants do not have their operating hours listed in the dataset. Customer_Name, Rating, & Review_timing each have 7 missing values. Review and Pictures have 5 missing values. Review_count, Followers, Review_length, and

Polarity also have 5 missing values. The Customer_Name column lists the name of the customer who wrote the review, while the Review column contains the text of the review. The Rating column lists the rating given by the customer on a scale of 1-5, and the Review_timing column lists the date and time the review was written. The Pictures column lists the number of pictures attached to the review. The Review_count column lists the total number of reviews written by customers, while the Followers column lists the number of followers or fans of the restaurant's page or social media handle. The Review_length column lists the length of the review in characters or words. Finally, the Polarity column is a measure of the sentiment expressed in the review, & is classified as either positive or negative

- **Anomaly Detection -**
  Results of anomaly detection, each variable seems to have different skewness and kurtosis values, indicating different levels of deviation from a normal distribution.

  The cost variable has a positive skewness value of 1.143450 and a positive kurtosis value of 1.534478, indicating that the data is moderately skewed to the right and slightly more peaked than a normal distribution. This suggests that most of the observations are concentrated towards the lower end of the cost scale.

  The ratings variable has a negative skewness value of -0.707928 and a negative kurtosis value of -0.946219, indicating that the data is moderately skewed to the left and flatter than a normal distribution. This suggests that most of the observations are concentrated towards the higher end of the rating scale.

  The followers variable has a very high positive skewness value of 10.110880 and an extremely high positive kurtosis value of 151.860501, indicating that the data is highly skewed to the right and extremely peaked compared to a normal distribution. This suggests that there may be a few accounts with a very high number of followers that are outliers in the data set.

  The polarity variable has a negative skewness value of -0.572453 and a positive kurtosis value of 0.545082, indicating that the data is moderately skewed to the left and slightly more peaked than a normal distribution. This suggests that most of the observations are concentrated towards the positive polarity end of the scale, indicating that the sentiment is generally positive.

  Overall, the summary suggests that the variables have different degrees of deviation from normal distribution, with followers' variables showing extreme skewness and kurtosis. It's important to take into account these anomalies when analyzing the data to avoid biases or incorrect conclusions.

  We have used an isolation forest to visualize and remove the Anomalies from our data.

- **Categorical Encoding –**
  Categorical encoding is a technique used to represent categorical data in a numerical format that can be used in machine learning models. In the case of the Zomato dataset, the 'Cuisine' column is a categorical variable that contains 44 different cuisine categories. One popular technique for categorical encoding is one hot encoding.

  One hot encoding involves creating a binary vector for each category in the 'Cuisine' column. Each vector has a length equal to the total number of categories, in this case, 44. The value of each element in the vector is either 0 or 1, with 1 indicating that the dish belongs to that particular cuisine category and 0 indicating that it does not.

  For example, if a dish belongs to the 'North Indian' cuisine category, its one hot encoding vector would have a 1 in the element corresponding to the 'North Indian' category, and 0s in all other elements.

- **Textual Data Preprocessing –**

  The following are the textual data preprocessing techniques applied to the 'Review' column in the Zomato dataset:

  Expand Contraction: Contraction expansion is a technique where contractions like "can't" are expanded into "cannot". This helps to standardize the text and reduce variations in the text.

  Lower Casing: Lowercasing is the process of converting all text to lowercase. This is done to standardize the text and to reduce the number of unique words in the text.

Removing Punctuations: Punctuations are removed from the text as they do not add any value to the text analysis and only contribute to the noise in the text.

Removing Stopwords and Removing White spaces: Stopwords are common words in a language like "the", "a", "an" which do not contribute much to the meaning of the text. They are removed to reduce the size of the text and to focus on the more meaningful words. White spaces or extra spaces between words are also removed to standardize the text.

Rephrase Text: Rephrasing is the process of rewriting text in a different way, while retaining the meaning. It is done to standardize the text and reduce variations in the text.

Tokenization: Tokenization is the process of splitting the text into individual words or tokens. It is a crucial step in natural language processing and is done to prepare the text for further analysis.

Text Normalization: Text normalization is the process of converting text into a standard format. It involves techniques like stemming, lemmatization, and normalization of numbers and dates. This helps to standardize the text and reduce variations in the text.

Part of speech tagging: Part of speech tagging is the process of labeling each word in the text with its corresponding part of speech like noun, verb, adjective, etc. It is done to provide additional context for the words and to prepare the text for further analysis.

Text Vectorization: Text vectorization is the process of converting the text into a numerical representation that can be used by ML models. Techniques like bag of words, term frequency-inverse document frequency (TF-IDF), and word embeddings are used for text vectorization. This helps to represent the text in a way that can be processed by machine learning models.

- **Feature Manipulation and Selection –** In this section, we have manipulated features and dropped some unnecessary columns which are of literally no use. The selected features in Zomato restaurant clustering and sentiment analysis include a mix of numerical and categorical variables. The numerical features include 'Cost', 'Rating', 'Pictures', 'Review_count', 'Followers', 'Review_length', 'Polarity', and 'cuisine_count'. These variables provide important information about the cost of the restaurant, the quality of the food and service, the popularity of the restaurant, and the sentiment of the reviews. The categorical features are 'Cuisines', 'American', 'Andhra', 'Arabian', 'Bakery', 'Beverages', 'Biryani', etc. These variables represent the type of cuisine served at each restaurant. The 'pos_tags' feature indicates the part of speech tag of each word in the restaurant review text. This feature can be used to extract additional information about the review, such as the presence of adjectives or adverbs, which can be used to infer sentiment.

- **Dimensionality Reduction –**
  Dimensionality reduction is needed because it can help to reduce the amount of time and resources required for training a model. By reducing the number of dimensions, a machine learning model can more quickly learn patterns in the data and improve its accuracy. Reducing the number of dimensions can also help to avoid overfitting and reduce the need for large amounts of data to train the model. It helps in reducing the complexity of the data, removing redundant features. Reducing the number of dimensions can also help improve the accuracy of predictive models by reducing noise and improving the signal-to-noise ratio, while also making it easier to visualize high dimensional data.

IV. MODEL IMPLEMENTATION

We have implemented the following five models for clustering the restaurants:

- **K-Means Clustering:**
  We plotted the within-cluster sum of squares against the number of clusters to determine the optimal number of clusters. Based on the elbow method, we selected n_cluster=3. We then computed the silhouette score for cluster 2 to cluster 10 and found that n_cluster=3 is giving a good score of 0.63.

- **Principal Component Analysis:**
  We used PCA to reduce the dimensions of the dataset to two and plotted the data points on a scatter plot. From this plot, we observed that the data points can be divided into two clusters.

- **Hierarchical Clustering (Agglomerative Clustering):**
  We used Agglomerative Clustering to generate a dendrogram to visualize the hierarchical relationship between the data points. We then set the threshold distance by selecting a horizontal line that cuts the tallest vertical line in the dendrogram. Based on this threshold, we selected n_cluster=2. We then computed the silhouette score, which was found to be 0.634, & the Davies Bouldin Score, which was found to be 0.444.

- **DBSCAN Clustering:**
  We used DBSCAN Clustering to cluster the data points based on their density. We determined the optimal value for epsilon by selecting the point of maximum curvature on the distance plot. We then set eps=0.48 and min_samples=50. The silhouette score for this model was found to be 0.465.

- **Content Based Recommendation System:**
  Recommendation system that suggests restaurants to users based on the similarity of their reviews. Content-based recommendation systems can suggest 10 other restaurants with similar cuisines and cost, which have received positive reviews from other users. After clustering, the sentiment analysis is performed on the reviews within each cluster to determine the overall sentiment of the reviews. For example, if most of the reviews for a restaurant in the "Chinese cuisine" cluster have positive sentiments, then that restaurant will be considered a top-rated restaurant in that cluster. Finally, the system generates a list of top 10 restaurants with similar reviews based on the clustering and sentiment analysis. Based on the example you provided, the system can suggest other restaurants like "Shanghai Chef 2", "Flechazo", "Mathura Vilas", "Kritunga Restaurant", and "The Fisherman's Wharf" to the user, as they all have received a rating of 5.0 and are similar to "Chinese Pavilion" in terms of cuisine and ambiance.

V. MODEL EXPLAINABILITY

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that is particularly useful for clustering datasets with complex structures and varying densities. It works by identifying clusters based on the density of points in a given area.

In our project, DBSCAN is a suitable algorithm because restaurants in a given area may have varying densities, and the number of restaurants may not be known in advance. DBSCAN can discover the optimal number of clusters by detecting regions of high density and identifying clusters as regions with a high density of points. It is effective for sentiment analysis because it can group similar reviews together based on their proximity in the feature space. This is particularly useful when dealing with large volumes of text data, where traditional clustering algorithms may struggle.

The Silhouette Coefficient is a measure of how well each data point fits within its assigned cluster, as compared to neighboring clusters. A high Silhouette Coefficient indicates that the clustering is effective in grouping similar data points together and separating dissimilar data points.

A Silhouette Coefficient of 0.465 suggests that the DBSCAN algorithm is effective in clustering Zomato restaurant data and performing sentiment analysis, as it indicates that the clusters are well-defined and distinct. Therefore, DBSCAN is a suitable algorithm for Zomato restaurant clustering and sentiment analysis

VI. RESULT

The results obtained from analyzing a Zomato restaurant dataset using different clustering algorithms and a content-based recommendation system. The document reports that K-Means Clustering, PCA, Hierarchical Clustering, and DBSCAN Clustering were used to group the data points into clusters based on various criteria, and the optimal number of clusters and parameters were selected based on different validation metrics. Additionally, the document mentions that a Content-Based Recommendation System was developed to suggest restaurants to users based on the similarity of their reviews.
Regarding the effectiveness of the clustering algorithms, we analyze that DBSCAN Clustering achieved a silhouette score of 0.465, which is the highest among the evaluated algorithms. However, it is important to note that each algorithm may have different advantages and limitations depending on the characteristics of the dataset and the research question. Therefore, the choice of the best model may depend on the specific goals and requirements of the analysis.

VII. FUTURE WORK

Following are some future works on Zomato Restaurant clustering and Sentiment analysis.

AI-driven algorithms can be used to automatically generate summary reports of restaurant reviews in various languages and identify common trends in customer feedback.

Sentiment Analysis: Sentiment analysis can be used to identify what people really think about a restaurant based on reviews. This can help customers determine which restaurants are worth their time, and which ones should be avoided.

Online Learning: Currently, the system is trained on a batch of historical data. Future work can involve developing an online learning system that can adapt to the changing preferences of the users in real-time.

Multi-language support: Zomato is a global platform and supports multiple languages. Future work can involve developing a multi-language content-based recommendation system that can handle different languages and provide recommendations in the user's preferred language.

Computer Vision: Utilize computer vision techniques to identify objects and classify food items in restaurant photos.

Deep Learning: Use deep learning algorithms to compare reviews between two different restaurants and generate comparison results.

VIII. CONCLUSION

The conclusion of this Zomato restaurant clustering and metadata sentiment analysis project is that it is possible to use natural language processing and machine learning algorithms to build a model that can accurately cluster restaurants based on their reviews and sentiments. This project has helped identify customer preferences and restaurant ratings in order to better understand the impacts of customer feedback on the restaurant industry. This model can then be used to improve the decision-making process of a restaurant owner or manager in terms of advertising, pricing, customer acquisition, and other important business decisions. With this data, business owners can make more informed decisions about the quality of their restaurants and better understand the customer experience. We have seen that AI-based solutions provide a powerful tool for business owners to gain insight into the performance of their restaurants. After that sentiment analysis can be used to gain insights into customer preferences, providing data-driven understanding into how customers perceive different aspects of service quality. Furthermore, the insights provided by this project can be used for further business growth strategies.