Capstone Project - 4

ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS

Presented by:
Mrugesh Patel

Content

1. Problem Statement
2. EDA and feature engineering
3. Clustering
4. Sentiment Analysis
5. Conclusion

Problem Statement

• The Project focuses on Customers and Company. One has to analyze the sentiments of the reviews given by the customer in the data and make some useful conclusion in the form of Visualizations. Also, cluster the Zomato restaurants into different segments in order to help customers find the best place to eat in the locality. The data is visualized as it becomes easy to analyze data at instant. Also the metadata of reviewers can be used for identifying the critics in the industry.

• So the problem statement is to find the best clusters and analyze the review data to understand the sentiment.

Data Reading and Exploration

• After Initial exploration of the datasets It was observed that, there are 5 attributes for 105 unique restaurants are provided, out of which Collections column has 51 Null values, and Timings has 1 Null value. Review Dataset has 10000 entries of reviews with 6 attributes for each review. Except Pictures, every column has Null values, although the Null value fraction is less than 1%.

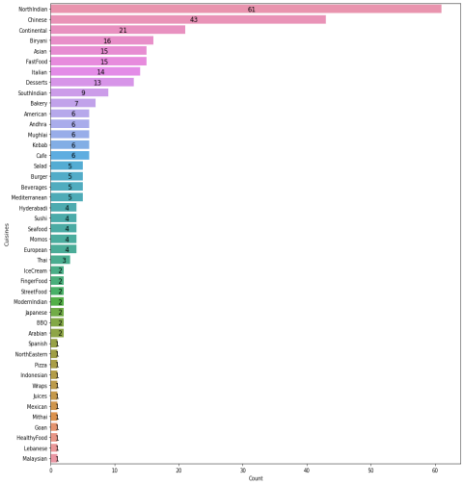
Feature Engineering

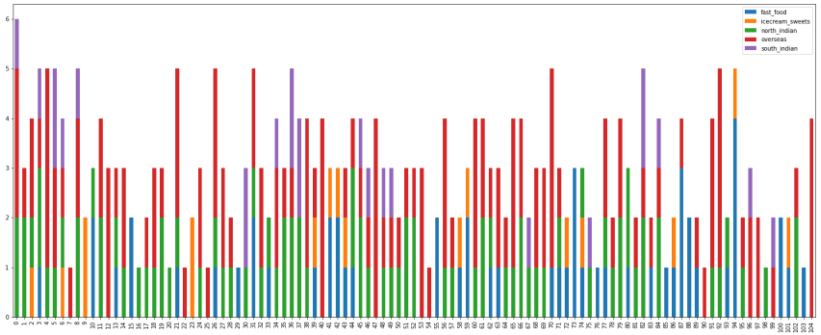
- Cost Values have commas in the values, so the format of the Cost values need to change.
- The Links Column is of no use, it is just an information.
- Collections has more than 50% Null values, So the column needs to be dropped.
- There Are more than one cuisines for most of the restaurants, We must find the unique cuisines and their count.
- There are 77 unique entries in all different format for Timings column. Now it will be impractical to clean and transform 77 unique values for 105 total values. So for further analysis, Timings will not be considered.

EDA

❖ Cuisines

- Only 8/44 cuisines have double-digit counts
- 44 cuisines is impractical for dataset of 105 restaurants for clustering
- Practical approach to categorize these cuisines in to more understandable cuisine category, based on the attribute to cluster the restaurants easily.
- After categorizing, using CountVectorizer for converting them to numeric form for clustering
- As we can see most restaurants has multiple categories mixed, although some has one dominating category.

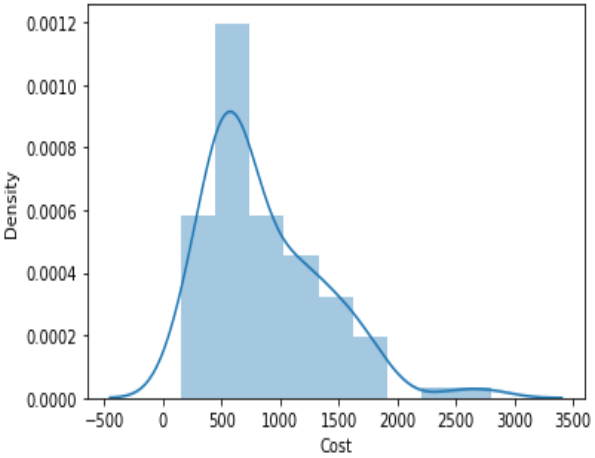




EDA

❖ Cost

- Cost values distribution is very close to normal distributions as we can see in the plot above.
- Most values are between the range of 500 to 1000, with mean of 861.428 and median of 700.
- The costliest restaurant has Cost value of 2800 and minimum value is 150.
- The values are of very large scale compared to the other columns, so normalizing the dataset before clustering also can be checked while clustering.

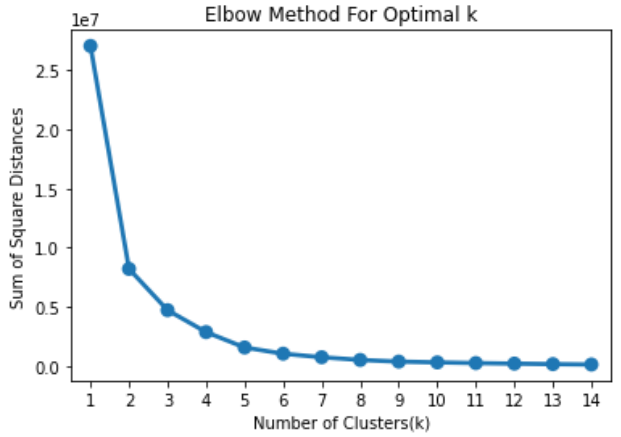


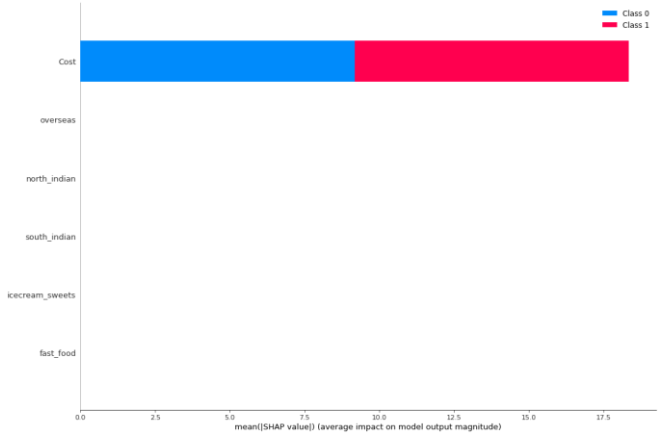
Clustering

❖Kmeans (actual data)

• For Kmeans optimum number of clusters is 2 according to Elbow method, f1 score of 1.0.

• But in terms of feature importance, the explainer revealed that only the cost is the important feature for clustering for actual data.

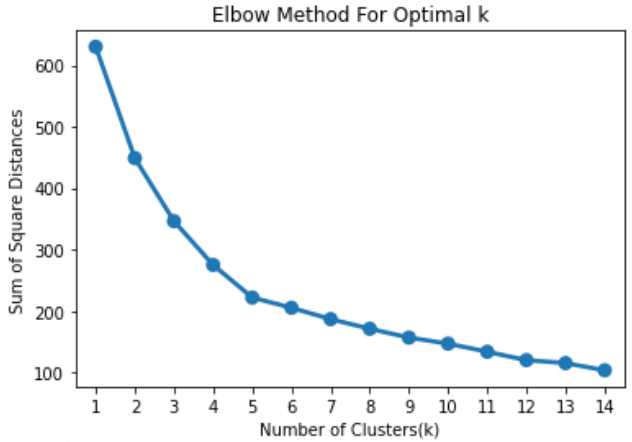


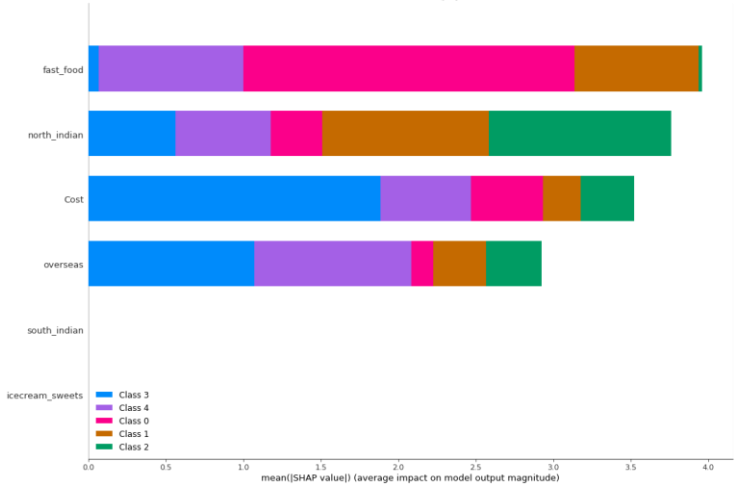


Clustering

❖ Kmeans (Normalized data)

For normalized dataset better results in terms of feature importance with optimum number of cluster being 5, with fast_food and north_indian category being the most important features according to SHAP explainer, but the classification model achieved 0.51 f1 score.

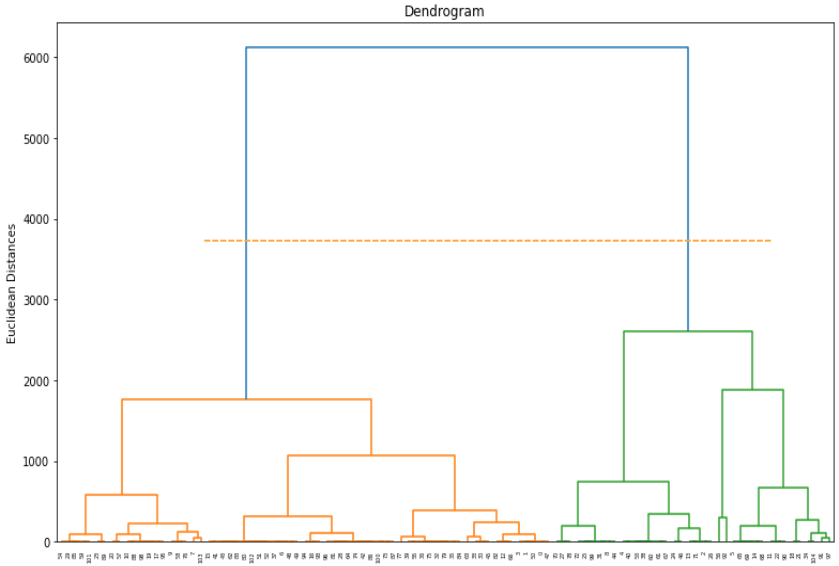




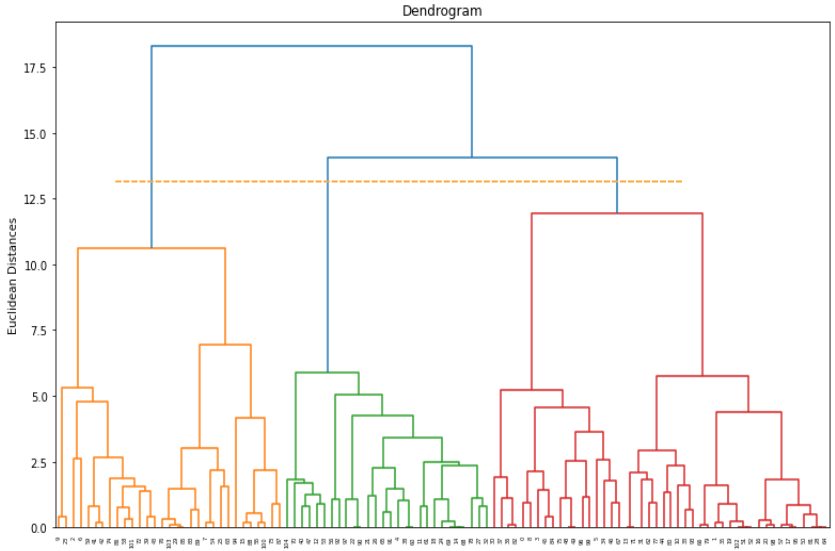
Clustering

❖ Hierarchical Dendrogram

Actual data



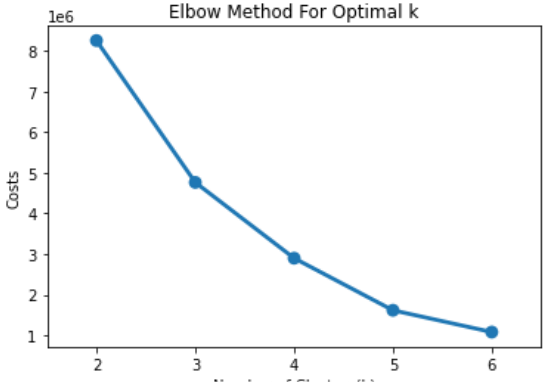
Normalized data

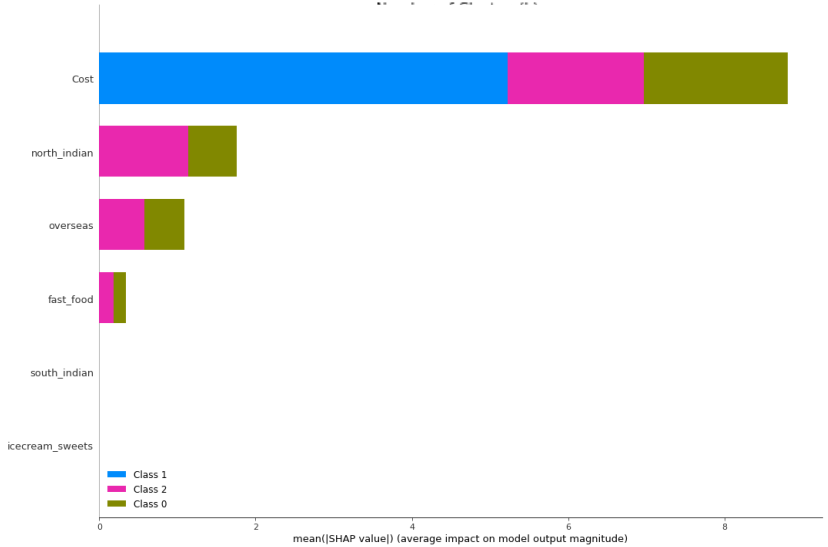


Clustering

❖ KPrototype (Actual data)

For actual data according to elbow method and silhouette score method optimum number of clusters is 3, with f1 score for K-Prototypes clusters is 0.857. The feature importance for the model is shown below



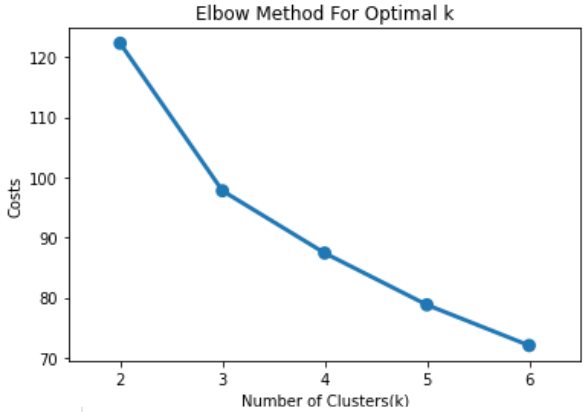


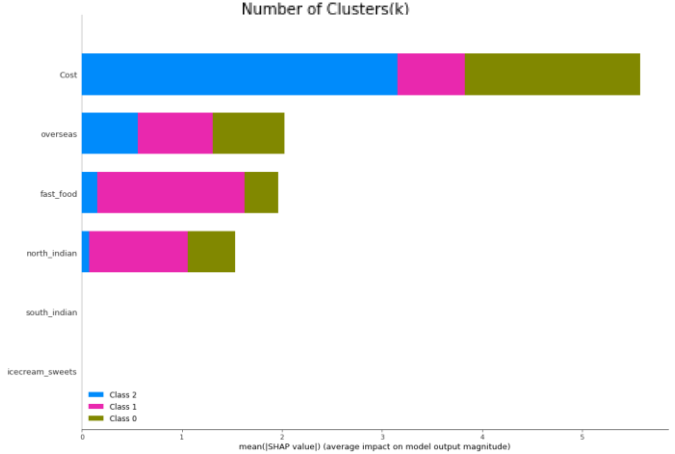
Clustering

❖ KPrototype (Normalized data)

For normalized dataset, produced similar results as actual values with 0.93 f1 score.

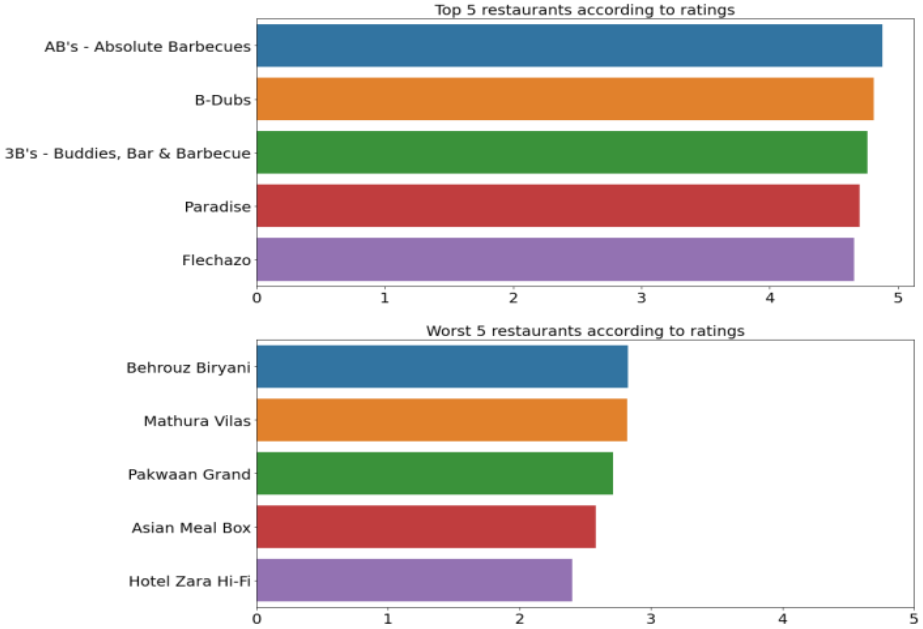
And SHAP explainer shows that cost is the primary feature for clustering, with overseas category as secondary feature. Class 2 has highest effect by cost, while class 1 is affected by fast_food and north_indian categories. Class 0 is mostly shared among overseas and cost.





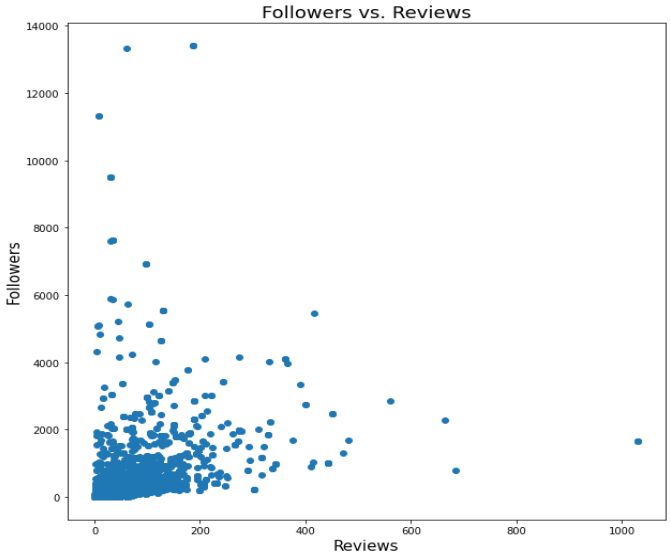
Sentiment Analysis

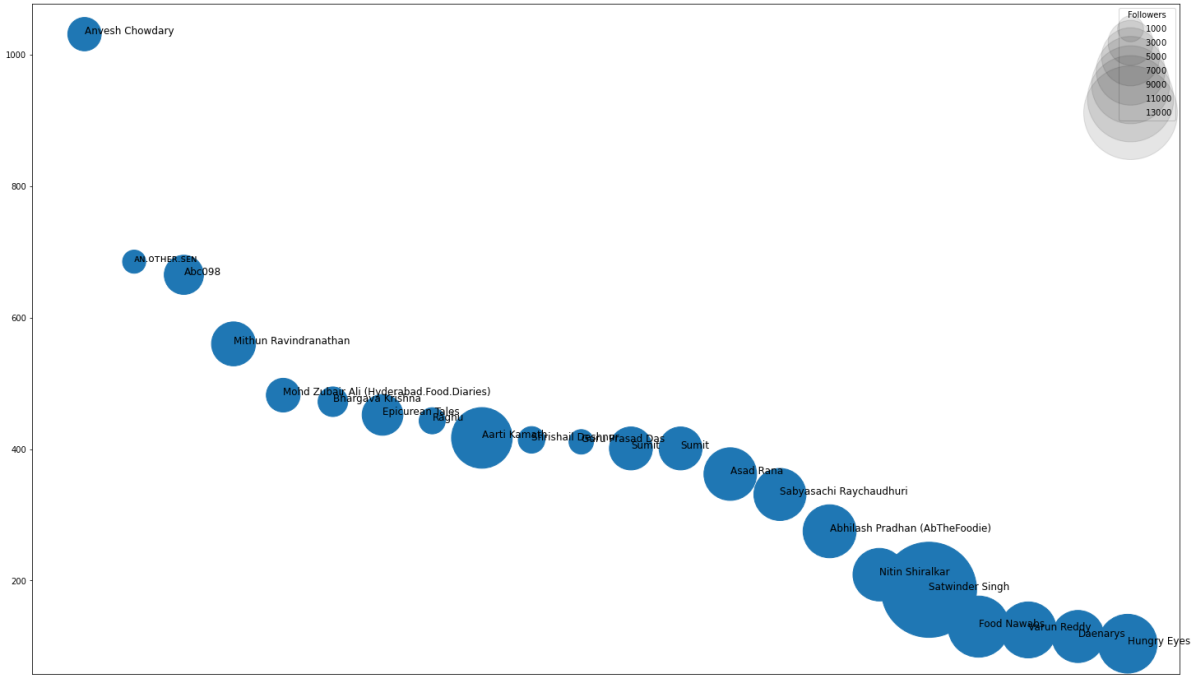
❖ **Top/worst Apps according to rating**



Sentiment Analysis

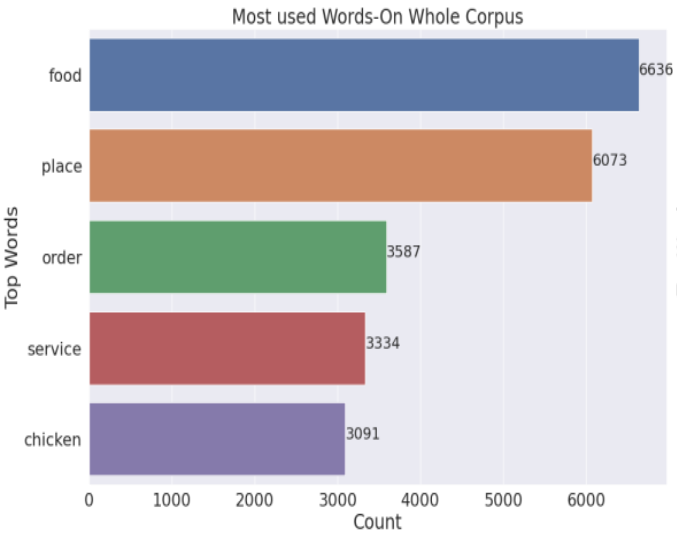
❖Critics

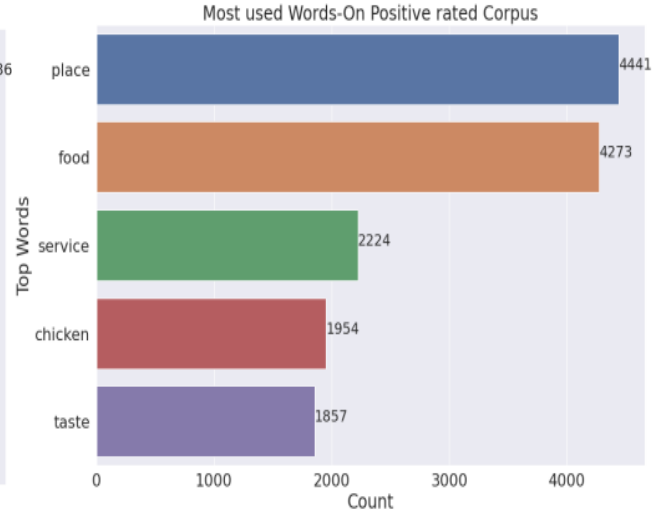


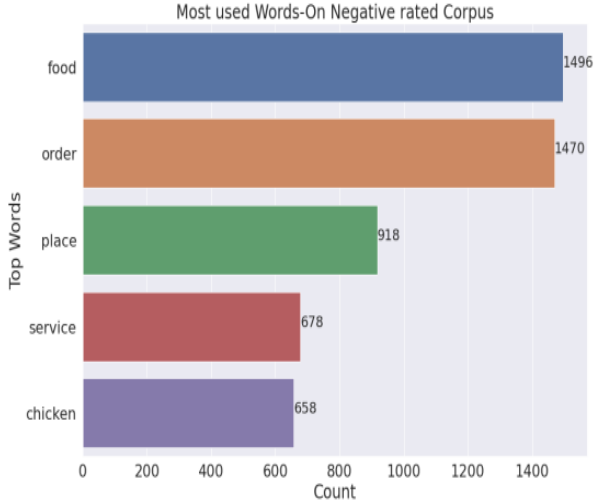


Sentiment Analysis

❖Top Words for all Sentiments



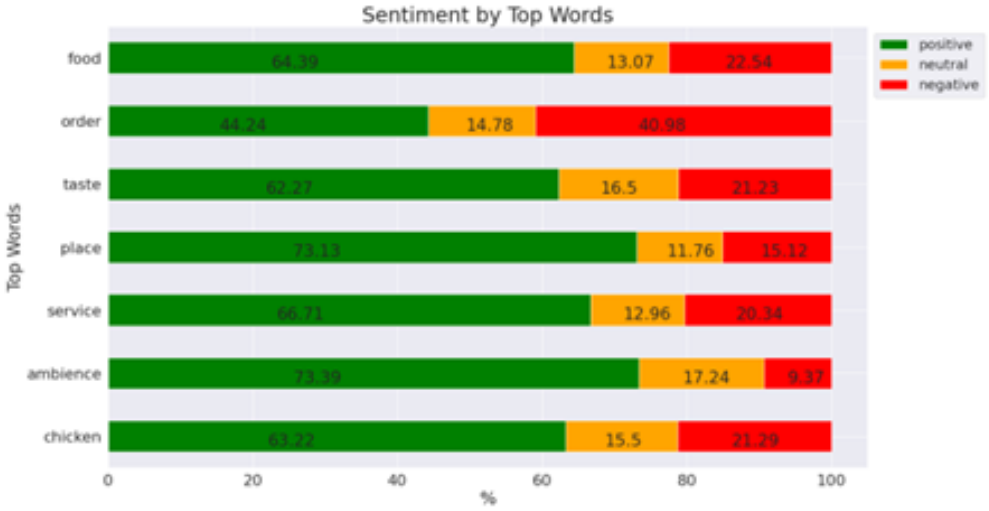




Sentiment Analysis

❖Sentiment By Top Words

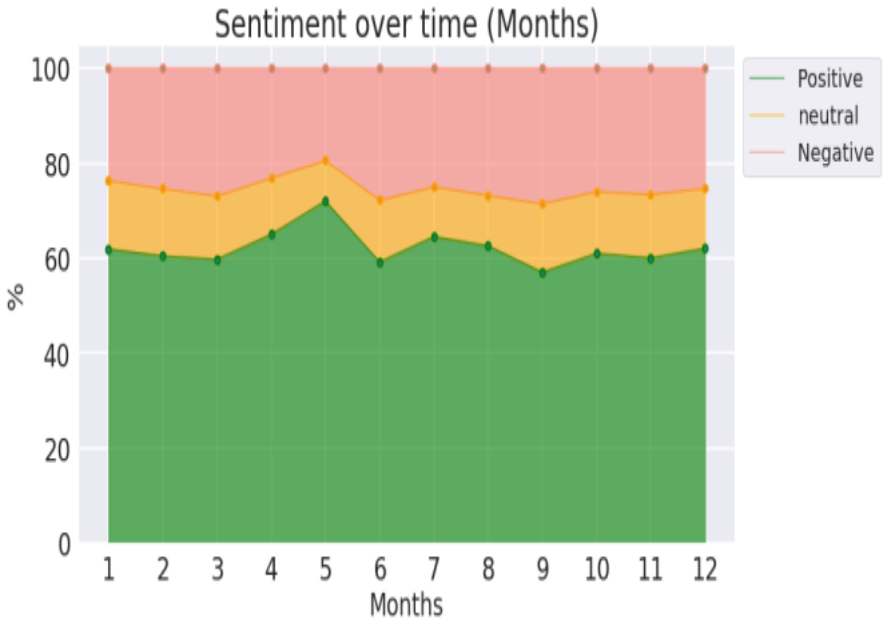
Considering the top words and their associatively with the respective corpus, the sentiment for each top words is plotted in stacked bar plot as shown below.



Sentiment Analysis

❖Sentiment Over Time

Plotting the mean sentiment over time, month-wise, showed that there is spike in May month in positive reviews fraction. Apart from that there is no apparent variation.



Future Scope

- The dataset for clustering has only few restaurants data. Finding clusters with high difference between them is hard for smaller set. So more restaurant data can be added.
- Formatting of Timing is very messy in the dataset. It can be helpful for clustering considering timing as well. So timing with better formatting can be obtained and used for clustering.

Conclusion

❖ Clustering

✓ Three methods for clustering KMeans, Hierarchical and KPrototye with actual values and normalized data was used for clustering. For finding optimum number of clusters silhouette and Elbow method were used.

✓ To validate further, classification machine learning model was fitted for clustered data for both KMeans and KPrototypes method, and f1 score was checked. To understand feature importance, SHAP explainer was used.

✓ Considering both f1 score and feature importance , efficacy of the method used for clustering was decided.

✓ KMeans method produced excellent f1 score with 2 clusters, but only considered cost as a feature, as the scale of values of cost is high. Same Method with normalized values improved in terms of feature importance, but f1 score dropped to 0.51, with number of optimum clusters as 5.

✓ Hierarchichal dendogram confirmed optimum clusters as 2 for actual values, but for normalized values it was observed to be 3.

✓ Finally KPrototype was used, and produced good results with f1 score of 0.857 for actual values and 0.93 for normalized values, with optimum number of clusters as 3.

✓ The most important feature was cost , with overseas type categories as a second important feature. Overseas category included chinese, asian, italian etc.

Conclusion

❖ Sentiment Analysis

✓ Top 5 restaurants according to rating are, ‘AB's - Absolute Barbecues', 'B-Dubs', '3B's - Buddies, Bar & Barbecue', 'Paradise', 'Flechazo'.

✓ And the Worst 5 restaurants according to rating are, 'Behrouz Biryani', 'Mathura Vilas', 'Pakwaan Grand', 'Asian Meal Box', 'Hotel Zara Hi-Fi'.

✓ Top words from the corpus of positively rated reviews are, 'good, ‘food', 'place', 'order', 'service', 'chicken'.

✓ Top words from the corpus of negatively rated reviews are, 'food', 'order', 'place', 'service', 'chicken'.

✓ From Sentiment by top words analysis it was observed that, words with highest positively rated fraction are place, ambience. And words with highest negatively rated fraction are order, taste.

✓ According to number of reviewers and number of followers, top critics are Avnesh Chowdhary, Mithun Ravidranathan, Satwender Singh, Food Nawabs, Aarthi Kamath.

✓ Plotting the mean sentiment over time, month-wise, showed that there is spike in may month in positive reviews fraction. Apart from that there is no apparent variation.

Thank You