**ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS**

**Akhilesh Kumar**

**Mentor: Richa Sharma**

**AlmaBetter**

**Abstract:**  In today’s digital world, food apps like Zomato are widely used because it provides a platform for people to share their opinion about the restaurants and cafes they have visited. This paper includes analysis of client ratings and reviews in Zomato utilizing content mining. Utilizing content mining, break down the content audits/reviews from the client with a specific end goal to create productive results and legit surveys. The rating has a review of the restaurant which can be used for sentiment analysis. Based on this, writers want to discuss the sentiment of the review to be predicted. The method used for preprocessing the review is to make all words lowercase, tokenization, remove numbers and punctuation, stop words, and lemmatization. Then after that, we create a word to vector with the term frequency-inverse document frequency (TF-IDF). The data that we process are 10,000 reviews. After that, we make positive reviews that have a rating of 3.5 and above, negative with reviews that have a rating of 3 and below. We have used Split Test, 80% Data Training and 20% Data Testing. The metrics used to determine random forest classifiers are precision, recall, accuracy, F1 score.

**Keywords:** *logistic regression, random forest, decision trees, precision, recall, accuracy, F1 score.*

**Introduction:** In today’s digitized modern world, popularity of food apps is increasing due to its functionality to view, book and order for food by a few clicks on the phone for their favorite restaurant or cafes, by surveying the user ratings and reviews of the previously visited customers.Food app like Zomato provides a secular part where user can rate their experience of the visited restaurant or café. Zomato also provides columns for writing classified user reviews. Sharing on the internet is something we usually do. Giving a review is also a useful activity so that other people on the internet can find out something else and see opinions about things. The usual things reviewed by someone in the form of experiences, places, objects, and others. Give a review we usually use text to explain something that we experience with an item, place, or event that we normally experience. Zomato is a site where someone can give a review of a restaurant, how the restaurant is and someone's opinion about the restaurant. Restaurant customer satisfaction can be analyzed by their review on Zomato. Sometimes, restaurants see the reviews in Zomato, but they don't get if the reviews are positive or negative to their restaurants. Reviews on Zomato are still in the form of text and can be classified with positive, negative, or neutral ratings. Zomato doesn’t have an analysis of how users interact with the reviews and what words will indicate whether they like it or not. We need to extract the words in review and analyze it so we can know how users interact in Zomato and get customers' satisfaction by their review. In this paper, we propose a method to analyze user’s sentiment of Zomato Restaurants. We are using Random Forest Classifier to classify the sentiments of users based on their review. We also find words that affect the classifier model. Also, we focus on mining customer reviews, authenticating them and classifying them into positive and negative reviews.

**Logistic regression:** Logistic regression (LR) is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X. However, unlike linear regression the response variables can be categorical or continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a least squares to fit the final model. This means the researcher has more freedom when using LR and the method may be more appropriate for non normally distributed data or when the samples have unequal covariance matrices. Logistic regression assumes independence among variables, which is not always met in morphoscopic datasets. However, as is often the case, the applicability of the method (and how well it works, e.g., the classification error) often trumps statistical assumptions. One drawback of LR is that the method cannot produce typicality probabilities (useful for forensic casework),

**Random forest:** Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. As the name suggests, a Random Forest is a tree-based ensemble with each tree depending on a collection of random variables. More formally, for a p-dimensional random vector X = representing the real-valued input or predictor variables and a random variable Y representing the real-valued response, we assume an unknown joint distribution . The goal is to find a prediction function for predicting Y. The prediction function is determined by a loss function and defined to minimize the expected value of the loss .

**Decision trees:** Decision trees produce models that are easy to understand if the number of branches remain relatively small. If there are many variables then the tree can become complex and uninterpretable. Decision trees work by doing successive binary splits (some algorithms do produce more than two branches at each split). The first split will yield the biggest separation or distinction in two groups of data. Each subgroup is then split until some stopping criteria are reached. Algorithms differ partly on how they measure the separation distance between groups. Two different algorithms on the same data set would very likely give different branches and rules. Decision trees can result in very easy to understand decision rules.

**Multinomial Naive Bayes:** The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

**Model Deployment:** For clustering, we have decided on 4 clusters after the Silhouette score plot. We have tried on 3 clusters but were not able to draw relationships between the variables and on which variables the clusters were split. Hence, we decided on clustering into 4 groups.

KMeans clustering and Hierarchical clustering have been used for the clustering model. All features in the dataset are used for this exercise.

For sentiment analysis, Multinomial NB, Logistic regression, Decision trees and Random Forest have been used. After tuning the hyperparameters and using cross validation technique, the optimum hyperparameters were chosen for the models. In the Sentiment Analysis we tried by reducing features using regularisation though it's causing overfitting, we did a different experiment, though the result remained the same.

**Results and discussion:**

**Table 1: Performance metrics for diff. models on training set using TF-IDF vectorizer**

| Algorithm | Class label | Performance parameters | | | |
| --- | --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| Multinomial NB | 0 | 0.84 | 0.90 | 0.63 | 0.74 |
| 1 | 0.82 | 0.96 | 0.88 |
| Logistic regression | 0 | 0.85 | 0.75 | 0.90 | 0.82 |
| 1 | 0.93 | 0.83 | 0.88 |
| Decision trees | 0 | 0.79 | 0.73 | 0.66 | 0.69 |
| 1 | 0.82 | 0.86 | 0.84 |
| Random forest | 0 | 0.74 | 0.59 | 0.95 | 0.73 |
| 1 | 0.75 | 0.62 | 0.75 |

**Table 2: Performance metrics for diff. models on test set using TF-IDF vectorizer**

| Algorithm | Class label | Performance parameters | | | |
| --- | --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| Multinomial NB | 0 | 0.81 | 0.85 | 0.59 | 0.70 |
| 1 | 0.80 | 0.94 | 0.86 |
| Logistic regression | 0 | 0.80 | 0.69 | 0.85 | 0.76 |
| 1 | 0.90 | 0.78 | 0.84 |
| Decision trees | 0 | 0.79 | 0.70 | 0.63 | 0.66 |
| 1 | 0.80 | 0.85 | 0.62 |
| Random forest | 0 | 0.70 | 0.55 | 0.90 | 0.69 |
| 1 | 0.91 | 0.58 | 0.71 |

**Table 3: Performance metrics for diff. models on training set using bag of words**

| Algorithm | Class label | Performance parameters | | | |
| --- | --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| Multinomial NB | 0 | 0.86 | 0.84 | 0.75 | 0.80 |
| 1 | 0.87 | 0.92 | 0.89 |
| Logistic regression | 0 | 0.85 | 0.75 | 0.90 | 0.82 |
| 1 | 0.93 | 0.83 | 0.88 |
| Decision trees | 0 | 0.77 | 0.67 | 0.71 | 0.69 |
| 1 | 0.83 | 0.80 | 0.81 |
| Random forest | 0 | 0.77 | 0.63 | 0.95 | 0.75 |
| 1 | 0.96 | 0.67 | 0.79 |

**Table 4: Performance metrics for diff. models on test set using bag of words**

| Algorithm | Class label | Performance parameters | | | |
| --- | --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 score |
| Multinomial NB | 0 | 0.83 | 0.79 | 0.72 | 0.75 |
| 1 | 0.85 | 0.89 | 0.87 |
| Logistic regression | 0 | 0.81 | 0.69 | 0.85 | 0.76 |
| 1 | 0.90 | 0.78 | 0.84 |
| Decision trees | 0 | 0.75 | 0.66 | 0.67 | 0.67 |
| 1 | 0.81 | 0.81 | 0.81 |
| Random forest | 0 | 0.72 | 0.58 | 0.89 | 0.70 |
| 1 | 0.91 | 0.63 | 0.74 |