Sentiment Analysis and classification of Restaurant Reviews

Ariba Waseem[1,2]* and Tayyaba Zubair[2,1]†

[1]*Computer Science, Institute of Business Administration, Karachi, 75270, Sindh, Pakistan.
[2]Computer Science, Institute of Business Administration, Karachi, 75270, Sindh, Pakistan.

*Corresponding author(s). E-mail(s):
a.waseem.18054@khi.iba.edu.pk;
Contributing authors: t.zubair.18049@khi.iba.edu.pk;
†These authors contributed equally to this work.

Abstract

In the last two years, Covid-19 has entirely redefined the idea of "normality" through a significant cultural shift towards online platforms, creating new consumer trends. The newly emerged trend of online feedback on the restaurant plays a substantial role for the businesses to capture the customer's requirements and quality improvement hints. These reviews indicate the restaurant's performance on the most notable measures such as taste, price, service, and place or ambiance. People express their sentiments on many platforms through blogs or posts, and customers on a large scale rely on these reviews before making any decision. Analyzing the feedback produced by the consumers uncovers many hidden patterns and provides insights on customers' behavior and expectations in regards to these notable measures from the restaurant. This study examines the criteria of customer ratings and proposes an integrated approach that leverages text mining and machine learning to associate the ratings with the reviews quantitatively. Visualization techniques are used to screen out the impactful factors which influenced the restaurant rating based on the customer's experience.

Keywords: Text Mining, Machine Learning ,Sentiment Analysis, Restaurant Reviews , Sankey Chart

1 Introduction

The 20th century brought a massive change in the use of the internet, such as online activities via social media platforms. Ease of using online sites and 24x7 availability sets an edge over other services, especially the trend of online reviews on any product. The visibility of online reviews has diverted the customer's trust more to consumer reviews rather than fancy marketing advertisements. This leaning trend has a major impact on restaurant customers whose decisions on selecting cuisine rely on past experiences or word-of-mouth. These experiences are user-generated content through blogging, videos, photos, and stories. It helps customers decide on good restaurants based on the available feedback on their services, taste, ambiance, and price. Customer satisfaction needs to be analyzed to know what they are looking for and what makes them feel good about the restaurant services.

Consequently, restaurants are required to keep track of their customer reviews to analyze them for improving services. Sentiment Analysis helps to do so. Sentiment Analysis helps to know about the customer's expectations from the restaurant's services and satisfaction. It helps the restaurant to improve the customer's experiences by incorporating them with the customer's suggestions. Sentiment Analysis or opinion mining determines the polarity of reviews, whether it is positive or negative. It can automatically detect the nature of feedback given by the customer on the restaurant's performance. Opinion mining also helps in ranking the restaurant based on customers' experiences and ratings.

This paper draws an image on the idea of NLP(Natural language processing) to the Kaggle dataset, extending the concepts from the sentiment analysis field. Sentiment Analysis a subfield of NLP comprises different Machine Learning Algorithms tries to find and extract useful insights from the text fields. NLP a component of AI break downs the natural language giving the ability to computers to understand and interpret the human language. It provides many methods to extract data from text such as sentiment, text clarification, topic modeling, keyword extraction, and aspect mining.

For this purpose different visualization techniques such as Sankey Charts, Bar Graphs have been used to visualize the insights produced from Machine Learning Models. The data cleaning is carried out through NLP techniques such as lexical analysis, semantic analysis, and syntactic analysis. The model is trained on three different classification algorithms such Random Forest, Support Vector Machine (SVM), Logistic regression to calculate the accuracy on setting different parameters. The model's performance is evaluated on various metrics such as Precision, Recall, F-score.

The paper visualizes the insights obtained from these models by using visualization techniques to present the idea.

The aim of this research work mainly focuses on the visualization of text analysis produced through different machine learning techniques. The overall results are shown by using Sankey Charts and Network Diagram.

The paper is organized as follows: Part I discusses the brief overview of

past studies performed on sentiment analysis for restaurant reviews. Part II describes the Methodology used for extracting information. Part III presents the results of the model. Part IV visualizes the insights obtained from the trained model. Part V concludes the suggestions and implies future applications.

2 Related Work

Georgiana Denisa Stefan presented a method in which sentiment embedding are fed to a convolutional neural network with the architecture including an input layer, convolutional layer, a max-pooling layer, and a fully connected SoftMax layer. The accuracy achieved after 3 epochs is 66.03Mohanad Alhadethy, Mortadha Hamad, and Refed Adnan used Twitter data of customer reviews and performed sentiment analysis of reviews by using Naive Bayes algorithm. The accuracy achieved is 73

Rachmawan Adi Laksono, Kelly Rossa Sungkono, Riyanarto Sarno worked on the classification of Surabaya restaurant customer satisfaction using Naive Bayes and sentiment analysis using TextBlob. The accuracy achieved by Naive Bayes was 72.06

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, Iqbal H. Sarker worked on a deep learning technique BiLSTM to classify the restaurant reviews positive and negative. A corpus consisting of 8435 reviews is constructed the results show that the BiLSTM technique produced the highest accuracy of 91.35Masrur Adnan; Riyanarto Sarno; Kelly Rossa Sungkono used data from TripAdvisor and perform Text Classification using Decision-Tree-J48 to classify data in positive and negative and obtained accuracy up to 45.6

Kanwal Zahoor, Narmeen Bawany, Soomaiya Hamid uses 4000 record data from Facebook Page "the SWOT'S guide to Karachi's restaurants" perform sentiment analysis by Naive Bayes Classifier, Logistic Regression, Support Vector Machine (SVM), and Random Forest. Then they used the text categorization technique to classify reviews according to feedback about food taste, ambiance, service, and value for money. The best accuracy achieved was 95

Priya Kamath B., Geetha M., and Dinesh Acharya U proposed an approach in which they used 1000 purely unstructured restaurant reviews as input, preprocessed them, after cleaning they created BoW(Bag of Words) for feature extraction. and then classify them using Decision Tree, SVM, Random Forest, K-NN, Logistic Regression, and Naïve Bayes to check the effectivity of algorithms and found at the classification model has no significant influence in the correct classification of reviews. (Priya Kamath B., 2021)

3 Data Set

Our dataset of restaurant reviews has 977 records. The dataset has two columns. Column 1 of Reviews having 977 unique values, and Column 2 of Like representing if anyone liked it or not. It is Binary either 1 or 0.

4 Methodology

In this paper the data set used is from Kaggle. The methodology includes, understanding of data, Cleaning of Data, Preparation of Data, Sentiment Analysis and then Category classification.

Data Understanding:

Our dataset of restaurant reviews has 977 records. The dataset has two columns. Column 1 of Reviews having 977 unique values, and Column 2 of Like representing if anyone liked it or not. It is Binary either 1 or 0.

Data Preparation:

For the processes like sentiment analysis, Data preparation is the initial and major step. Because the provided reviews are in natural language and require some serious cleaning, to obtain high accuracy. It includes:

Loading Dependencies:

We made use of NLTK for processing the Review, Sklearn for classification methods, feature extraction and score Calculations, WordCloud, PIL, Seaborn and Matplotlib for data visualization, Pandas for loading data and NumPy for generating random probabilities for train-test split and heatmaps.

Data Preprocessing:

For text cleaning, we convert all letters to lower/upper case, removing punctuation, removing stop words, tokenization, and lemmatization. Moreover, we used the built-in preprocessing capabilities of CountVectorizer and Tfidf Vectorizer as well.

Feature Extraction:

In this process we analyze the data, reviews minutely to find out the features (i.e., words) which would be most useful in the classification. Then these features would be further used to train the classifier.

Review Classification and Categorization

After segregating data and train-test split, preprocessed data is passed through different classification algorithms. The model is trained using four different supervised machine learning techniques; SVM, Random Forest, Naive Bayes, Logistics Regression. The highest accuracy of the model is noted. Here the reviews are classified as positive and negative. Through category classification, the reviews were classified into four sub-categories.

1)Taste
2)Ambiance
3)Price
4)Service

The reviews are given a rank out of 5, (1 being least satisfied and 5 being strongly satisfied), based on the following four categories for every restaurant. The highest accuracy rate is observed through these classification methods. The results are then analyzed through different data visualization techniques. The data visualization covers many important aspects of obtained results to convey the story to viewers.
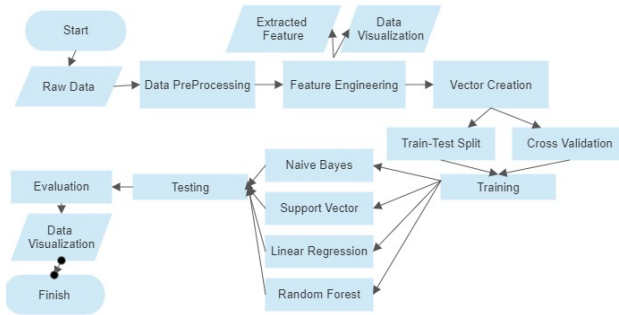
Fig. 1: Process Flow

5 Results and Discussion

In this project, we have trained the model using Machine Learning algorithms, in which Random Forest worked well. This project aims to visualize the analysis obtained by the models to conclude the useful findings.

Models	Accuracy
Naïve Bayes	92.2
Linear Regression	90.3
SVM	93.1
Random Forest	94.2

Fig. 2: Accuracy of Different Machine Learning Algorithms

The figure below represents the distribution of restaurant reviews primarily into 4 main categories then the type of sentiment attached to it as whether it is good or bad.

From our analysis, we found out that for the most customers, taste ranks first among all the positive reviews, which infers that a customer judges a restaurant on taste rather than other services. The possible explanation for this is the customer pays for the food, and they expect to have a good meal irrespective of other services. It is also being observed that the restaurants are ranked 5/5 on the good taste for food.

The place is the second-highest judging factor for a restaurant, and cost is the least one. Restaurants are ranked above the average for setting up a good environment to enjoy a meal. The restaurants having a bad place are ranked
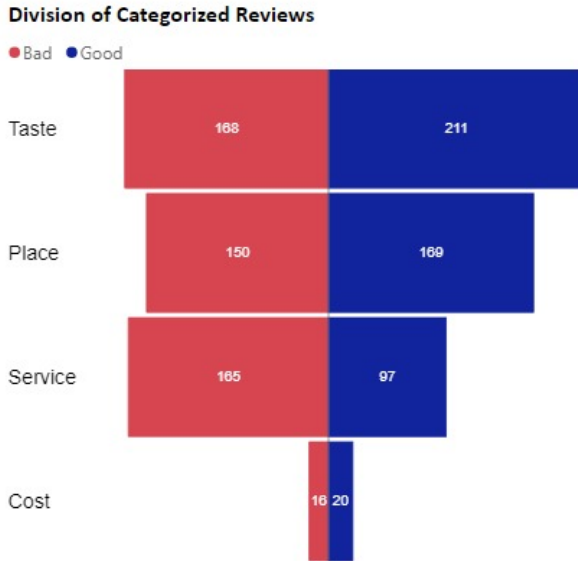
Fig. 3: Classification of Reviews After Sentiment Analysis

below the average irrespective of other factors. From here we can suggest that the food taste and the environment matter a lot. People expect to have a fancy and decent place, the unhygienic places make customer angry.

Figure Below visualize the analysis of Restaurant from the results of Random Forest.
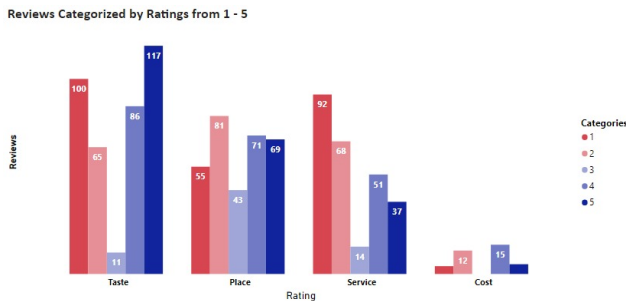
Fig. 4: Analysis of Restaurant Reviews by Random Forest according to Ratings

Our solution from the analysis presents the particular features from the reviews, categories of each review and ranking based on the categories. Owners of restaurants can make good use of each visualization to extract the important information and act upon those suggestions. From reviews, they can analyze

their current brand image, and customer concern towards the restaurant. They can get to know, why a particular customer like or dislike the food? What are the services that make them happy? Through these analysis, they can compare with the other similar restaurants, and get to know why they are receiving good/bad reviews.

Even though the model performance is good enough to categorize the reviews, and visualize the important features, there is still a room for improvement. The paper includes all the relevant methodology through which one can analyze the brand image and improve it.
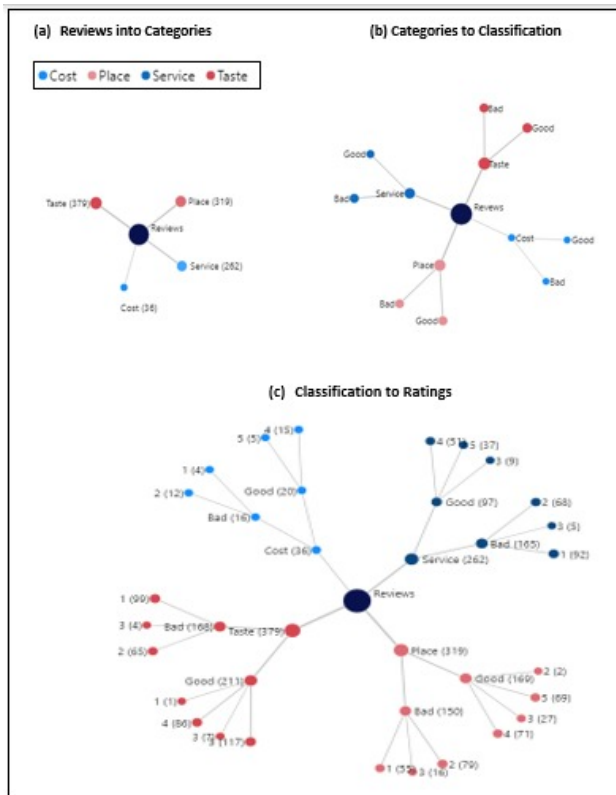

Fig. 5: Classification of Reviews After Analyses

6 Conclusion

This paper represents a framework for restaurant customer's reviews sentiment analysis and visualization on review categories. The data set used for this purpose has taken from kaggle, which consists of reviews from different customers for different restaurants. The model was based on NPL techniques for

tokenization and extracting the textual features from the reviews. The framework is designed by implementing the four Machine Learning algorithms NBC, SVM, LR, and Random Forest.

The visualization of the final output consists of Sankey Chart and Tree Map. For Future work, we tend to extend the model classification and visualization to the real-time data, which would automatically visualize the reviews as soon the customer posts the review.
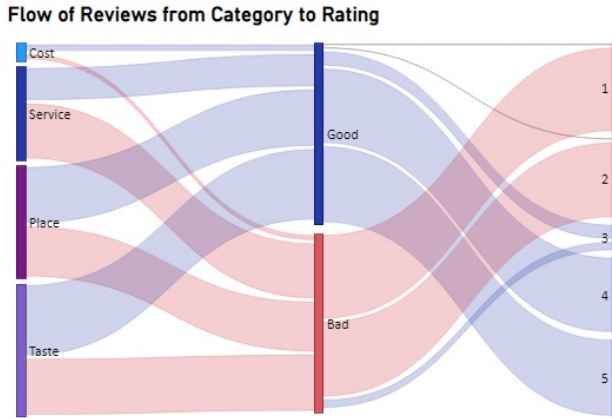
Fig. 6: Connectivity Between Raw Reviews to Category, Classes and Ratings
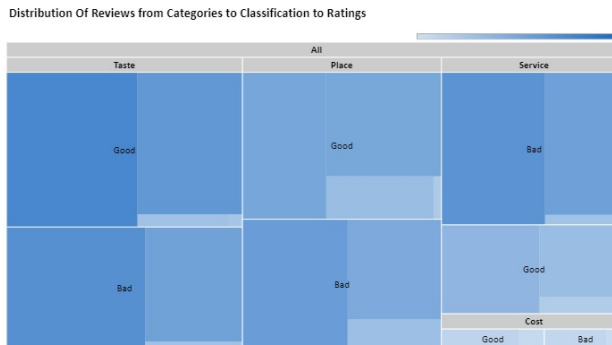
Fig. 7: Hierarchical Patterns

References

[1] Eftekhar Hossain, O. S. (2020). SentiLSTM: A Deep Learning Approach for Sentiment Analysis of Restaurant Reviews.

[2] Kanwal Zahoor, N. B. (2020). Sentiment Analysis and Classification of Restaurant Reviews using Machine., (p. 7).

[3] Mohanad Alhadethy, M. H. (2021). Sentiment Analysis of Restaurant Reviews in Social Media using Naïve Bayes. Systems Analysis Modelling Simulation · January 2021, 8.

[4] Priya Kamath B., G. M. (2021). Empirical Evaluation of various ML algorithms for classification of online Restaurant reviews. 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), (p. 4).

[5] M. Lazarevic, Stability and stabilization of fractional order time delay systems, Scientific Technical Review, 61 (2011) 31-45.

[6] Rachmawan Adi Laksono, K. R. (2019). Sentiment Analysis of Restaurant Customer Reviews on Trip Advisor using Naive Bayes. 12th International Conference on Information,Communication Technology and System (lCTS) 2019, (p. 7).

[7] Adnan, M., Sarno, R.,Sungkono, K. R. (2019). Sentiment Analysis of Restaurant Review with Classification Approach in the Decision Tree-J48 Algorithm. 2019 International Seminar on Application for Technology of Information and Communication (iSemantic). IEEE.

[8] Boya Yu, Jiaxu Zhou, Yi Zhang, Yunong Cao.Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews