Food Inspector - Predict Restaurant Inspection

Daksh Jain - E19CSE121, Mihir Sirpaul - E19CSE217,
Sidharth Mittal - E19CSE411, Vivan Singh Chouhan - E19CSE121

Introduction

Nowadays most of the cities have hundreds of restaurants, so food inspection has become a necessity. like in our country India, where the population is high and the number of restaurants is also high, so manual food inspection is very random and inefficient so many places which are very unhygienic are left over and a lot of people consume that unhygienic food. So instead of manual inspection, many people post their reviews on different social websites like Zomato, swiggy, Justdial and based on those reviews, we can focus on the restaurants which are tagged as unhygienic in the reviews. So, sir our main target is to develop a model which will gather data from these social websites and will narrow down our search. Our algorithm will detect words, phrases, and different such patterns or features to help make the restaurants as unhygienic and also help the food inspection department to focus on these restaurants.

Motivation -

The content that social media establishes creates new opportunities for people and companies. Currently, the restaurants tend to place a strong emphasis on customer feedback in order to increase their business.

Consumers or users are frequently asked to do a survey that is advertised at their app and websites which users are prompted for. This strategy enables restaurants to collect information on their user's demographics, satisfaction, and efficiency of their restaurants. This feedback, though, can be problematic. To begin with, the questions asked of the customer may be biassed. Furthermore, the number of the sample size is frequently too small because many users usually ignore the feedback forms due to which there is less data about these restaurants and also the manual surveys are often financially and statistically impossible and inefficient which gives a hard time to the food inspection office to classify them as hygienic or unhygienic .

Social media platform like zomato,swiggy, twitter and facebook produces a large amount of data(reviews,surveys or comments) on a daily basis, so this data is easily accessible and readable but the problem is that combining this amount off data from these different websites are very time consuming and inefficient.

Big departments and organizations read thousands of online posts per day, making it logistically inefficient to devote people's time to understand, respond, or collect the information. Furthermore, technology can store and analyse text, it cannot comprehend its meaning or viewpoint. As a result, organisations want a system that is both efficient and cost-effective for extracting easily available information from social media.

Background -

Since most of the restaurants in India are now accessible through apps like zomato,swiggy and justdial. So it has become very easy to write reviews for the restaurants throughout the whole city. Since most of these apps focus on the review from the customers whether for delivering food or dining. So we can easily get data from these websites which is the most important aspect of our project i.e, our training data. In countries outside India like USA,Uk many dedicated websites are there for the customer reviews like YELP reviews and many more, but in our country there are very few websites to get this data and if there are these websites, it is not easy to get the reviews for these websites. This training data will help us classify these reviews by searching and matching patterns or words which will help our model to do this predictive task.

What is the motive –

The goal of our project is to provide a way for extracting business intelligence from social media reviews. We use different websites and apps to review to quantify textual data using a discrete scale of one to five to indicate consumer satisfaction . To communicate negative to positive, a numerical number is assigned. Businesses may track customer satisfaction patterns over time and make informed decisions based on their findings and learn what the general public thinks about the company. Furthermore, our technique seeks to extract essential aspects from evaluations that contribute to highly satisfied/dissatisfied ratings, allowing us to see their strengths more clearly and their flaws. Our model will help the food inspection department to label the restaurants as hygienic or unhygienic.

Challenges -

Health inspections are largely random, as they are in most cities, which can result in greater time spent on spot checks at clean restaurants that have been following the laws so there are missed opportunities to enhance health and hygiene at restaurants with more severe food safety concerns. Meanwhile, millions of customers visit these restaurants each year and leave reviews about their experiences. These studies' findings have the potential to improve the City's inspection operations and change the way inspections are focused. Our main challenge is to identify whether the reviews posted by the customers are negative or positive for the restaurants and then on the basis of this identification we tag the restaurants as hygienic and unhygienic.

So, our first challenge is the Data Collection that is to find or collect the appropriate dataset which is properly applicable to our model. The appropriate dataset means the dataset that mainly contains the Username ID and the respective reviews of the user on the restaurants where he visited. Then our second challenge is Data Preprocessing as it is considered to be the most important step in machine learning. As we have to clean the data by removing and pre-processing the things that we don’t want to train for the better accuracy ahead. Then our third challenge is going to be Train as well as Test The Data. According to the model that we are going to use, we need to identify how much percent of data we need to train and how much percent of data we need for the testing purpose. And our most important challenge is to build a model which will provide us with the predictions/evaluations on the basis of the training and testing data which also evaluate the accuracy of the model using the testing data. One most important challenge is how we are going to identify the reviews(which are in the text form) whether they are positive or negative and at the end give hygienic and unhygienic tags to the restaurants.

Related Work

DrivenData created a prediction challenge in collaboration with Yelp and Harvard, as well as the City of Boston, to link Yelp reviews and ratings to the results of Boston's sanitary inspections. The purpose was to leverage social media data to narrow the search for health code violations in Boston, identifying the words, phrases, ratings, and trends that indicate infractions and assisting public health inspectors in their work. Individuals and organisations are increasingly leveraging the material in various media for decision-making, thanks to the phenomenal rise of social media on the Internet.

Sentiment Analysis is a branch of research that includes natural language processing, data mining, and text mining (Farhadloo & Rolland, 2016). It is frequently used to evaluate words based on people's writing patterns in order to uncover positive, negative, or neutral sentiments. Sentiment Analysis is to determine how people feel about something based on their text.

Algorithm and the Flow Diagrams-

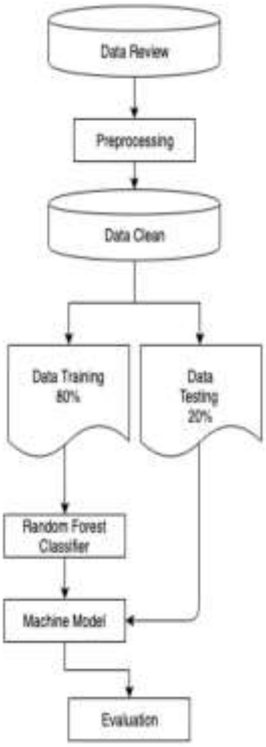
The data that we will be collecting will be analyzed using the random forest in Python Scikit library and analyse it with the precision-recall. The methods that we will be using for developing our models:-

Data Collection -

The data that we will be using will be extracted from different social media websites like swiggy,zomato,food panda and different twitter threads. The data that we will be collecting for a particular city will be analysed. We will be collecting several reviews from these websites and we will split them by their ratings. So we will be classifying the reviews as positive if their rating is above or equal to 3 and negative if their rating is below 3. There will be some imbalanced datasets but from research by Yusran, Juliana, and Bern these datasets will not have any significant effect on the accuracy of our model.

Workflow Process –

So, we will be processing our model in Python 3.6 using the anaconda software. For implementing the program of the Sentimental Analysis we are going to use the Random Forest method with the help of the Scikit-Learn library. The Work Flow Process of our project/model is shown in the following Block Diagram: -

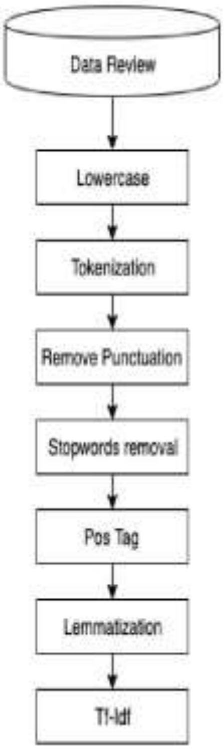


We pre-process text from the data review, and if there is clean data, we split it into 80 percent training data and 20 percent test data. Then we use Random Forest to train the remaining 80% of the data. We test the machine model for data testing and evaluate the accuracy and precision once it has been trained and at the last precision-recall to check how our machine model's metrics change.

Text Preprocessing –

This will be our first step for text mining. Text preprocessing is used to convert unstructured text documents into structured data that may be used in processes by removing noise, homogenising word forms, and reducing word volume (Putraranti & Winarko, 2014). Lowercase, tokenization, and other stages of text preprocessing were used in this investigation.

We need to understand the emotion associated with words because our data is made up of subjective reviews. As a result, we use sentiment analysis to determine whether each review conveys a favourable or negative mood.Sentiment analysis varies from text mining in that it aims to extract and categorise opinion rather than factual data.



Proposed Methodology

Features and Data Collection

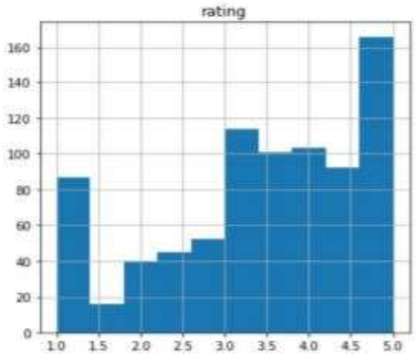
The most difficult task for us was the data collection we had to manually collect data from the restaurants i.e reviews. So we basically collected reviews from the top 15-20 restaurants of the major cities of India. In the data scraping process the different features of our dataset like Restaurant Code, Restaurant Name,Locality,Cuisines ,Votes ,Average Price, Aggregate scores were available in multiple datasets but the main problem was the reviews of the customers. So each of us picked the top restaurants from different cities and collected multiple reviews of each restaurant. So we were able to collect 3000 reviews from different

restaurants. Every row represents reviews and other features for a particular restaurant.
Our dataset has 17 features for each row representing a particular restaurant:-

1.) Location of the restaurant:- Address, Locality,City,Phone No.
2.) Type of Restaurant - Name , Online Ordering, Book Table,rest_type,Cuisines,Listed_in(type)
3.) Rating of the restaurant - Reviews,Aggregate Rating, Rating from each Customer.

Feature Engineering and Data Preprocessing
Pre-processing is a set of procedures that eliminates extraneous language such punctuation, repetitive letters, and plural forms.After successfully scraping the data from the zomato website which included the User reviews and their respective ratings. We started cleaning our data,since we know that all the features were not necessary for our model so we dropped many unnecessary features like we can see in our dataset there is no need for url features and many more like that so we dropped these features. Next step was to remove the duplicates from our data since it creates data inconsistency and decreases the accuracy of our model.Now removing the Nan values from our dataset.Now changing the column names to smaller text since it consumes a lot of time to write these big column names like changing ‘approx_cost(for two people)’ to cost, ‘listed_in(type)’ to type ,’listed_in(city)’ to city etc.Now we performed some transformation in our data like zomato[‘cost’] = zomato[‘cost’].astype(str)(converting cost to str type for consistency in our data) and identify whether the column is of float type, int type or str stype so that we can perform the operations over these columns easily and there is no inconsistency in our dataset or in any column of any restaurant.Next task in our data preprocessing was the conversion of ratings to scale of 5 since we know many customers can give a rating higher than 10 so we converted those ratings by dividing them by 2 and adding this corrected rating to our dataset .Now We know many customers give ratings inside their reviews also like ⅘ for food etc. so we had to drop these reviews so that the algorithm doesn’t get confused. Now we had the review list which contained the reviews, ratings in a list type so we had to retrieve the text from the list using slicing and then we stored our data in a pandas dataframe.Encoding was done on columns like book table, online order, rest type, and listed in using the sklearn library's One Hot Encoding (city).

Data Distribution
We used ratings for each review as a proxy for action taken to look at the distributions of the features because action done is directly reliant on the rating given to the restaurant.We created a graph of each restaurant and the ratings as weights. The colour red denotes a greater score and the colour green denotes a lower score, with a minimum of 0. A higher score implies that a restaurant has more violations, both in terms of severity and number.
After analysing the whole dataset we have distributed the features like rating(out of 5) on the basis of the positive and negative case, which means that how many ratings are negative(0-2 ratings) or positive(3-5 ratings) for the particular restaurant. If the intensity of the negative rating is high for the restaurants of that particular area then that area will be considered as the unhygienic for the food or if the if the intensity of the positive rating is high for the restaurants of that particular area then that area will be considered as the hygienic fo the food.



Algorithm and Proposed Method

Linear Regression
By fitting a linear equation to observed data, linear regression seeks to model the relationship between two attributes. One variable is treated as an explanatory variable, while the other is treated as a dependent variable. A modeller might, for example, use a linear regression model to match people's weights to their heights. A modeller should first

evaluate whether or not there is a link between the variables of interest before attempting to fit a linear model to observed data.This does not necessarily indicate that one variable causes the other (for example, greater competitive paper scores do not necessarily imply higher college grades), but rather that the relationship between the two variables has some significant association strength. If the given dependent and independent variables seem to have no relationship (i.e., the plot shows no increasing or decreasing trends), then using and fitting a LR model to data is very unlikely to give a useful model.

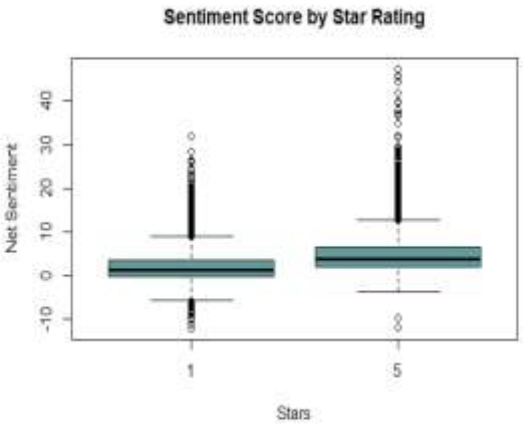
The correlation coefficient, having a value between -1 and 1 shows the strength of the link of the observed data for the two variables, is a useful numerical measure of association between two variables. The equation for a linear regression line is y = ax + c, with x as the independent variable and y as the dependent variable. The intercept is c, while the slope of the line is a.

Sentiment Analysis

We must comprehend the emotion associated with words because our data comprises subjective reviews. As a result, we use sentiment analysis to determine whether each review expresses a good or negative emotion. As a result, we use sentiment analysis to determine whether each review expresses a good or negative emotion. According to the Indian Institute of Management,"user-generated text content to extract opinion and subjectivity knowledge."Sentiment analysis is different from text mining since it aims to extract and categorise information instead of current knowledge.

When you look at specific reviews, it becomes evident why technology can't understand language: sarcasm. When we were extracting reviews from zomato we found specific user reviews which makes the data confusing and inconsistent, like one user wrote , "This food is michael jackson bad that means it's amazing" This review contains both a very favourable adjective, "amazing," and a negative term, "bad," as understood by a native English speaker. While a phrase composed in this fashion communicates considerable disappointment and sarcasm to an English-speaking mind, a machine is unable to recognise such sarcasm and sentence structure. As a result, we must create algorithms that can recognise patterns and learn from language.

We cannot simply check for the existence or sabbatical of a word to take out the overall meaning of a sentence; rather, we must comprehend it.The links between terms and their combinations As a result, we deploy sentiment analysis to assign numerical numbers to terms/concepts in order to reflect the emotion.We assign an affinity score to each term, a numerical value that reflects the term's amount of positive or negative emotion.   To reflect negative or positive emotion, the system assigns a sentiment score of {-1,1} to each phrase . The algorithm takes into account a cluster of words around each polarised term, including four words preceding the polarised term and two words following it. The words that surround the polarised term are labelled as neutral terms, negators, amplifiers, or de-amplifiers. De-amplifiers convey a weaker display of emotion and so decrease the polarity, whereas amplifiers show a stronger expression of emotion and thus raise the polarity. If the number of negators is odd, the sign of the polarity is Negators ip.The programme accounts for sentiment communicated in sentences by forming a cluster around each polarised word. We next add the score as a column to our matrix by summing the sentiment of each polarised cluster inside each review. The frequency of each individual phrase, the customer experience themes, and the net sentiment make up our Complete Matrix.As expected, we find that the 5 stars have a larger net sentiment than the one stars. However, there is still overlap between the net sentiment of one and the 5 stars, making it impossible to make a clear distinction based just on net sentiment.



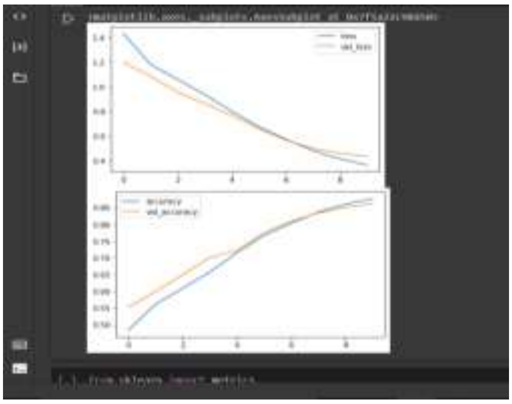
Result & Analysis

Accuracy and Metrics

From our dataset we split our data into 90% training and 10% testing data.

- **Smaller Dataset(3000 rows)**
  Accuracy :The results of this model is having **86.19%** accuracy

  ROC Curve :



```
[ ] from sklearn import metrics
```

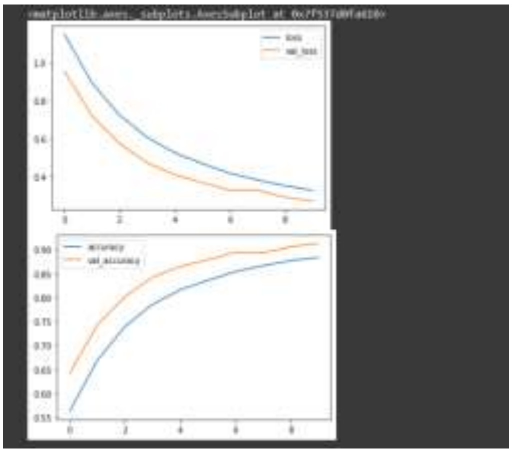
Metrics: .
Test R2: 0.7668
Train R2: 0.8748

Test MSE: 0.0217
Train MSEL 0.0129

Test RMSE: 0.1475
Train RMSE: 0.1136

- Bigger Dataset(25000) :
  Accuracy : The result of this model is having **91.31%** accuracy.

  ROC Curve:



Metrics:
Test R2: 0.8322
Train R2: 0.8899

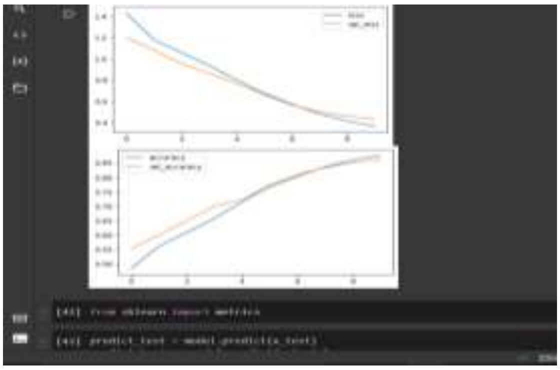
Test MSE: 0.0144
Train MSE: 0.0098

Test RMSE: 0.1200
Train RMSE: 0.0989

Now we tried changing the optimizers and the layers of our model using the smaller dataset to test the accuracy in these different cases:

- **Changing the Epoch**:
  We first trained the model using the adam optimizer and the epoch set to 15, here are the results:
  Accuracy :
  The results of this model is having **88.94%** accuracy.
  Roc Curve:

Metrics:



```
[41] from sklearn import metrics
[41] predict_test = model.predict(x_test)
```

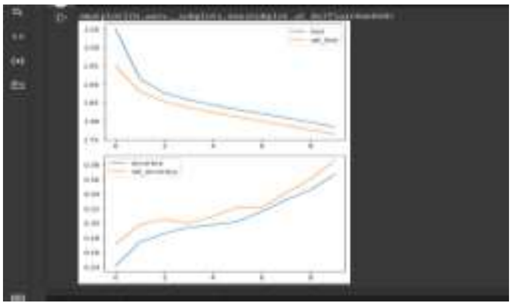
Test R2: 0.7668
Train R2: 0.8748

Test MSE: 0.0217
Train MSE: 0.0129

Test RMSE: 0.1475
Train RMSE: 0.1136

- **Changing the Optimizer and the Layers:**
  We changed the optimizer from adam to sdg to test out the accuracy for our model.
  Accuracy : 44.01%

ROC Curve:



Section 5: Conclusion

We are able to find the level of the satisfaction of the Food lovers using sentiment analysis and predictive modelling and also successfully able to give the hygienic and unhygienic tag to the restaurants.

The hygiene and freshness of the kitchen has a significant impact on the degree of pleasure of food customers as well as the staff's friendliness. The interesting thing is that the food's taste isn't talked about often in junk food restaurant reviews.

This could be due to the customer's expectations when they come into a junk food restaurant or any other restaurants . Although the quality and taste of a fast food restaurant are usually known before entering, the setting in which the consumer finds himself may be unexpected or vary. As a result, while looking to improve customer happiness, junk food restaurants should stress a friendly, clean environment.

Problem Statement

The main problem that we identified is with the Food Inspection Department. As the number of restaurants is high, manual food inspection is very random and inefficient so many places which are very unhygienic are left over and a lot of people consume that unhygienic food which is the problem and also so much time is wasted for the department through manual inspection.

Our project is helping the Food Inspection Department by providing the details and tags about the restaurants whether that restaurant is hygienic or unhygienic which saves a lot of time of the department by avoiding them to do the manual inspection of the restaurants.

REFERENCES

Putraranti, N. D., & Winarko, E. (2014). Support Vector Machine. https://doi.org/10.22146/ijccs.3499

Farhadloo, M., & Rolland, E. (2016). Fundamentals of sentiment analysis and its applications. https://doi.org/10.1007/978-3-319-30319-2_1