"Sentiment Analysis on Reviews of Online Food Ordering System"

Developed For
FCAIT, *i*MSc(IT)

Project Report (Sem – VI)
Submitted For
The Partial Fulfillment Towards
The Degree of
Integrated Master of Science (Information Technology)
***i*MSc(IT)**

By
Dhruvi Acharya, A01
Vanshika Amin, A04
Dhwani Sanghvi, A52

Under the Guidance of

External Guide

Internal Guide
Disha Shah
FCAIT, *i*MSc(IT),
Ahmedabad

Faculty of Computer Applications & Information Technology
***i*MSc(IT) Programme, Ahmedabad.**

CERTIFICATE

This is to certify that
1) Dhruvi Acharya
2) Vanshika Amin
3) Dhwani Sanghvi

Student/s of Semester- VI Integrated Msc(IT) [TY iMSc(IT)], FCAIT, GLS University has/have successfully completed the
Mini Project
on
"Sentiment analysis on reviews of online food ordering system"
as a partial fulfillment of the study of Third year Semester-VI,
Integrated Master of Science (Information Technology) [iMSc(IT)]

Date of Submission: 15th march 2022

Prof. Disha Shah
(Project Guide)

Prof. Tripti Dodiya
(Project Co – ordinator)

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our teacher “Prof. Disha Shah” as well as our Dean “Prof. Tripti Dodiya” who gave us the golden opportunity to do this wonderful project on the topic “Sentiment Analysis on reviews of online food ordering application”, which also helped us in doing a lot of Research and we came to know about so many new things. Secondly, we would also like to thank our parents and friends who helped us a lot in finalizing this project within the limited time frame.

TABLE OF CONTENTS

ABOUT SENTIMENT ANALYSIS

(1) Definition of Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing (NPL) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback , and understand customer needs.

(2) What is Sentiment Analysis ?

Sentiment analysis is the process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers.

(3) Why is Sentiment Analysis important ?

Improving sales and retaining customers are core business goals. According to research by Apex Global Learning, every additional star in an online review leads to a 5-9% revenue bump. There's an 18% difference in revenue between businesses rated as three-star and five-star ratings. Sentiment analysis can help you understand how people feel about your brand or product at scale. This is often not possible to do

manually simply because there is too much data. Specialized SaaS tools have made it easier for businesses to gain deeper insights into their text data. This could include everything from customer reviews to employee surveys and social media posts. The sentiment data from these sources can be used to inform key business decisions.

(4) Benefits of Sentiment Analysis

The applications of sentiment analysis are endless. So, to help you understand how sentiment analysis could benefit your business, let's take a look at some examples of texts that you could analyze using sentiment analysis.
Then, we'll jump into a real-world example of how Chewy, a pet supplies company, was able to gain a much more nuanced (and useful!) understanding of their reviews through the application of sentiment analysis.

- Sorting Data at Scale - Can you imagine manually sorting through thousands of tweets, customer support conversations, or surveys? There's just too much business data to process manually. Sentiment analysis helps businesses process huge amounts of unstructured data in an efficient and cost-effective way.
- Real-Time Analysis - Sentiment analysis can identify critical issues in real-time, for example is a PR crisis on social media escalating? Is an angry customer about to churn? Sentiment analysis models can help you

- immediately identify these kinds of situations, so you can take action right away.
- Consistent criteria - It’s estimated that people only agree around 6065% of the time when determining the sentiment of a particular text. Tagging text by sentiment is highly subjective, influenced by personal experiences, thoughts, and beliefs.
- By using a centralized sentiment analysis system, companies can apply the same criteria to all of their data, helping them improve accuracy and gain better insights.

(5) Limitations of Sentiment Analysis

Subjective opinions and feelings expressed towards your brand represent only a fraction of the information your care about when managing the reputation of your organization. Very often, online mentions of your brand will have an impact on your reputation regardless of whether any sentiment is expressed in the text.

• Sentiment Neutral Statements with Negative or Positive Reputational Impact

• Incorrectly Targeted Sentiment

• Sarcastic tone in the comment

• Review Language is Dissimilar to Social Media and News Language

(6) Applications of Sentiment Analysis

Sentiment analysis is one of the most popular ways to analyze text, such assurvey responses, customer support issues, online reviews, and live chats, because it can help companies stay on top of customer satisfaction.

- Social media monitoring
- customer support management
- analyzing customer feedback
- Brand monitoring and reputation management
- Listen to voice of the customer
- Listen to your employees
- Product analysis
- Market and competitor research

(8)

PROJECT PROFILE

Our project is basically sentiment analysis using reviews on food ordering system. We have performed following tasks:

1) we have firstly classified the reviews on the basis of rating provided in the dataset
   We have done classification into 5 categories such as best, good, average, bad and worst.
   The lowest rating belongs to worst category and the highest rating belongs to best category and so wise.
   We have made a webpage to display the result of above mentioned functionality. We are providing the one word review as well as star ratings to the corresponding restaurant the user searches for.

2) The second task we performed is sentiment analysis on the reviews. We retrieved the restaurant dataset using pandas library and and performed data processing and data cleaning functions.

   We used data processing functions for :
   - to drop unnecessary columns
   - changing datatypes
   - names of the columns.

We performed data cleaning functions :

- remove punctuations
- remove stopowords
- remove urls and all other unnecessary data

Then we classified the reviews into 3 categories such as positive negative and neutral on the basis of ratings. The lowest review belongs to negative category and highest rating to positive category. We have then trained the model and tested it using machine learning algorithms.

We have tried making unsupervised predictions using supervised learning method.

DEVELOPMENT TOOLS AND TECHNOLOGIES USED

(1) Modules Imported

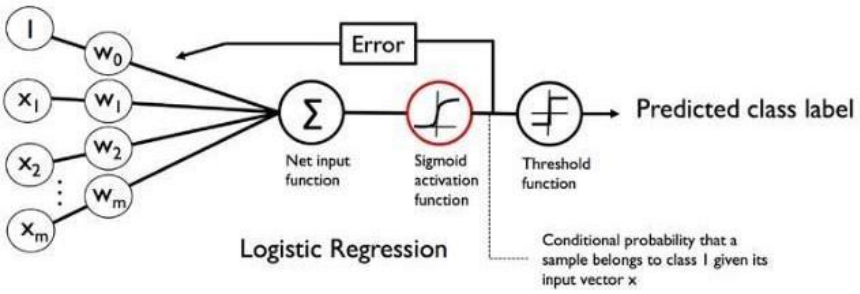
For text processing and cleaning

- Matplotib
- Re
- string
- Numpy
- Pandas For taining and testing model
- from sklearn.feature_extraction.text import TfidfVectorizer
- from sklearn.linear_model import LogisticRegression
- from sklearn.pipeline import Pipeline
- from sklearn.metrics import accuracy_score,precision_score

(2) Algorithms Used

1. LOGISTIC REGRESSION
   Logistic Regression is a classification model that is very easy to implement and performs very well on linearly separable classes. It is one of the most widely used algorithms for classification in industry too, which makes it attractive to play with.

Very simplistically explained, Logistic Regression works as follows:



First we will define the input for our algorithm. The imput will be each sample in whatever dataset we are working with. Each sample will consist of several features. For example, if we're working with housing price prediction, the features for each sample could be the size of the house, number of rooms, etc. We'll call the input vector X.
For the algorythm to learn, we need to define variables that we can adjust accordingly to what we want to predict. We will create a vector of weights (W) that the model will adjust in order to predict more accurately. The process of adjusting those weights is what we call learning.
For every input sample, we will perform a dot product of the features by the weights XW. This product is sometimes referred as net input. This will give us a real number. Since in this particular problem we want to classify (positive/negative), we need squash this number in the range [0, 1].

This will give us the probability of a positive event. A function that does precisely that is called sigmoid.

2. TF-IDF VECTORIZER
   TF-IDF stands for Term Frequency — Inverse Document Frequency and is a statistic that aims to better define how important a word is for a document, while also taking into account the relation to other documents from the same corpus.This is performed by looking at how many ties a word appears into a document while also paying attention to how many times the same word appears in other documents in the corpus.

   The rationale behind this is the following:
   - a word that frequently appears in a document has more relevancy for that document, meaning that there is higher probability that the document is about or in relation to that specific word
   - a word that frequently appears in more documents may prevent us from finding the right document in a collection; the word is relevant either for all documents or for none. Either way, it will not help us filter out a single document or a small subset of documents from the whole set.

   So then TF-IDF is a score which is applied to every word in every document in our dataset.

And for every word, the TF-IDF value increases with every appearance of the word in a document, but is gradually decreased with every appearance in other documents. And the maths for that is in the next section.

TF-IDF Formula Explained :-
Now let's take a look at the simple formula behind the TF-IDF statistical measure. First let's define some notations:
• N is the number of documents we have in our dataset
• d is a given document from our dataset
• D is the collection of all documents
• w is a given word in a document

First step is to calculate the term frequency, our first measure if the score.
$Tf(w,d)=\log(1+f(w,d))$
Term Frequency Formula
Here f(w,d) is the frequency of word w in document d.

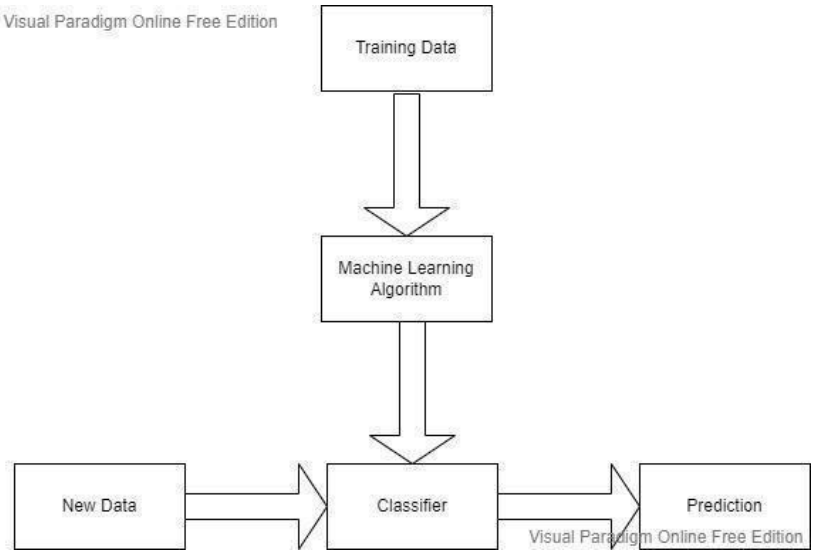
Second step is to calculate the inverse term frequency.
$Idf(w,d)=\log(N/f(w,d))$
Inverse Document Frequency Formula

With N documents in the dataset and f(w, D) the frequency of word w in the whole dataset, this number will be lower with more appearances of the word in the whole dataset.

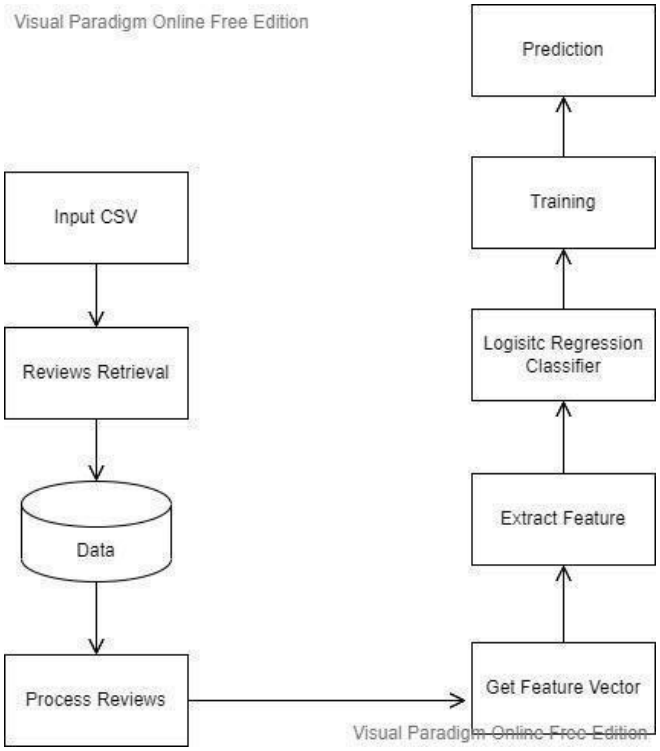
Final step is to compute the TF-IDF score by the following formula:
Tfidf(w,d,D)=tf(w,d)*idf(w,D)

SYSTEM FLOW DIAGRAMS

1.

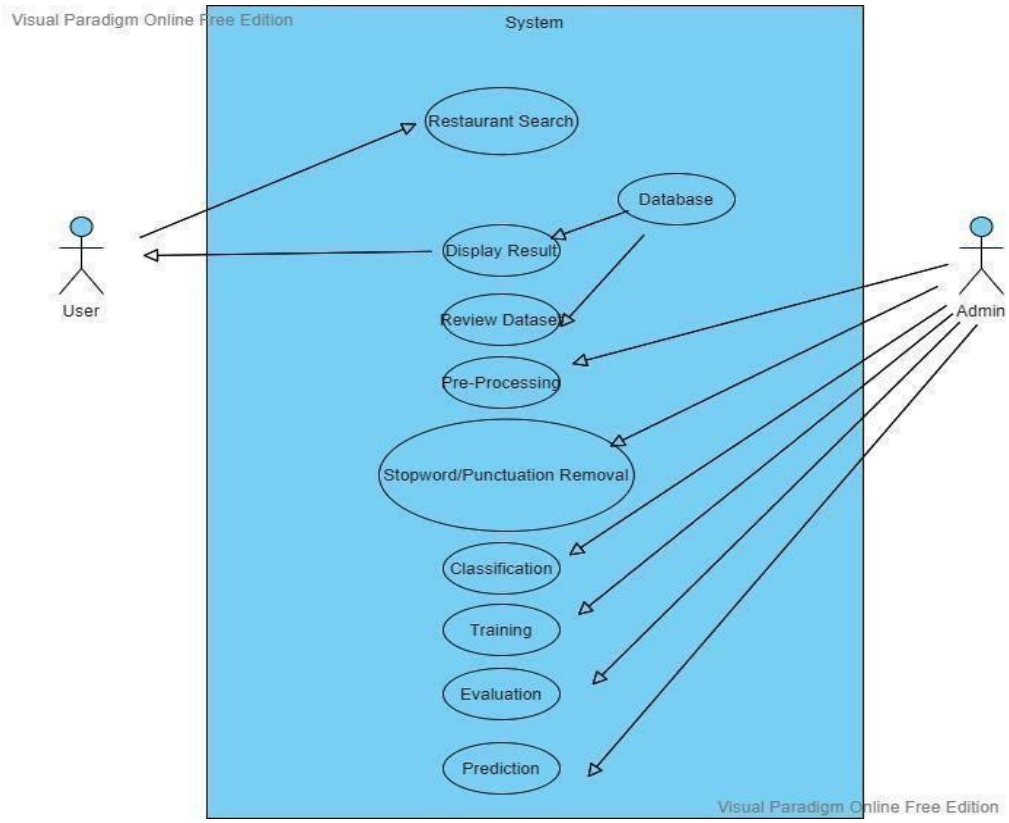


2.

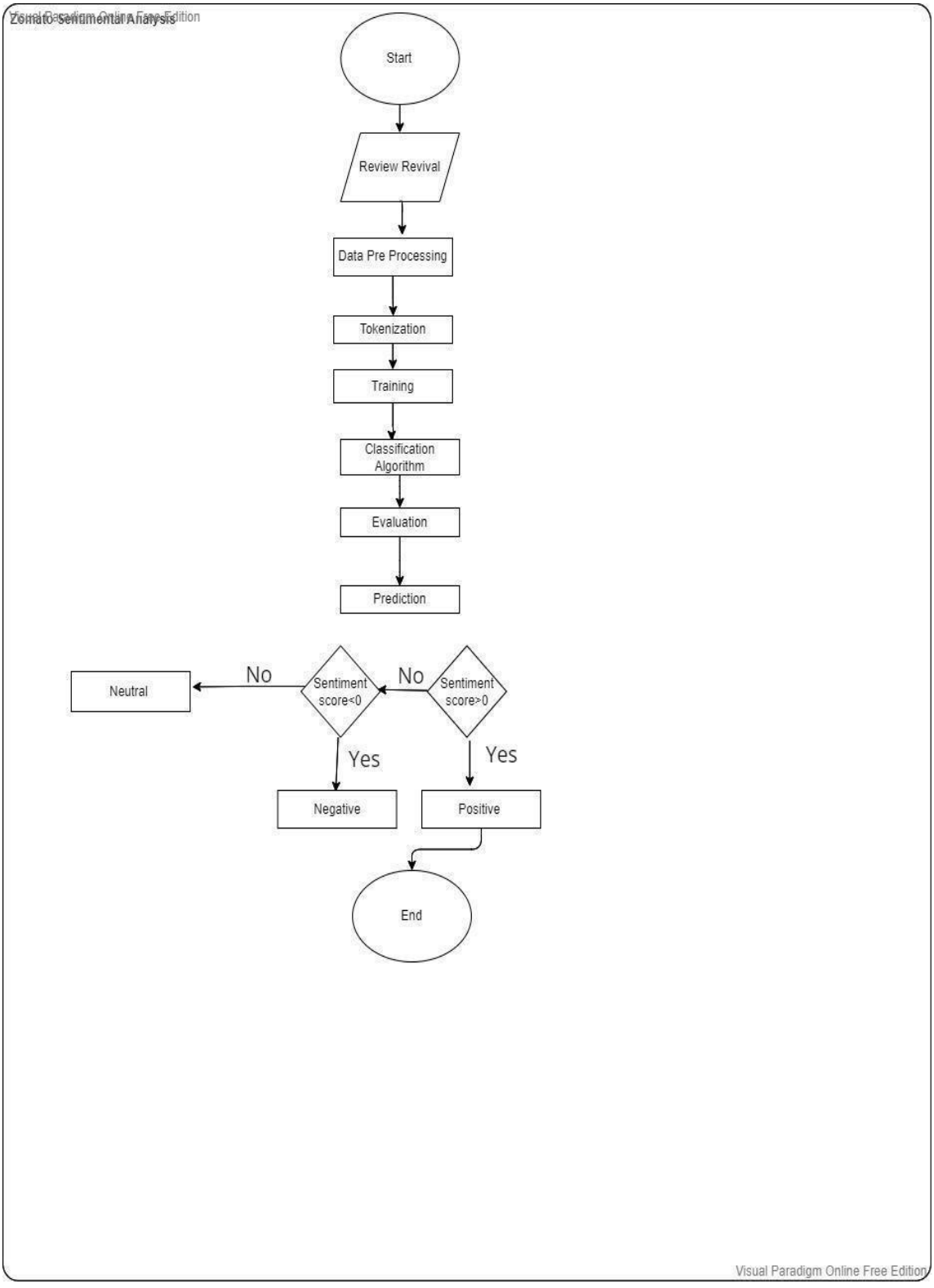


UML DIAGRAMS

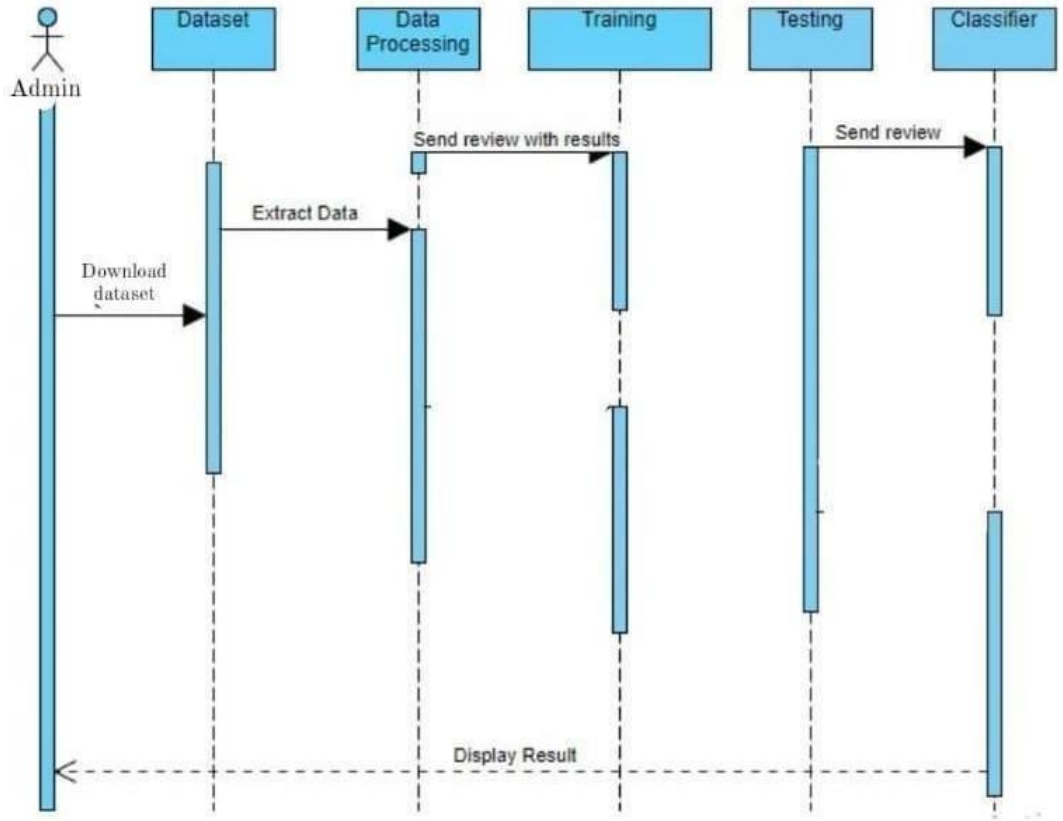
1. Use Case Diagram



2. Activity Diagram



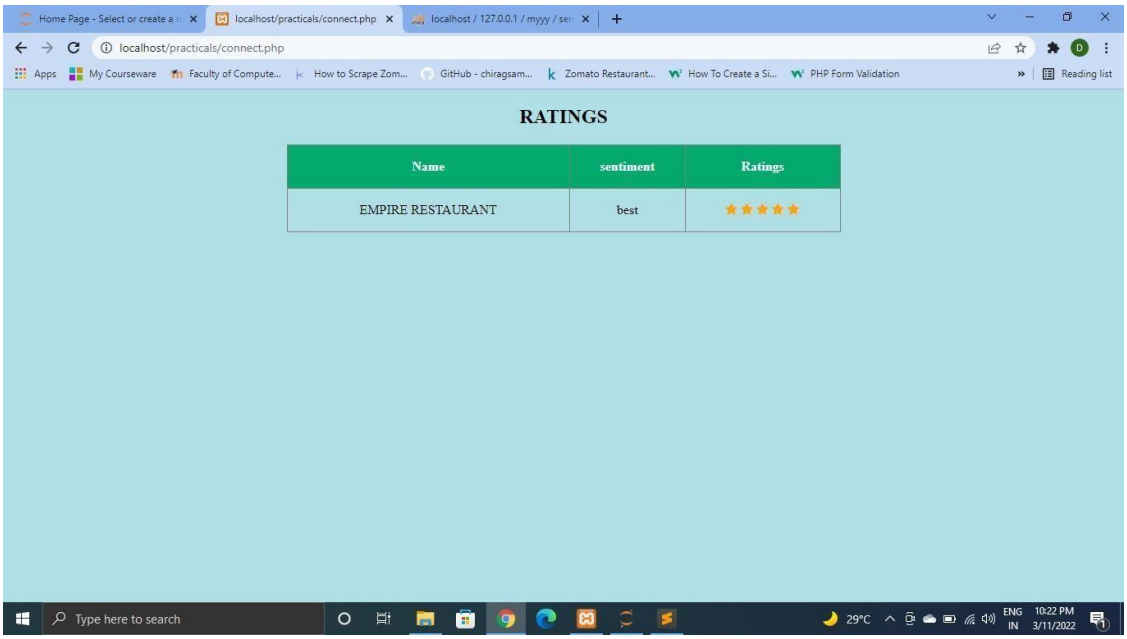
3. Sequence Diagram



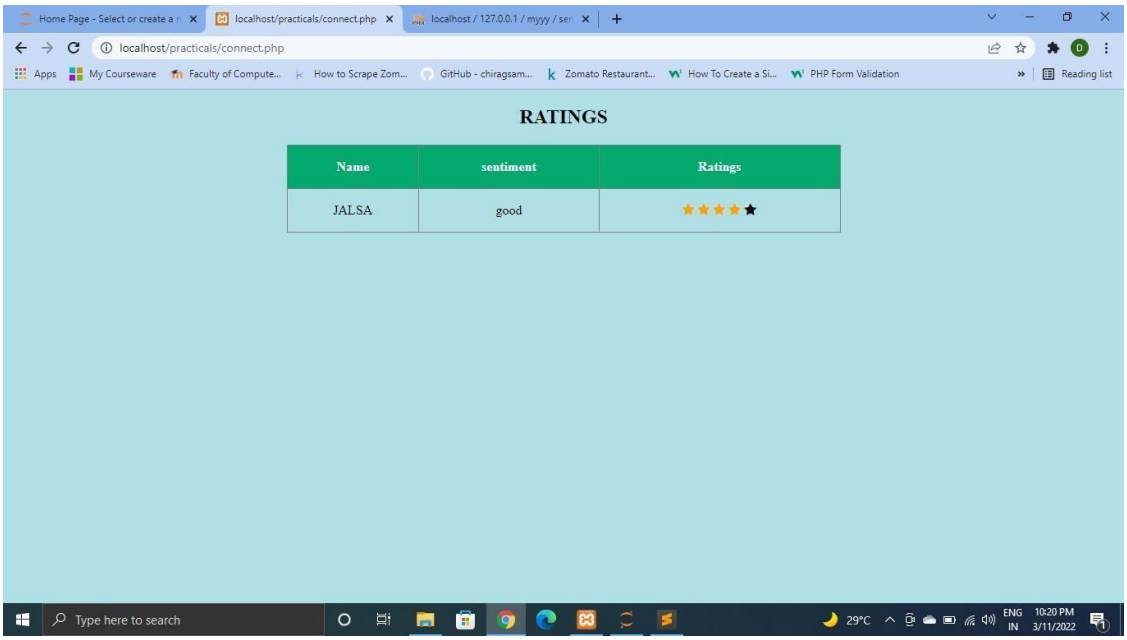
SCREENSHOTS

1. Frontend

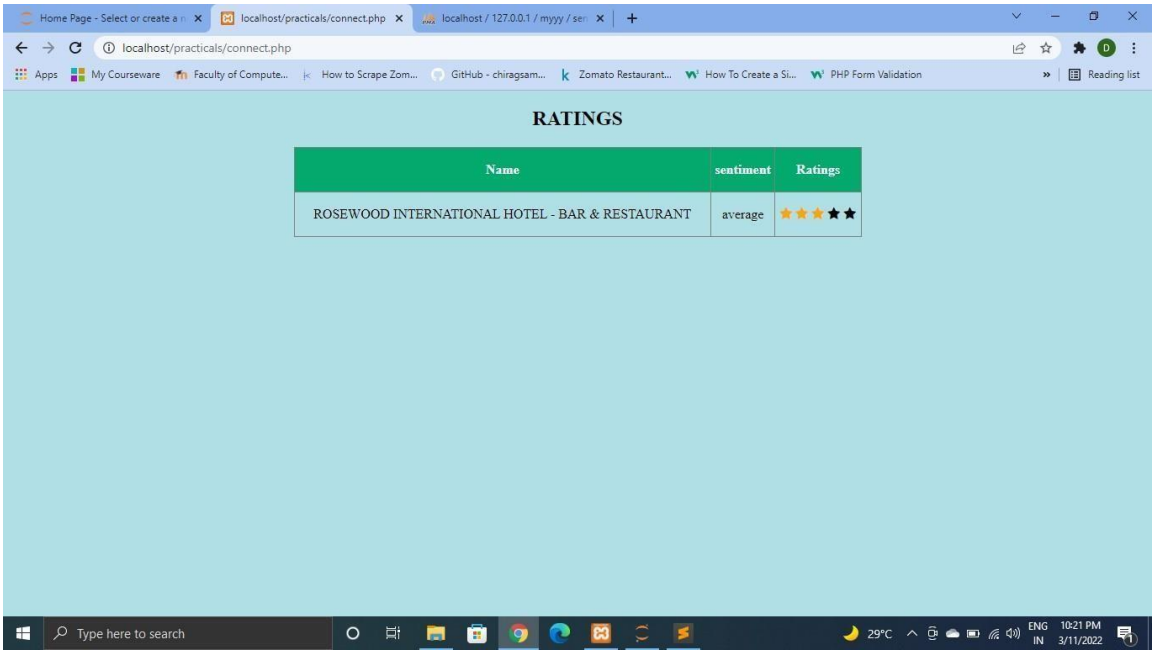
- **Best (5 stars)**



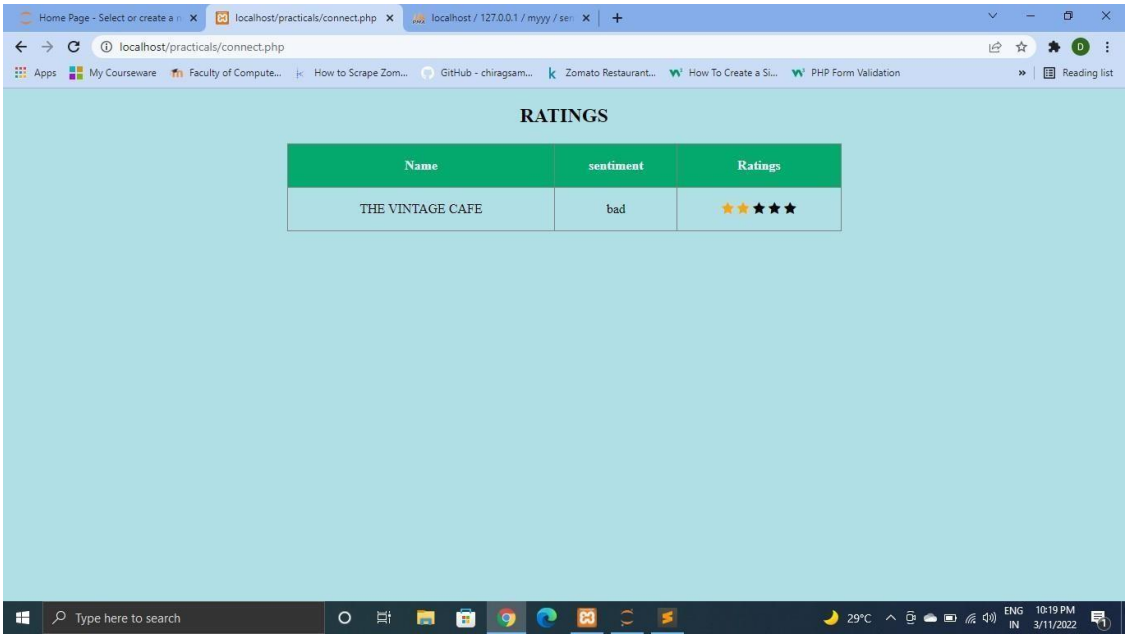
- **Good (4 stars)**



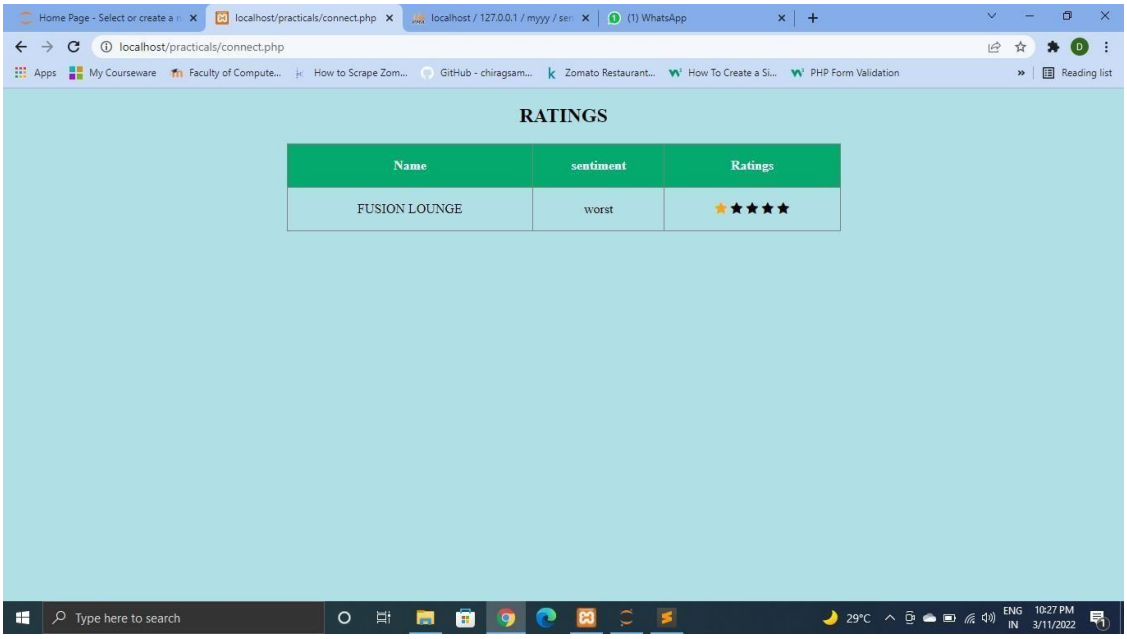
- **Average (3 stars)**



- **Bad (2 stars)**

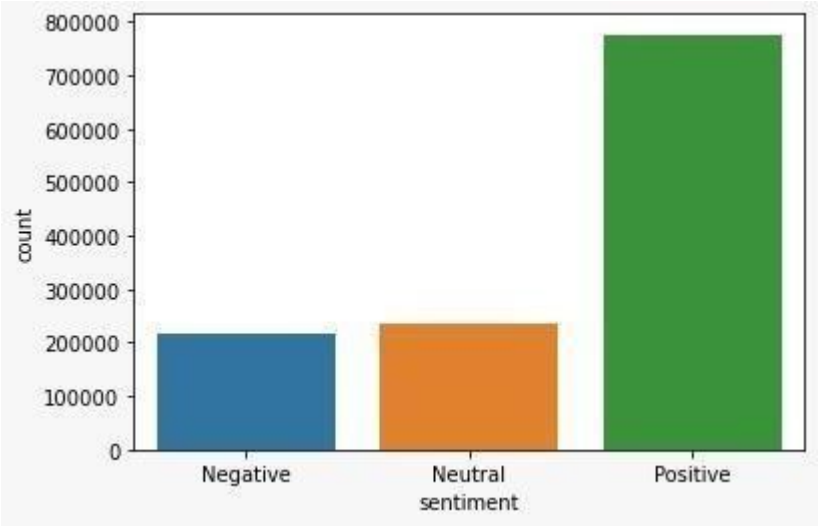


- **Worst (1 star)**



2. Backend

- **Classification**



- **Accuracy Score**

```
In [66]: from sklearn.metrics import accuracy_score,precision_score,recall_score
         print("Accuracy : ",accuracy_score(predictions,Y_test))
         print("Precision : ",precision_score(predictions,Y_test,average='weighted'))
         print("Recall : ",recall_score(predictions,Y_test,average='weighted'))

         Accuracy :  0.8916988662072468
         Precision :  0.8983228217170921
         Recall :  0.8916988662072468
```

- **Predictions**

```
In [66]: from sklearn.metrics import accuracy_score,precision_score,recall_score
         print("Accuracy : ",accuracy_score(predictions,Y_test))
         print("Precision : ",precision_score(predictions,Y_test,average='weighted'))
         print("Recall : ",recall_score(predictions,Y_test,average='weighted'))

         Accuracy :  0.8916988662072468
         Precision :  0.8983228217170921
         Recall :  0.8916988662072468
```

CONCLUSION

We have completed our project using python as language. PHP and HTML for output representation. We have used MySQL for database needs. Although there was a problem in integration of python mysql and php, through number of tutorial we were able to integrate it.

We were able to classify the restaurant on the basis of their reviews into 5 classes and we have displayed the same to the user along with star ratings.

At last we have performed sentiment analysis using supervised learning algorithms and bow are able to make predictions on unsupervised data.

REFERENCES

https://github.com/llSourcell/logistic_regression/blob/master/Sentiment%20analysis%20with%20Logistic%20Regression.ipynb

https://towardsdatascience.com/tf-idf-explained-and-python-sklearn-implementation-b020c5e83275

https://monkeylearn.com/sentiment-analysis/

https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

https://en.m.wikipedia.org/wiki/Sentiment_analysis

https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining?amp=1

https://www.lexalytics.com/technology/sentiment-analysis

https://www.whoson.com/customer-service/top-ten-benefits-of-sentiment-analysis/

https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/