GOVERNMENT DEGREE COLLEGE

BEGUMPET- HYDERABAD

(Affiliated to Osmania University)

STUDENT STUDY PROJECT

ON

Sentimental Analysis of Product Reviews

DEPARTMENT OF COMPUTERS

PRESENTED BY

STUDENTS

B.SC (Data Science) III Year

1.	M. Aishwarya	:	1085-20-539-046
2.	K.Sanjali	:	1085-20-539-038
3.	P. Sanghavi	:	1085-20-539-056
4.	P.Taraka Lakshmi	:	1085-20-539-050
5.	M.S. Jagruthi	:	1085-20-539-041

CERTIFICATE

This is to certify that the project work entitled "***Sentimental Analysis of Product Reviews***" using **Natural Language processing** is presented by **B.SC (Data Science) III Year Students** in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer science by the Osmania University, Hyderabad during the academic year 2022-2023.

The results embodied in this report have not been to any other University or Institution for the award of any degree.

Under the Guidance of
Dr.G.RAJITHA DEVI
Assistant Professor Head, Dept. of CS

ACKNOWLEDGEMENT

I sincerely take it as a privilege to thank the management of our college **GOVENMENTT DEGREE COLLEGE (W), Begumpet** for providing

required facilities during our Project.

I derive my great pleasure in expressing our sincere gratitude to our Principal, **Dr K.Padhmavathi** for his timely suggestions, which helped us to complete the Project successfully.

I take it as a privilege to thank our guide **Dr.GURRAM RAJITHA DEVI Asst.Professor in Computer Science Department**, for the ideas that led to complete the Project and we also thank him for continuous guidance, support and unfailing patience, throughout the course of this work. Her valuable comments during this period have been valuable and worth for a lifetime. It is very auspicious moment we would like to express our gratitude to **Dr G .RAJITHA DEVI**, our beloved**, Lecturer. in Computer Science** for his consistent encouragement during the progress of this Project.

I also thankful to both teaching and non-teaching staff of Computer Science Department for their kind co-operation and all sorts of help bringing out this Project successfully.

DECLARATION

I hereby declare that the project report entitled **“Sentimental Analysis of Product Reviews”** using “**Natural Language processing**” is the work done by us in the campus at **GOVT DEGREE COLLEGE FOR WOMEN(A) Begumpet** during the academic year **2022-2023** and is submitted in partial fulfillment of the requirements for the degree of **Bachelor of science** in **Data Science** by the **Osmania University**, Hyderabad.

ABSTRACT

Sentiment analysis is defined as the process of mining of data, view, review or sentence to predict the emotion of the sentence through natural language processing (NLP). The sentiment analysis involve classification of text into three phase “Positive”, “Negative” or “Neutral”. Sentiment analysis is a broadly employed method for finding and extracting the appropriate polarity of text sources using Natural language Processing (NLP) methods. This paper focuses on examining the efficiency of machine learning technique (Logistic Regression) for classification of online reviews available in the E-Commerce websites.It Aims To Automate The Process Of Gathering Online, End Users Reviews For Any Given Product Or Services And Analyzing Those Reviews In Terms Of Sentiments Expressed About Specific Features . This Involves the Filtering Of Irrelavent and Unhelpful Reviews, Quantification Of The Sentiments Of Thousands Of Useful Reviews. Also we provide a visualization for our result summarization.

CONTENTS

1. Introduction

Every day we come across various products in our lives, on the digital medium we swipe across hundreds of product choices under one category. It will be tedious for the customer to make selection. Here comes 'reviews' where customers who have already got that product leave a rating after using them and brief their experience by giving reviews. As we know ratings can be easily sorted and judged whether a product is good or bad. But when it comes to sentence reviews we need to read through every line to make sure the review conveys a positive or negative sense. In the era of artificial intelligence, things like that have got easy with the Natural Language Processing(NLP) technology.

1.1 Motivation

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral.Understanding people’s emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before.It is quite hard for a human to go through each single line and identify the emotion being the user experience.Now with technology, we can automatically analyzing customer feedback, from survey responses to social media conversations, brands are able to listen attentively to their customers, and tailor products and services to meet their needs.

1.2 Problem Statement

This is the Problem Statement to classify the customer comments. This would be helpful for the organization to understand Customer feedbacks.

Web-portals get vast amount of feedback from the customers. To go through all the feedback's can be a tedious job. You have to categorize opinions expressed in feedback forums. This can be utilized for feedback management system. We Classification of individual comments/reviews.and we also determining overall rating based on individual comments/reviews. So that company can let a complete idea on feedback's provided by customers and can take care on those particular fields. This makes more loyal Customers to the company, increase in business , fame ,brand value ,profits.

1.3. Objectives

1. Reviews Preprocessing and Cleaning
2. Story Generation and Visualization from reviews
3. Extracting Features from Cleaned reviews
4. Model Building: Sentiment Analysis

2. Literature Survey

A detailed literature survey was done as the initial step of our project. And the following are the papers which we referred as a part of it.

- Amazon Review Classification and Sentiment Analysis by Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande

- An Approach towards Feature Specific Opinion
Mining and Sentimental Analysis Across E-Commerce Websites Lijisha T, Archana Balakrishnan, Sumayya K , Arjun P , Renjith Sunny

- Sentiment Analysis Based Requirement Evolution Prediction by Lingling Zhao and Anping Zhao

- Comparative Study of Machine Learning Approaches for Amazon Reviews by Abhilasha Singh Rathora , Amit Agarwalb , Preeti Dimri

- Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning Callen Rain Swarthmore College Department of Computer Science

- SENTIMENT ANALYSIS ON AMAZON PRODUCT REVIEWS by Kopparthi Harika, K. Mani Veera Venkata Ratna Kumari, M.Sai Anusha, M.Anila ,S. Tejaswini

- Sentiment analysis using product review by data Xing Fang and Justin Zhan

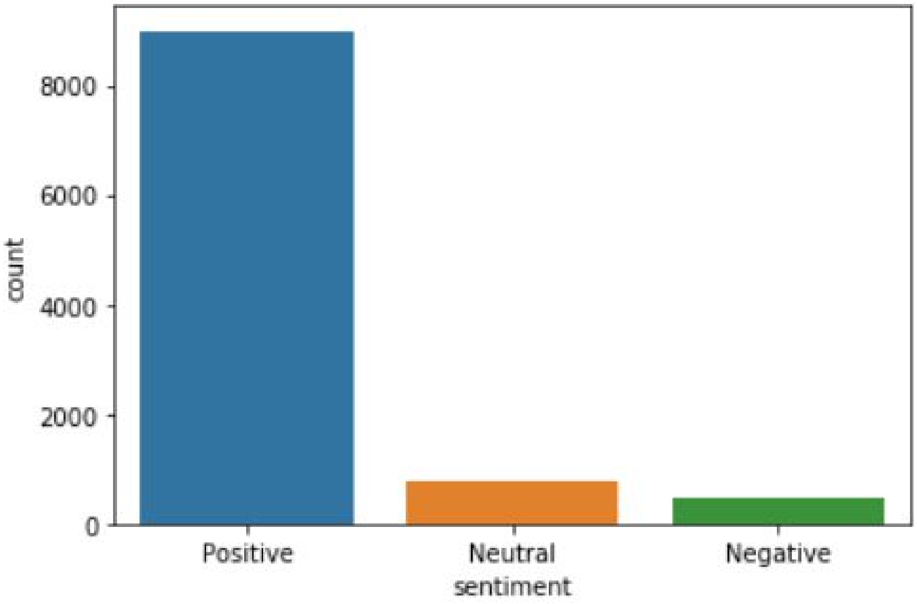
- Sentiment Analysis for Amazon Reviews by Wanliang Tan,Xinyu Wang

3.Analysis

3.1. Methodology

3.1.1. Data Collection

Our dataset comes from Consumer Reviews of Amazon Products . Our dataset comes from Consumer Reviews of Amazon Products. This dataset has 10261 data points in total. Each example includes the type, name of the product as well as the text review and the rating of the product. To better utilize the data, first we extract the rating and review column since these two plays an essential part in analysis. Then , we found that there are some data points which has null values when we went through the data. After eliminating those examples, we have 10227 data points in total. Besides, to have a brief overview of the dataset, we have plot the distribution of the ratings. In it shows that we have 5 classes - rating 1 to 5 as well as the distribution among them.



Also, these five classes are actually imbalanced as class 1 and class 2 have small amount of data while class 5 has more than 8000 reviews.

Here is one sample from our dataset:

Review text: ‘This product so far has not disappointed. My children love to use it and I like the ability to monitor control what content they see with ease.’ Rate: ‘5’ In the subsection ‘3.3 Features’, we will illustrate how we convert a review text into an input vector, and we simply take the rate of a review as its label.

3.1.2. Data Preprocessing

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Major Tasks in Data Preprocessing:

Data cleaning

Data integration

Data reduction

Data transformation

Data cleaning:

Data cleaning is the process to remove incorrect data, incomplete data and inaccurate data from the datasets, and it also replaces the missing values. There are some techniques in data cleaning.

Handling missing values:

Standard values like “Not Available” or “NA” can be used to replace the missing values.

Missing values can also be filled manually but it is not recommended when that dataset is big.

The attribute’s mean value can be used to replace the missing value when the data is normally distributed

wherein in the case of non-normal distribution median value of the attribute can be used.

While using regression or decision tree algorithms the missing value can be replaced by the most probable value.

Noisy:

Noisy generally means random error or containing unnecessary data points. Here are some of the methods to handle noisy data.

Binning:

This method is to smooth or handle noisy data. First, the data is sorted then values are separated and stored in the form of bins. There are three methods for smoothing data in the bin.

Smoothing by bin mean method: In this method, the values in the bin are replaced by the mean value of the bin;

Smoothing by bin median: In this method, the values in the bin are replaced by the median value;

Smoothing by bin boundary: In this method, the using minimum and maximum values of the bin values are taken and the values are replaced by the closest boundary value.

Regression:

This is used to smooth the data and will help to handle data when unnecessary data is present. For the analysis, purpose regression helps to decide the variable which is suitable for our analysis.

Clustering:

This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

Data integration:

The process of combining multiple sources into a single dataset. The Data integration process is one of the main components in data management. There are some problems to be considered during data integration.

Schema integration:

Integrates metadata(a set of data that describes other data) from different sources.

Entity identification problem:

Identifying entities from multiple databases. For example, the system or the use should know student _id of one database and student_name of another database belongs to the same entity.

Detecting and resolving data value concepts:

The data taken from different databases while merging may differ. Like the attribute values from one database may differ from another database. For example, the date format may differ like "MM/DD/YYYY" or "DD/MM/YYYY".

Data reduction:

This process helps in the reduction of the volume of the data which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space. There are some of the techniques in data reduction are Dimensionality reduction, Numerosity reduction, Data compression.

Dimensionality reduction:

This process is necessary for real-world applications as the data size is big. In this process, the reduction of random variables or attributes is done so that the dimensionality of the data set can be reduced. Combining and merging the attributes of the data without losing its original characteristics. This also helps in the reduction of storage space and computation time is reduced. When the data is highly dimensional the problem called "Curse of Dimensionality" occurs.

Numerosity Reduction:

In this method, the representation of the data is made smaller by reducing the volume. There will not be any loss of data in this reduction.

Data compression:

The compressed form of data is called data compression. This compression can be lossless or lossy. When there is no loss of information during compression it is called lossless compression. Whereas lossy compression reduces information but it removes only the unnecessary information.

Data Transformation:

The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements. There are some methods in data transformation.

Smoothing: With the help of algorithms, we can remove noise from the dataset and helps in knowing the important features of the dataset. By smoothing we can find even a simple change that helps in prediction.

Aggregation: In this method, the data is stored and presented in the form of a summary. The data set which is from multiple sources is integrated into with data analysis description. This is an important step since the accuracy of the data depends on the quantity and quality of the data. When the quality and the quantity of the data are good the results are more relevant.

Discretization: The continuous data here is split into intervals. Discretization reduces the data size. For example, rather than specifying the class time, we can set an interval like (3 pm-5 pm, 6 pm-8 pm).

Normalization: It is the method of scaling the data so that it can be represented in a smaller range. Example ranging from -1.0 to 1.0.

Some of the Data Preprocessing tasks in sentiment analysis are:

- Checking for Duplicates
- Checking for Null Values
- Making text lowercase

- Removing hyperlinks
- Removing punctuation marks
- Eliminating Stop Words

3.1.3. Data Resampling

Due to the imbalance of our dataset, we have tried data resampling in some of our experiments. Data resampling is a popular way of dealing with imbalanced data. In this project, we tried to oversample the data of class 1,2 and 3 by repeatedly sampling those reviews because these three classes have far less samples than the other two. Hence we use SMOTE technique to oversample our data by Synthetic approach.

SMOTE (Synthetic Minority Oversampling Technique)

It is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It increases the number of cases in your dataset in a balanced way.

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

3.1.4. Word Vectorization

TF-IDF Vectorizer

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.It transforms text to feature vectors that can be used as input to estimator.

It assigns a value to a term according to its importance in a document scaled by its importance across all documents in your corpus, which mathematically eliminates naturally occurring words in the English language, and selects words that are more descriptive of your text.

3.1.5. Classication Model

Logistic Regression

Logistic regression is a classification algorithm that uses the weighted combination of the input features and passes them through a sigmoid function. Here, we use One-vs-all is a strategy for multi class classification that involves training N distinct binary classifiers, each designed to recognize a specific class.In this we consider one class as 1 and rest all as 0, we train the model and get the requisite weights. We store the value of weights in a dictionary format for each classifiers. Then by the help of Sigmoid Function we calculate the probability. Highest probability takes a presidency and we classify that data to corresponding classifier.

The probability distribution that defines multi-class probabilities is called a multinomial probability distribution. A logistic regression model that is adapted to learn and predict a multinomial probability distribution is referred to as Multinomial Logistic Regression.

3.2 System Analysis

System analysis is the term used to describe the process of collecting and analyzing facts in respect of existing operation of the solution of the situation prevailing so that an effective computerized system may be designed and implemented of proved feasible. It also diagnosis the problems and using that information recommends improvement to the system. System analysis is the reduction of the entire system by studying the various operations performed and the relationship with the system and requirement of its successor. A system can be defined as an orderly grouping of independent component linked together according to a plan to achieve a specific objective. System analysis may be considered as an interface between the actual problem and computer. Before a computer can perform, it is necessary to investigations are called system analyst. System analysis also embraces system design which is an activity concerned with the design of a computerized application based on the facts disclosed during the analysis stage. The same person who knows as the system analyst carries out both activities. In feasibility study in most cases project is being driven by a problem in the business.

3.2.1. Feasibility Study :

A feasibility study is an evaluation of a proposal designed to determine the difficulty in carrying out a designated task. Generally, a feasibility study precedes technical development and project implementation. In other words, a feasibility study is an evaluation or analysis of the potential impact of a proposed project. Feasibility Study is performed to choose the system that meets the performance requirements at least cost. The most essential tasks performed by a Feasibility Study are the identification and description of candidate systems, the evaluation of the candidate systems and the selection of the best of the candidate systems. The best system means the system that meet performance requirements at the least cost. The most difficult part of a Feasibility Study is the identification of the candidate systems and the evaluation of their performances and costs. The new system has no additional expense to implement the system. The new system has advantages such as we can easily access files from any client in the Network, accurate output for accurate input and this application is more user friendly. We can use this application not only in this organization but also in other firms. So it is worth solving the problem.

3.2.2. Technical Feasibility :

Technical Feasibility study is performed to check whether the proposed system is technically feasible or not. Technical feasibility centers on the existing computer system (hardware, software, etc.) and to what extent it can support the proposed addition. This involves financial consideration to accommodate technical enhancement. This system is technically feasible. All the data are stored in files. The input can be done through dialog boxes which are both interactive and user friendly.

3.2.3 Economical Feasibility :

Economic Feasibility Study is the most frequently used method for evaluating the effectiveness of a candidate system. More commonly known as cost/benefit analysis, the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with cost. This analysis phase determines how much cost is needed to produce the proposed system. As the organization has required machines and supporting programs for the application to execute itself.

3.2.4 Operational Feasibility :

Operational feasibility is a measure of how well a proposed system solves the problems, and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development. The operational feasibility assessment focuses on the degree to which the proposed development projects fits in with the existing business environment and objectives with regard to development schedule, delivery date, corporate culture, and existing business processes. To ensure success, desired operational outcomes must be imparted during design and development. These include such design-dependent parameters such as reliability, maintainability, supportability, usability, reducibility, disposability, sustainability, affordability and others. These parameters are required to be considered at the early stages of design if desired operational behaviors are to be realized. A system design and development requires appropriate and timely application of engineering and management efforts to meet the previously mentioned parameters. A system may serve its intended purpose most effectively when its technical and operating characteristics are engineered into the design. Therefore, operational feasibility is a critical aspect of systems engineering that needs to be an integral part of the early design phases.

3.3. Software Requirement Specifications

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation. The appropriation of requirements and implementation constraints gives the general overview of the project in regards to what the areas of strength and deficit are and how to tackle them.

- Python IDE 3.7 version 14
- Jupyter notebook
- Dataset : Sentiment analysis of Amazon reviews (Kaggle)

Hardware Requirements

Minimum hardware requirements are very dependent on the particular software being developed by a given a thought Python / PyCharm / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

- Operating system : windows
- Processor : minimum intel i3
- Ram : minimum 4 gb
- Hard disk : minimum 250gb

3.3.1. Purpose

The purpose of this document is to investigate a small part of this large problem: positive,neutral and negative attitudes towards products. Sentiment analysis attempts to determine which features of text are indicative of it's context and build systems to take advantage of these features. The problem of classifying text as positive,neutral or negative is not the whole problem in and of itself, but it offers a simple enough premise to build upon further.

3.3.2. Scope

Finding opinion sources and monitoring them on the Web can still be a difficult task because there are a large number of diverse sources, and every source may also have a big volume of opinionated text (text with opinions or sentiments). Selecting an attributes for sentiment classification using feature relation networks. In many cases, opinions are hidden in long forum posts and blogs. It is complex for a human reader to find

relational sources, extract relational sentences with opinions, read them, understand them, and organize them into usable forms. Thus, automated summarization syst
ems are needed. Using this summarization we can recognize the importance, quality, popularity of product and services. In this system we make summarization for movie. But, we can use this system anywhere, where text analysis is required. Sentiment analysis, also known as opinion mining, grows out of this need. It is a challenging natural language processing or text mining problem. Due to its tremendous value for practical applications, there has been an explosive growth of both research in academia and applications in the industry.

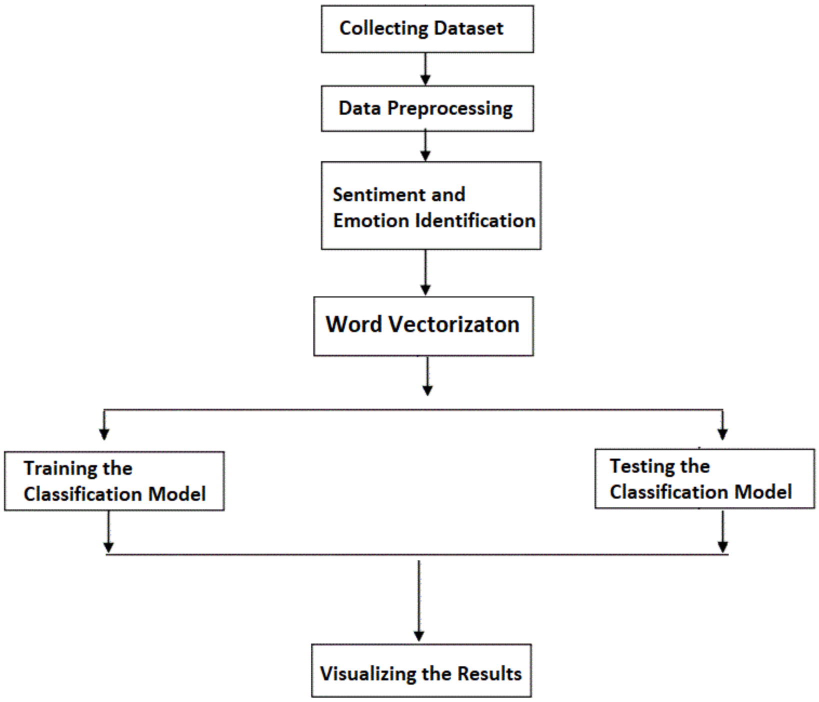
3.3.3 Overall Description

Customers rate a product depending on the level of satisfaction they have with it. Ideally, products are rated on a scale of 1-5. With 1 being the lowest rating and 5 being the highest. Depending on the rating, users leave a review of the product. There are eCommerce stores like Amazon, eBay, Overstock, Zappos, and others that display ratings of individual products. Then, there are app stores like Google Play and Apple App store that display the ratings of apps along with user comments. The ratings of a product are reflected in the comments. In the advanced sentiment analysis for the product rating system, comments are analyzed to detect the hidden sentiments.

Sentiment analysis using machine learning takes the help of a database comprising sentiment-based words that include both positive and negative keywords. The words used in the user comments section is compared to the words contained in the database and an evaluation is made. By comparing with the keywords in the database, the system specifies whether the product is good, bad, or worst. However, sentiment analysis uses computational linguistics that goes beyond the mere detection of words in a sentence. It matches sentiments to entities and also understands sarcasm to accurately recognize the emotional tone behind a sentence.

4. Design

4.1. Architecture


Collecting Dataset
Data Preprocessing
Sentiment and Emotion Identification
Word Vectorizaton
Training the Classification Model
Testing the Classification Model
Visualizing the Results

For the purpose of Sentiment and emotion identification, we make use of TextBlob which is a python library for Natural Language Processing (NLP).

TextBlob is actively used Natural Language ToolKit (NLTK) that supports complex analysis and operations on textual data. It returns Polarity and Subjectivity sentence.

Word Vectorization

We make use of TF-IDF (Term Frequency Inverse Document Frequency) Vectorizer which is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.It transforms text to feature vectors that can be used as input to estimator.

Training and Testing the Classification model -

These are the parallel activities in which our dataset is split and categorized as train and test data.Training data is used to train an algorithm or machine learning model to predict the outcome you design your model to predict while Test data is used to measure the performance, such as accuracy or efficiency, of the algorithm you are using to train the machine.

Visualizing the results -

Final Results are visualized to:

- Look at evaluation metrics.
- Look at performance charts like ROC, Lift Curve, Confusion Matrix, and others.
- Look at learning curves to estimate overfitting.
- Look at model predictions on best/worst cases.
- Look how resource-intensive is model training and inference (they translate to serious costs and will be crucial to the business side of things)

4.2. UML DIAGRAMS

Introduction

The Unified Modeling Language (UML) is a standard language for specifying, visualizing,constructing, and documenting the artifacts of software systems, as well as for business modeling and other non-software systems. The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

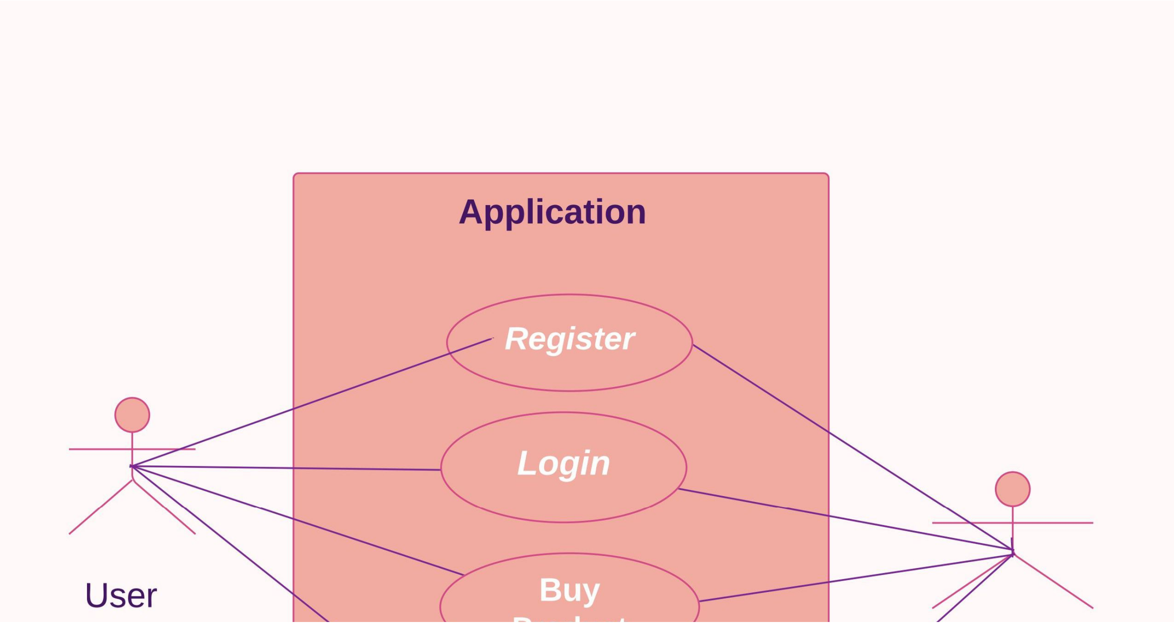
Use case Diagram

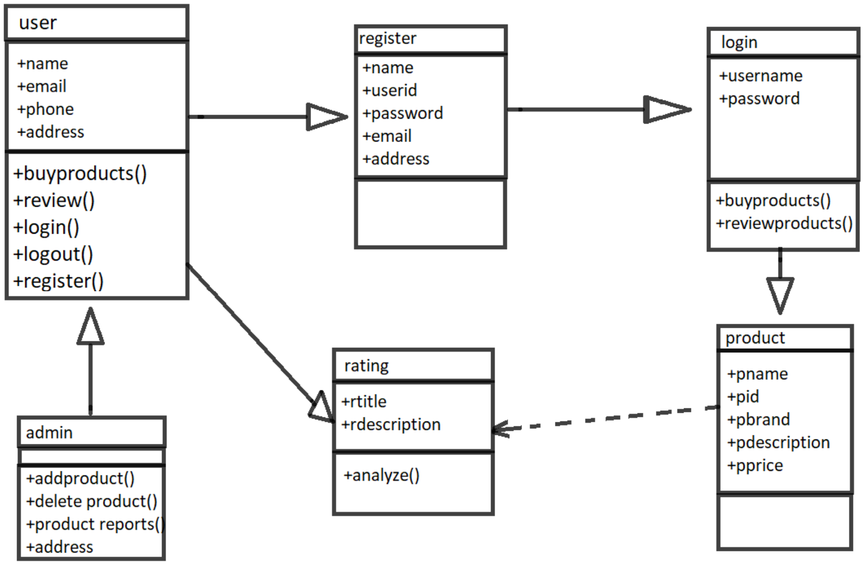
Use case Diagram shows a set of uses cases and actors with their relationships. It addresses the static use case view of a system. Use case diagrams gives a graphic overview of the actors involved in a system, different functions needed by those actors and how these different functions are interacted. These are important in organizing and modeling the behavior of a system. A more detailed description might characterize a use case as:

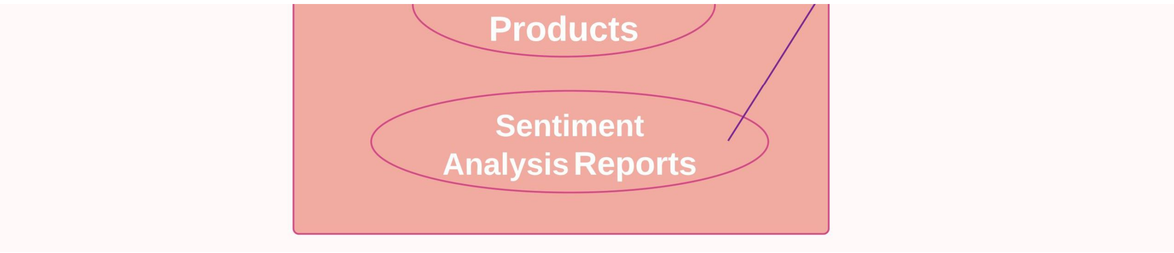
1.A pattern of behavior the system exhibits

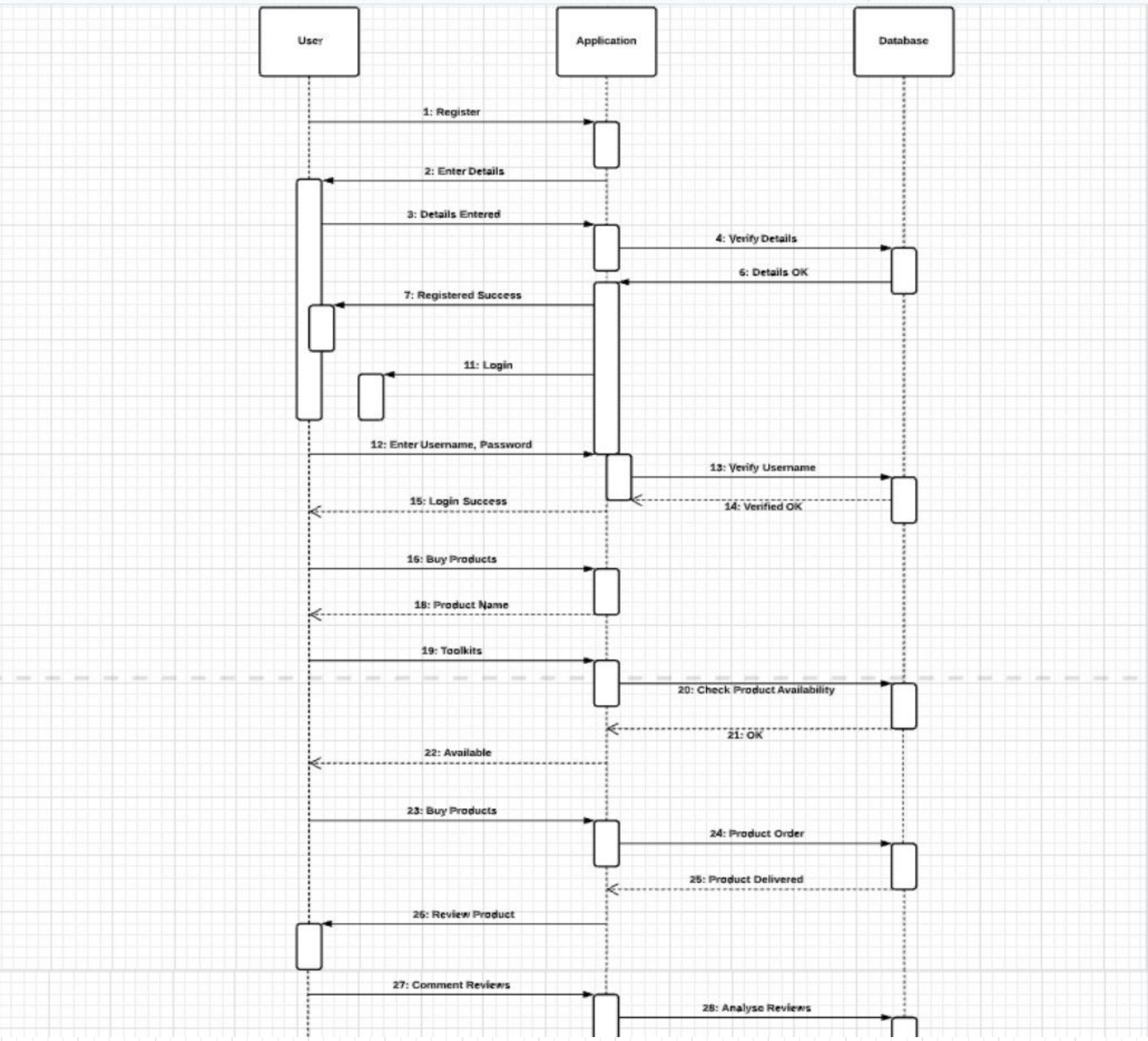
2.A sequence of related transactions performed by an actor and the system

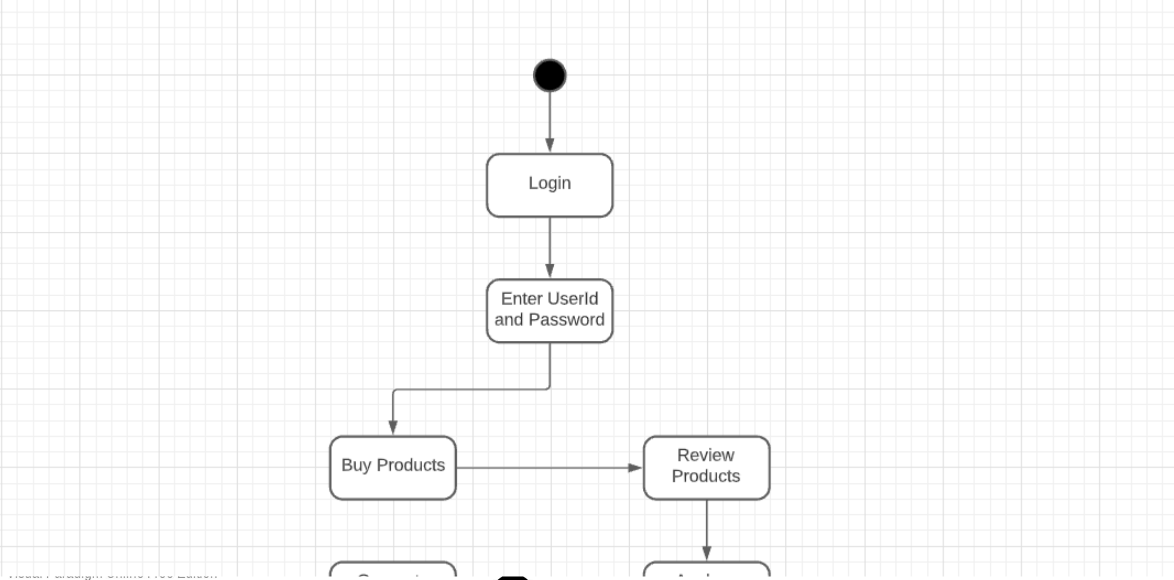
3.Delivering something of value to the actor.

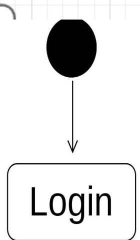







User
Application
Database
1: Register
2: Enter Details
3: Details Entered
4: Verify Details
6: Details OK
7: Registered Success
11: Login
12: Enter Username, Password
13: Verify Username
14: Verified OK
15: Login Success
16: Buy Products
18: Product Name
19: Toolkits
20: Check Product Availability
21: OK
22: Available
23: Buy Products
24: Product Order
25: Product Delivered
26: Review Product
27: Comment Reviews
28: Analyse Reviews


Login
Enter UserId and Password
Buy Products
Review Products


Login

5. Implementation

5.1. Introduction to Technologies used

5.1.1 Python

Python is currently the most widely used multi-purpose, high-level programming language.Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber.. etc.

5.1.2 Jupyter Notebook

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython 28 kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use

5.1.3 Numpy

NumPy (short for Numerical Python) is a commonly used Python data analysis package. By using NumPy, we can speed up our workflow, and interface with other packages in the Python ecosystem, like scikit-learn, that use NumPy under the hood.

NumPy provides an efficient interface to store and operate on dense data buffers. In some ways, NumPy arrays are like Python's built-in list type, but NumPy arrays provide much more efficient storage and data operations as the arrays grow larger in size.

5.1.3 Pandas

Pandas provide extended data structures to hold different types of labeled and relational data. This makes python highly flexible and extremely useful for data cleaning and manipulation. Pandas provide functions for performing operations like merging, reshaping, joining, and concatenating data.

Pandas generally provide two data structures for manipulating data, They are:

1. Series : Pandas Series is a one-dimensional labelled array capable of holding data of any type (integer, string, float, python objects, etc.).

2. Data Frame : Pandas DataFrame is a two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns).

5.1.4 SCIKIT LEARN

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

5.1.5 NLTK

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language.

NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems.

5.1.6 Seaborn

Seaborn is a visualization library for statistical graphics plotting in Python. It provides default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.

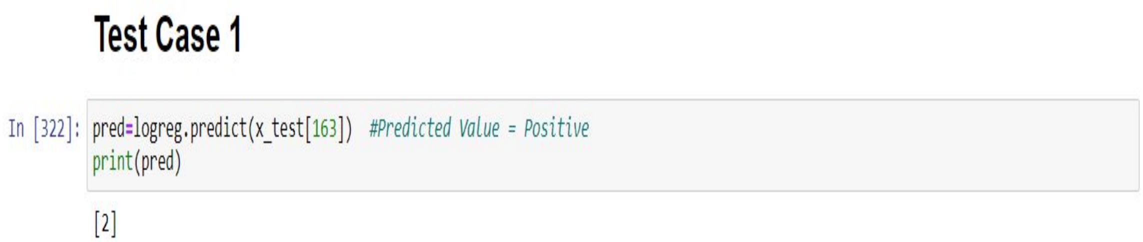
Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

5.2 Sample Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import nltk
import re
import string
import sklearn
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.feature_extraction.text import TfidfVectorizer
from imblearn.over_sampling import SMOTE
from textblob import TextBlob

df=pd.read_csv('Musical_instruments_reviews.csv')
df.head()
df.isnull().sum()
df['reviewText']=df['reviewText'].fillna('Missing')


def f(row):
    if row['overall'] == 3.0:
        val = 'Neutral'
    elif row['overall'] == 1.0 or row['overall'] == 2.0:
        val = 'Negative'
    elif row['overall'] == 4.0 or row['overall'] == 5.0:
        val = 'Positive'
    else:
        val = -1
    return val
df['sentiment'] = df.apply(f, axis=1)
df.head()

def review_cleaning(text):
    text = str(text).lower()
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    return text


df['reviews']=df['reviews'].apply(review_cleaning)
df.head()


df['polarity'] = df['reviews'].map(lambda text: TextBlob(text).sentiment.polarity)
df['sub'] = df['reviews'].map(lambda text: TextBlob(text).sentiment.subjectivity)
df['review_len'] = df['reviews'].astype(str).apply(len)
df['word_count'] = df['reviews'].apply(lambda x: len(str(x).split()))
df.head()

label_encoder = preprocessing.LabelEncoder()
df['sentiment']= label_encoder.fit_transform(df['sentiment'])
df['sentiment'].unique()
tfidf_vectorizer = TfidfVectorizer(max_features=5000,ngram_range=(2,2))
x= tfidf_vectorizer.fit_transform(df['reviews'])
y=df['sentiment']
print(f'Original dataset shape : {Counter(y)}')
smote = SMOTE(random_state=42)
X_res, y_res = smote.fit_resample(x, y)
print(f'Resampled dataset shape {Counter(y_res)}')

X_train, X_test, y_train, y_test = train_test_split(X_res, y_res, test_size=0.25,
random_state=0)
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression(C=10000.0, random_state=0)
logreg.fit(X_train, y_train)
```

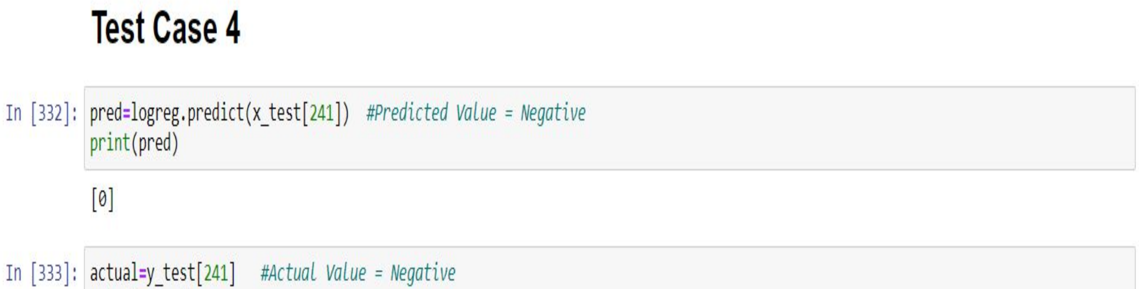
```
y_pred = logreg.predict(X_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_test,
y_test)))
```

6. Test Cases



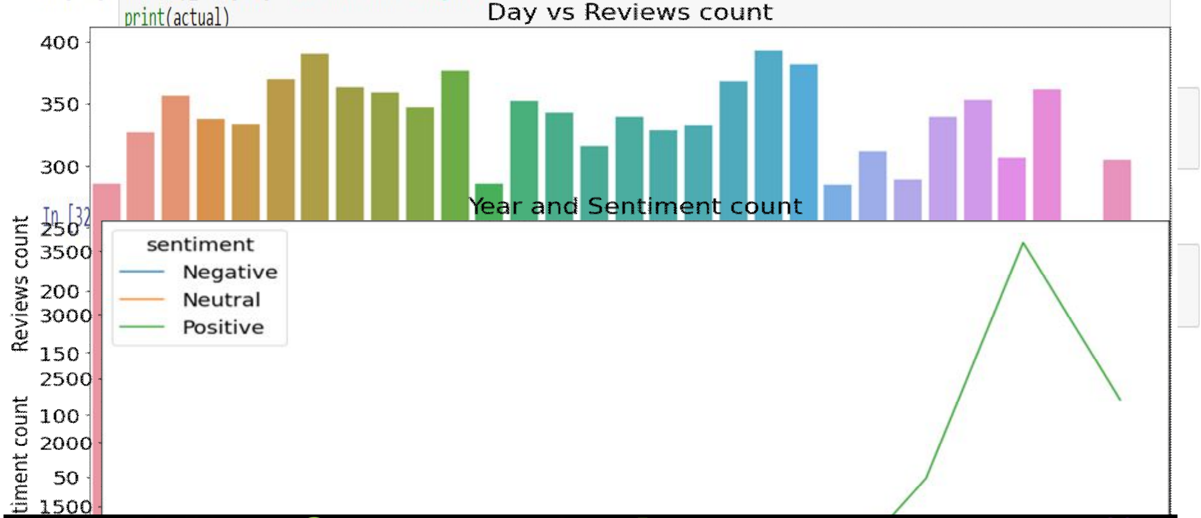
Test Case 1

```
In [322]: pred=logreg.predict(x_test[163])  #Predicted Value = Positive
          print(pred)

          [2]
```

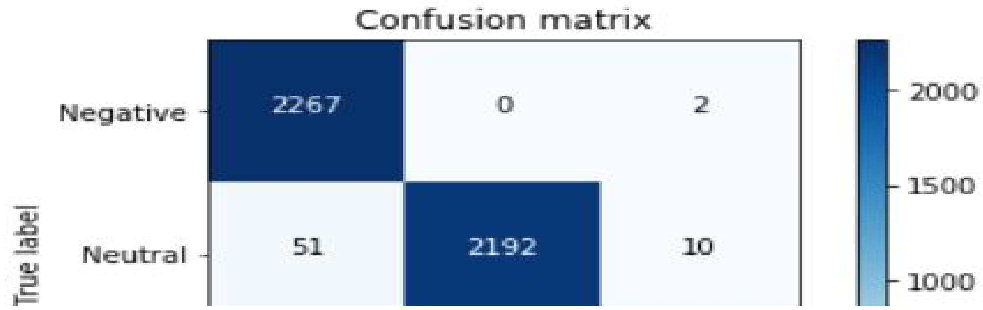


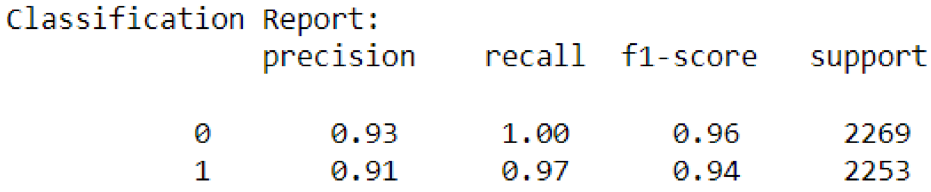
Test Case 4

```
In [332]: pred=logreg.predict(x_test[241])  #Predicted Value = Negative
          print(pred)

          [0]

In [333]: actual=y_test[241]   #Actual Value = Negative
          print(actual)
```

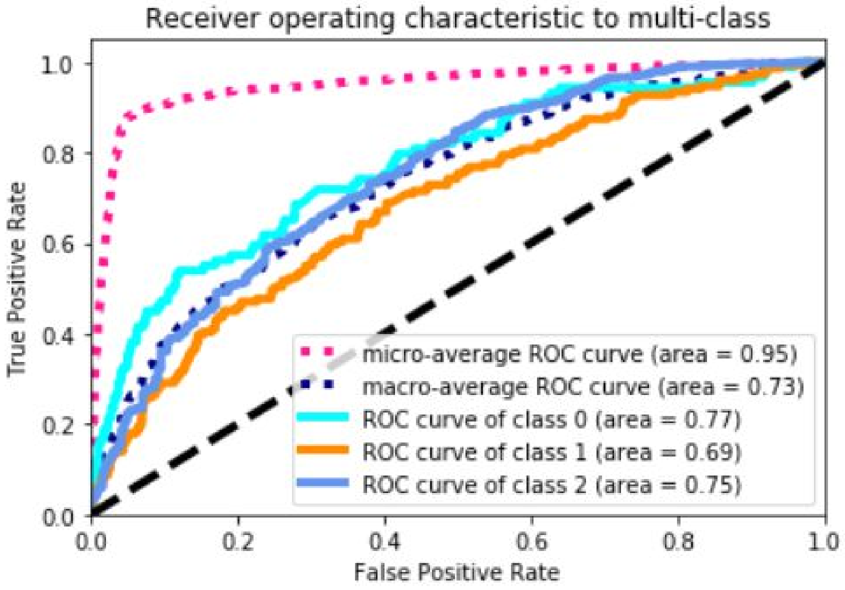


Confusion Matrix





```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.96      2269
           1       0.91      0.97      0.94      2253
```



Leveraging on machine learning and NLP, organizations can interact with their customers both rationally and emotionally, improve their customer experience, and provide tailored assistance.

In summary, We have done a pretty neat job on classifying all the classes starting from splitting the sentiments based on overall score,text cleaning, customize the stop words list based on requirement and finally handling imbalance with smote.

Here are few insights from the research:

- We considered ngram in sentiment analysis as one word can't give proper results and stop words got to be manually checked as they have negative words. It is advised to avoid using stop words in sentiment analysis.

- Most of our neutral reviews were actual critic of product from the buyers, so amazon can consider these as feedback and give them to the seller to help them improve their products.

Balancing the dataset gave us better accuracy score. Without balancing, we got good precision but very bad recall and in-turn it affected f1 score. So balancing the target feature is important.

9.Future Enhancement

Recent studies indicates that the number of people and companies using social media applications as a customer relationship management tool has dramatically increased (Bagheri et al., 2013; Fuchs et al., 2014; Kaplan and Haenlein, 2010). It is the norm to see a large number of reviews, complaints and compliments posted and shared just seconds after a new product is released. Analysing this information helps companies to accommodate this growing trend in order to achieve some business values like increasing the number of customers; enhancing customer loyalty, customer satisfaction and company reputation; and achieving higher sales and total revenue (Batrinca and Treleaven, 2014; Bravo-Marquez et al., 2014; He et al., 2015). On the other hand, this information can be used by the customers as testimonials by extracting the strengths and weaknesses of the distinguishable features of each product, as well as finding the satisfaction levels of other users of those products. Besides the benefits in entrepreneurship, an analysis of political pages provides information to political parties regarding people's view of their programmes. Social organisations may seek people's opinion on current debates or on matters like the next presidential candidate.

This information can be obtained by analysing the sentiment orientation of comments, the number of likes, shares or comments on posted topics. Applications of SA range from public voice analysis, crowd surveillance, customer care and social intelligence-based SA to exploit the publics' online content generation for analysing inputs such as pandemic spreading, emotion and responses towards local events. SA that focuses on microblogging is very typical because this is the main source that taps the public's voice. SA on microblogging data is more challenging compared to conventional texts such as documents review, due to the length, repeated use of some unofficial and atypical words and the rapid progress of language variation usage.

10.Bibliography

[1] C. Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. Swarthmore College, 2013.

[2] OnamBharti, Mrs, and Monika Malhotra."SENTIMENT ANALYSIS."

[3] Sangani, Chirag, and SundaramAnanthanarayanan. "Sentiment analysis of app store reviews."

[4] Callen Rain,"Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning" Swarthmore College, Department of ComputerScience

[5] K.Yessenov and S. Misailovic. Sentiment analysis of movie review comments. Methodology, pages 1–17, 2009.

[6] https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/

[7] https://towardsdatascience.com/document-feature-extraction-and-classification-53f0e813d2d3

[8] M. S. Elli and Y.-F. Wang. Amazon reviews, business analytics with sentiment analysis.

[9] https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

[10] Sentiment analysis using product review by data Xing Fang and Justin Zhan