Twitter Sentiment Analysis

Group Members :

Minahil Sadiq	SP20-BCS-023
Sheeza Ali	SP20-BCS-005
Ayesha Bibi	SP20-BCS-152
Muhammad Zubair	SP20-BCS-025

Group Leader :

Minahil Sadiq	SP20-BCS-023

Due Date: 12/09/2022

Course title: Machine Learning

Instructor: Dr. ALLAH BUX SARGANA

ABSTRACT

This project is about analyzing the sentiments of people regarding different topics. Sentiment analysis is defined as the process of computationally identifying and categorizing opinions expressed in a piece of text. Sentiment Analysis is important because it helps businesses to understand the emotion of their customers. It is useful in marketing, business because the fame of any product depends upon the reaction of people, what are the sentiments of people about certain product.

Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of group of words. It involves the use of data mining, machine learning (ML) and artificial intelligence (AI) to mine text for sentiment and subjective information.

In this era of social media, twitter is a major source of consumer insight. Everyday almost 500 million tweets are being produced related to different topics. In our project we are doing sentiment analysis of tweets taken by twitter. We did not get the tweets directly from twitter for our project, we took the data from an online website.

The central idea is to take the textual data and make it in the form, which is understandable by our Machine learning classifier, and then train our model to predict the sentiments of tweets as ‘Positive’, ‘Negative’, ‘Neutral’. As our text is in a natural language therefore, we have done Natural Language Processing on our tweets, It is a important part to do because the language which is understandable only by humans, now have to get analyzed by a learning classifier of machine learning, and our purpose was to get the maximum perfection therefore it was important part done to achieve maximum performance we can.

The **training regime** used in the project is ‘Batch Method’ in which all training instances are used at once to compute the hypothesis. And the **cross-validation technique** used is ‘Holdout Cross Validation’.

Our motivation for this problem was to learn deeply how unstructured data can be used and manipulate in the useful information, and we explore different learning algorithms of machine learning and pick the one that best suits for this problem of analyzing sentiments.

METHODOLOGY

Here are the detailed steps of methods which have been taken to achieve the goals:

1. GATHERING RELEVANT DATA:

The data we have for project was taken non-dynamically using existing data by website.

2. IMPORTING LIBRARIES:

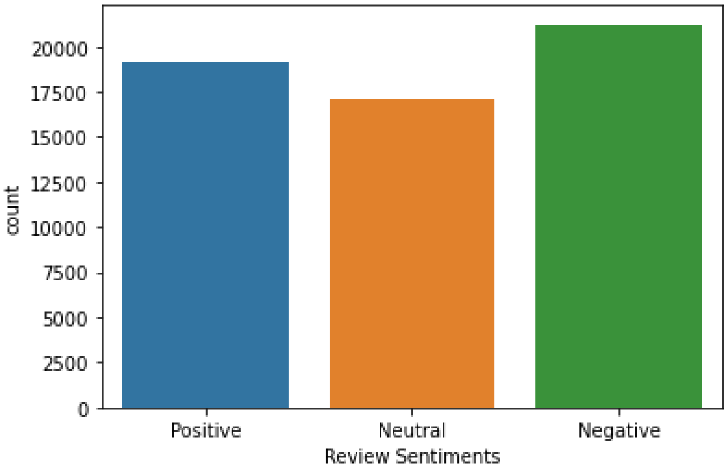
We import all the required libraries such as **pandas, numpy, seaborn, matplotlib, re, nltk, CountVectorizer, word_tokenize, PorterStemmer, WordNetLemmatizer, train_test_split, Classification_report, confusion_matrix**

3. CLEANING DATASET:

As the data obtained from social media, there was a lot of noise, we tried to remove the maximum noise we can and for this purpose we have remove null, duplicate and irrelevant data from our dataset.

4. VISUALIZING DATA:

To make the data easier to understand through pictorial representation we use to do Visualization.



The above plot represents the number of instances we have in each category of ‘sentiment’ after cleaning dataset.

5. REMOVING NOISE FROM TWEETS:

For removing noise from tweets, we perform certain tasks which are:

- Converting text into lower case
- Removing links

- Removing Punctuations
- Removing Stop-words
- Applying Stemming
- Applying Lemmatizing

6. VECTORIZATION AND TOKENIZATION:

In our project, we have used **CountVectorizer** which is defined as breaking down a sentence or any text into words by performing preprocessing tasks.

7. SPLITTING DATA:

We split the data with respect to the ratio of 80:20.

Python code:

```
#splitting data in train and test
X_train, X_test, y_train, y_test = train_test_split(x,y, random_state=1, test_size= 0.2)
```

8. TRAINING MODELS:

In this sentiment analysis, we have used four classifiers which are Naïve Bayes, Random Forest, Support Vector Machine (SVM) and Bayesian Logistic Regression.

1) SVM:

The "Support Vector Machine" (SVM) is a supervised machine learning technique that can solve classification and regression problems. It is, however, mostly employed to solve categorization difficulties. In our project for training purpose, we have used four SVM kernels:

- Linear Kernel
- Polynomial Kernel
- RBF kernel
- Sigmoid Kernel

Polynomial Kernel gives us the lowest accuracy which is 37%.

Radius Basis Function Kernel gives us the best accuracy among all four kernel of SVM which is 88.7%.

2) Naïve Bayes:

Naive Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks.

Bayes Theorem

Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities. Conditional probability is a measure of the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred.

3) Random Forest:

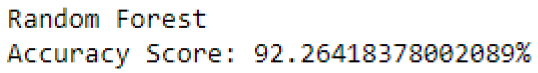
The random forest classifier is used to solve regression or classification problems. It is made up of a collection of decision trees, and each tree in the ensemble is comprised of data sample drawn from training set with replacement, called the bootstrap sample.

4) Bayesian Logistic Regression:

Bayesian logistic regression is the Bayesian counterpart to a common tool in machine learning, logistic regression. The goal of logistic regression is to predict a one or a zero for a given training item. An example might be predicting whether someone is sick or ill given their symptoms and personal information.

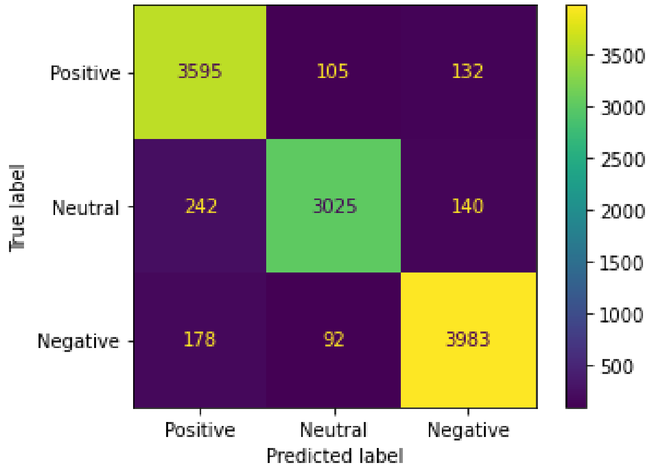
9. BEST FIT CLASSIFIER:

After the training and testing the data and checking the accuracy of the model. We have chosen "Random Forest" because it gives us better results as compared to other classifiers.



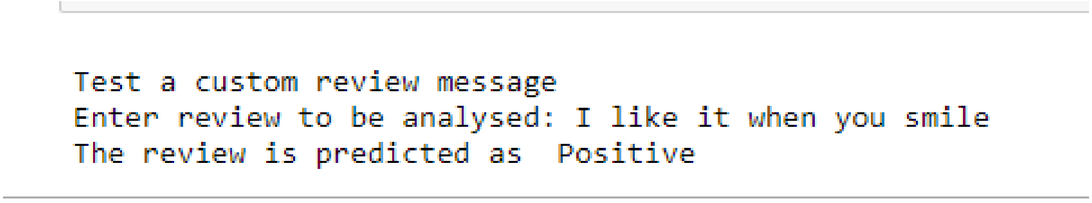
```
Random Forest
Accuracy Score: 92.26418378002089%
```

Confusion matrix:

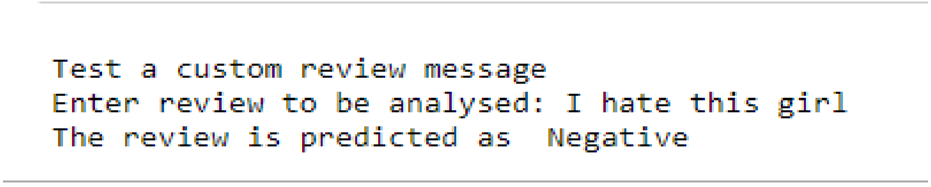


RESULTS

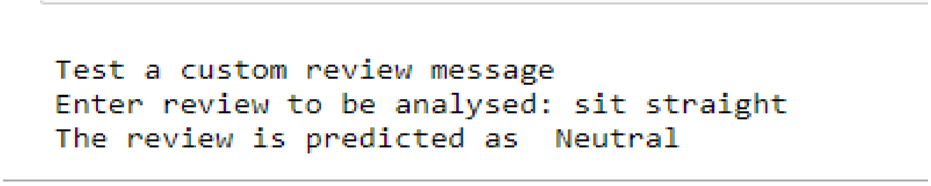
Results after testing our model on unseen data:



```
Test a custom review message
Enter review to be analysed: I like it when you smile
The review is predicted as  Positive
```



```
Test a custom review message
Enter review to be analysed: I hate this girl
The review is predicted as  Negative
```

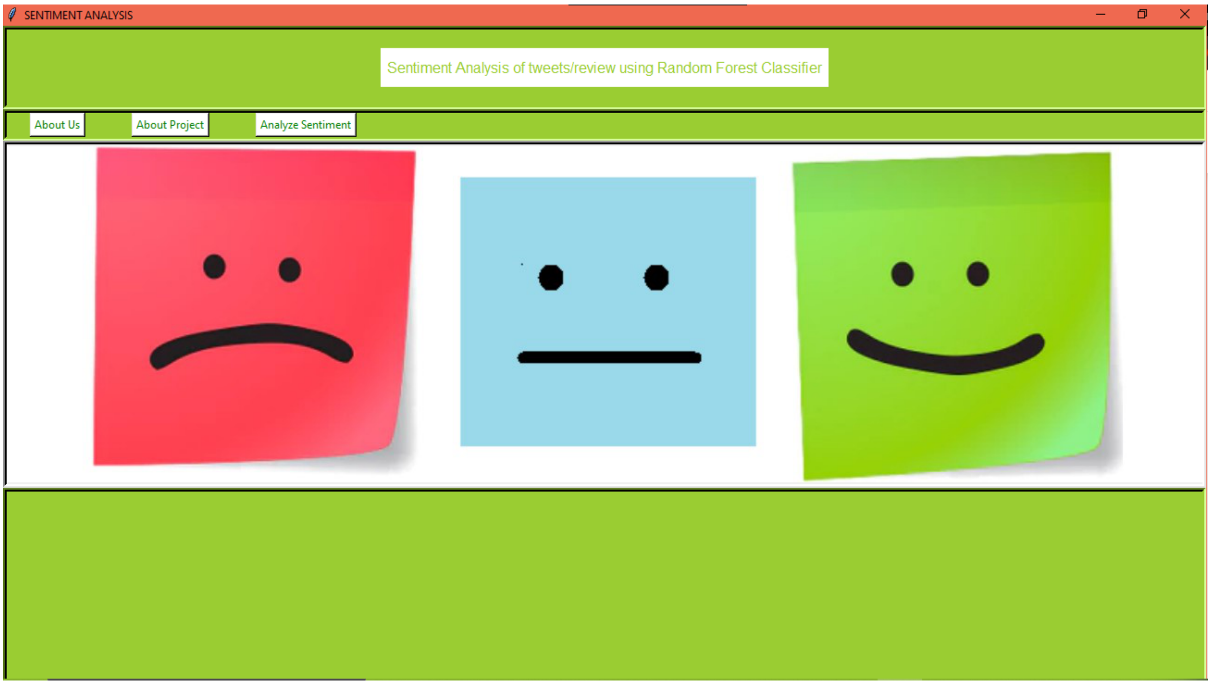


```
Test a custom review message
Enter review to be analysed: sit straight
The review is predicted as  Neutral
```

Graphical User Interface

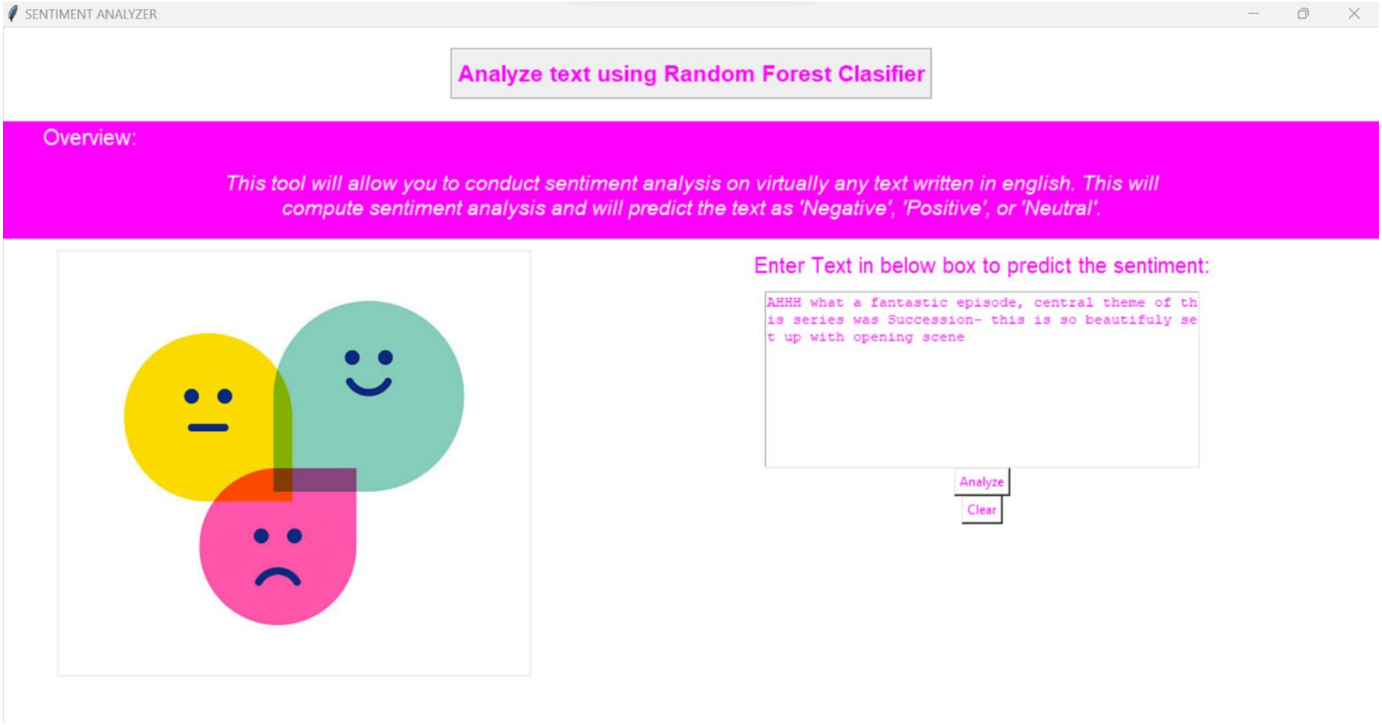
For the Graphical user interface, we have used tkinter.

<u>First Window To Appear</u>


SENTIMENT ANALYSIS
Sentiment Analysis of tweets/review using Random Forest Classifier
About Us
About Project
Analyze Sentiment

Clicking on Third Button This New Window Will Appear

Testing data is typed in the text box to analyze the results.



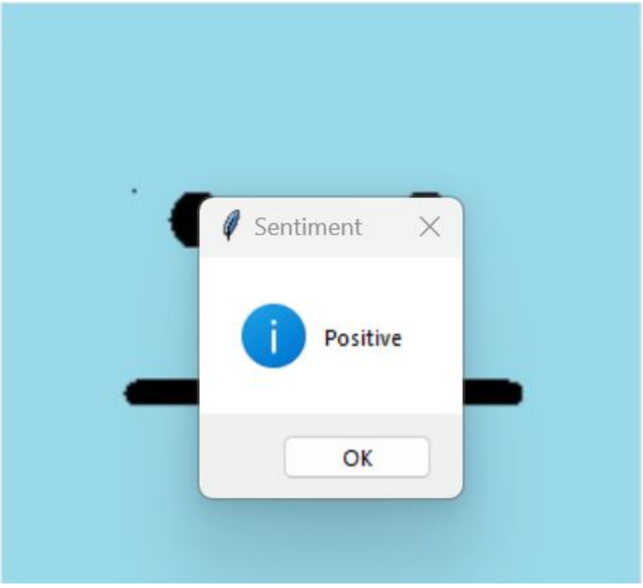
Overview:

This tool will allow you to conduct sentiment analysis on virtually any text written in english. This will compute sentiment analysis and will predict the text as 'Negative', 'Positive', or 'Neutral'.

Enter Text in below box to predict the sentiment:

```
AHHH what a fantastic episode, central theme of th
is series was Succession- this is so beautifuly se
t up with opening scene
```

Sentiment predicted:



Positive