Sentimental Analysis using twitter tweets

A project report submitted

in partial fulfilment of the requirement

for the degree in Bachelors of Computer Applications

By

Abhishek Barua

P1800108

Department of Computer Applications,

CERTIFICATE

This is to certify that the Project work titled “Sentimental Analysis using Twitter Tweets” is a bona fide work done by Abhishek Barua, P1800108 under my guidance during the final year of the course.

S. Deb
Head of the Department

Ms. Rajni Khyriem
Internal Guide

The Project presentation was held on …… at St. Edmund’s College, Shillong

Date:

External Examiner

St. Edmund’s College, Shillong.

Acknowledgement

I am immensely thankful for this wonderful opportunity through which I was able to test my skills in the subject. This project encouraged me to research further, and on the quest to acquire more knowledge, it has helped me to gain more confidence in applying everything I have learnt so far.

Firstly, I would like to express my sincere gratitude to my guide, Professor R. Khyriem, for her invaluable suggestions and ideas, great insights on the topic and enthusiastic efforts to help me in times of need. She has been greatly supportive and reliable. It is only through the helpful information she provided and her advices that I was able to complete my project successfully. I am thankful to her for sacrificing her precious time to guide me, for solving my doubts, answering my queries, for being so patient and for relentlessly helping me.

I am also grateful to the Department of Computer Applications and to all my respected lecturers who have always been there for me.

Last but not the least, I give my deepest thanks to my family and friends who have always supported and encouraged me with great enthusiasm. They are the source of my strength and without their moral support, I would not have been able to complete this project.

Table of Contents

Introduction: About the Project

Sentiment score is a scaling system that reflects the emotional depth of emotions in a piece of text.

Sentiment score detects emotions and assigns them sentiment scores, for example, from 0 up to 10 – from the most negative to most positive sentiment. Sentiment score makes it simpler to understand how customers feel.

Currently, many company are using sentimental analysis to determine about the sentiment among the user of their product in form or ratings, user review. This review helps the company to determine which product to show at the top and also on which product the company need to work more for a positive review from the user who are going to use it in the future.

This project is a machine learning application that can be used to check the sentiment of the tweets that a specific person tweets in twitter platform. This application where the user will have to enter the twitter ID of the person whose tweets sentiment they want to check. The application will then retrieve the tweets from the twitter database and tell if the tweets are positive, negative or neutral. The output then will be stored in an excel sheet that the user and user afterwards.

Synopsis

Objective:

The objective of the project is to create an application that will retrieve tweets from twitter database and check the sentiment of the tweets and save the tweets along with the sentiment of that tweet in an excel sheet.

Description:

This project will help the user to see the type of tweets that a specific person does in twitter.

The following are the outputs that are shown in this application

- It will show the percentage of negative, positive & neutral tweets that are done by a user in their first 200 tweets
- It also shows the three sentiment tweets in a separate table
- It also shows a bar graph which is plotted using the percentage calculated.

Modules used:

Pandas:

In computer programming, **pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for

Library features

- DataFrame object for data manipulation with integrated indexing.
- Tools for reading and writing data between in-memory data structures and different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of data sets.
- Label-based slicing, fancy indexing, and subsetting of large data sets.
- Data structure column insertion and deletion.
- Group by engine allowing split-apply-combine operations on data sets.
- Data set merging and joining.
- Hierarchical axis indexing to work with high-dimensional data in a lower-dimensional data structure.
- Time series-functionality: Date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging.

- Provides data filtration.

Matplot

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

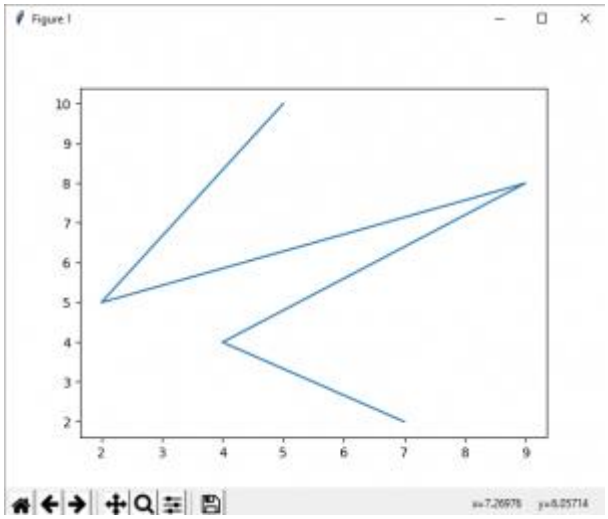
One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They're typically instruments for reasoning about quantitative information. Some of the sample plots are covered here.

Line plot :

```
# importing matplotlib module
from matplotlib import pyplot as plt

# x-axis values
x = [5, 2, 9, 4, 7]

# Y-axis values
y = [10, 5, 8, 4, 2]

# Function to plot
plt.plot(x,y)

# function to show the plot
plt.show()
```

Output :

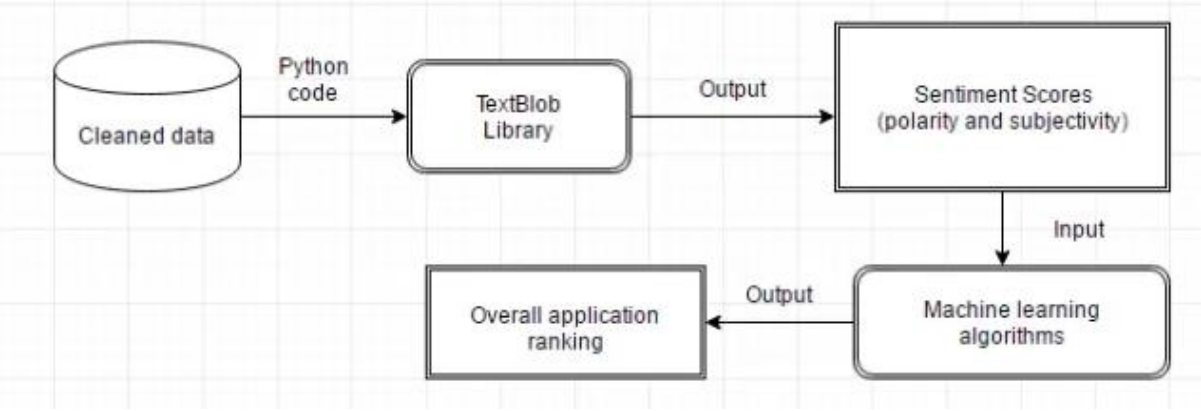


Textbolb

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Features

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Classification (Naive Bayes, Decision Tree)
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- **n**-grams
- Word inflection (pluralization and singularization) and lemmatization
- Spelling correction
- Add new models or languages through extensions
- WordNet integration



System Study

Existing systems:

Repustate

Repustate offers a sentiment analysis API that supports text in 24 languages and delivers reliable insights from social media data. Their software combines sentiment and semantic analysis to detect the emotions expressed in emojis, slang, abbreviations, and hashtags.

By creating customized sentiment analysis rules, you can adapt your model to specific industry vocabulary. Finally, you can integrate the API into your workflow through client libraries available in various programming languages.

Lexalytics

Lexalytics has been providing sentiment analysis solutions since 2003. They offer an on-premise solution (Salience) and cloud API (Semantria), which you can integrate with your workflows. You can also build fully customized models and adapt them to your business needs.

You can also create dashboards and visualizations from your data and gain deeper insights into customer emotions.

Proposed system:

Sentimental analysis using twitter tweets is an application the check the sentiment of the tweets done by a specific user. Now a days people tweets about their opinion or their thoughts in twitter which can be used by people to determine the way the person thinks.

User Requirements

User requirement:

Hardware :

Processor : intel i3 or above.

Operating system: Windows /Linux/Mac OS

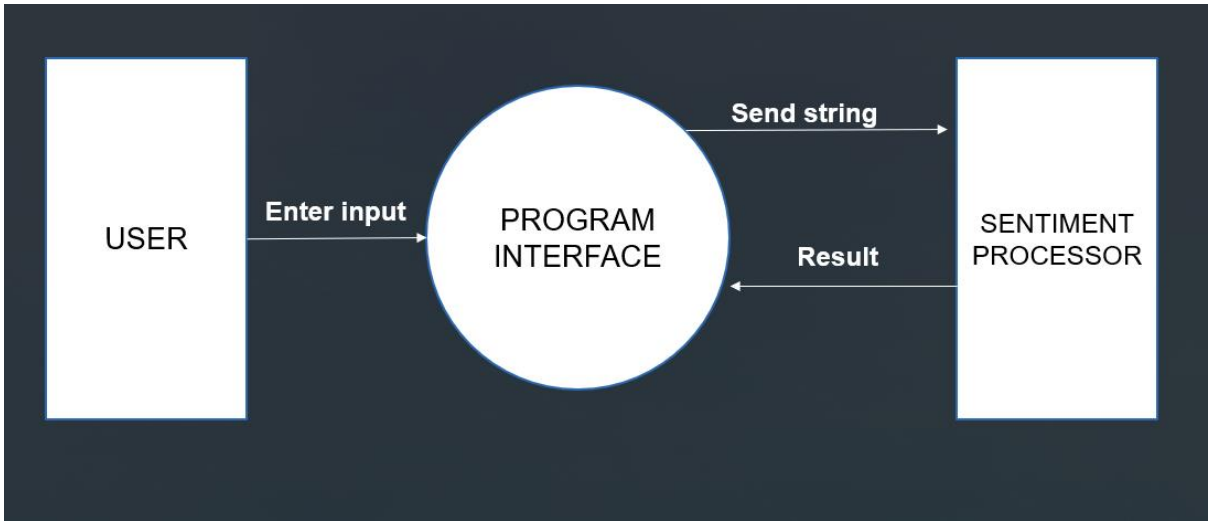
Software:

Python

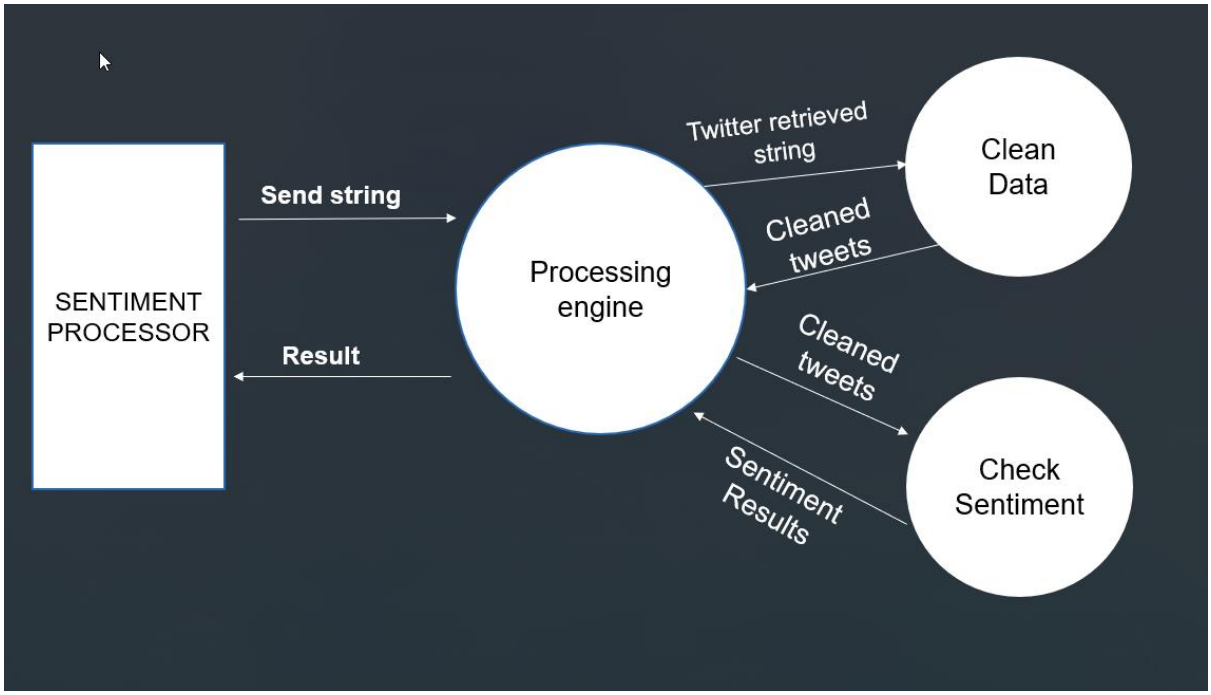
NLP (natural language processing)

Pycharm/ Jypyter

Data Flow Diagram (DFD)



Level 0



Level 1

Future Enhancements

There is enormous scope for sentimental analysis using twitter tweets in the coming future.

- The project can converted into webapp or mobile app.
- More moods can be added.
- Using google map and twitter developer account we can plot the places from where the different types of tweets are being tweeted.

Conclusion

The project is completed. The program is able to connect to the twitter server through twitter token keys , we are able to clean the tweets and also check the sentiment of each tweets and show the output. The program runs on a command line and its show the percentage of positive, negative and neutral tweets along with pie chart and the user can also view all the tweets under the three sentiments.

Bibliography

[1] R. Parikh and M. Movassate, “Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

[2] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38

[3] Neethu M,S and Rajashree R,” Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661

[4] Pak, Alexander, and Patrick Paroubek. ”Twitter as a Corpus for Sentiment Analysis and Opinion Mining.” LREC. 2010.

[5] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M..“Lexicon basedmethods for sentiment analysis”. Computational linguistics, 2011:37(2), 267-307.

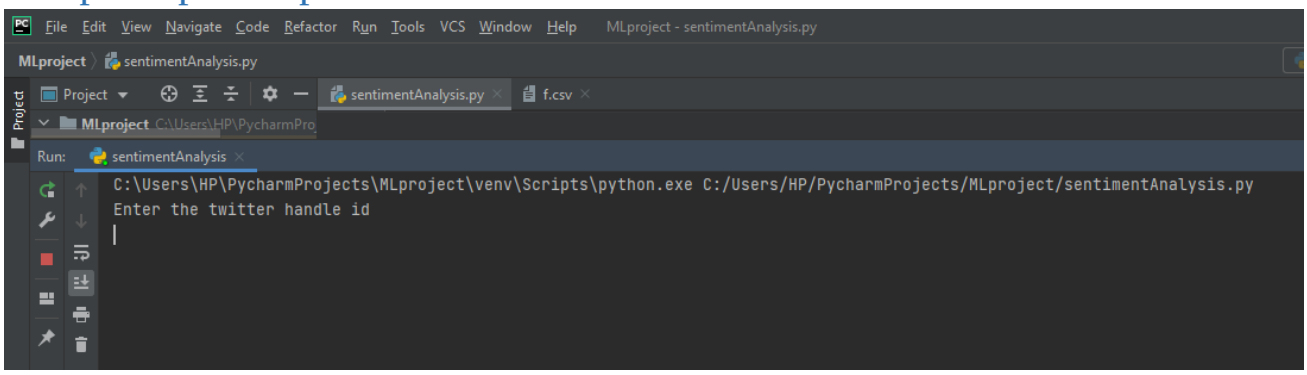
Source Code

```
import tweepy
import pandas as pd
from textblob import TextBlob
import re
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
consumer_key: str = "JvP6SALYgYN6FHKgAW7MBXeCv"
consumer_secret: str = "KJFczocadXq3IxLIotz3rLOlFcQBuP5QPsKSWXwtjzNYbkNSb4"
access_token: str = "137979383199494553 6-Eb7GYmlFJvCcE7TTBv5YCEXnDYep0Z"
access_token_secret: str = "0DQEthNO7hRB3rfNsCW894dkuEnrtGw1Y9KHID7V1uigH"
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit = True)
name = input('Enter the twitter handle id\n')
posts = api.user_timeline(screen_name=name, count=100, language = "en",
tweet_mode="extended")

df = pd.DataFrame( [tweet.full_text for tweet in posts], columns = ['tweets'])
# print(df.shape[0])
#creating a fn that clean the texts
def cleanTxt(text):
    text = re.sub(r'@[A-Za-z0-9]+' , '', text) # (sub = subtitute)it removes @ mentions and r
tells python that the text is raw
    text = re.sub(r'#', '', text) # removing the #
    text = re.sub(r'RT[\s]+', '', text) # rt for removing retweets and white spaces
    text = re.sub(r'https?:\/\/\S+', '', text) # removing the hyperlink
    return text
#cleaning the text
df['tweets']= df['tweets'].apply(cleanTxt)

#creating subjectivity
def getSubjectivity(text):
    return TextBlob(text).sentiment.subjectivity
#creating polarity
def getPolarity(text):
    return TextBlob(text).sentiment.polarity
#creating two columns for subjectivity and polarity
df['Subjectivity'] = df['tweets'].apply(getSubjectivity)
df['Polarity'] = df['tweets'].apply(getPolarity)
#showing new dataframe with new columns

#creating a fn to compute positive,negative & neutral analysis
def getAnalysis(score):
    if score < 0:
        return 'Negative'
```

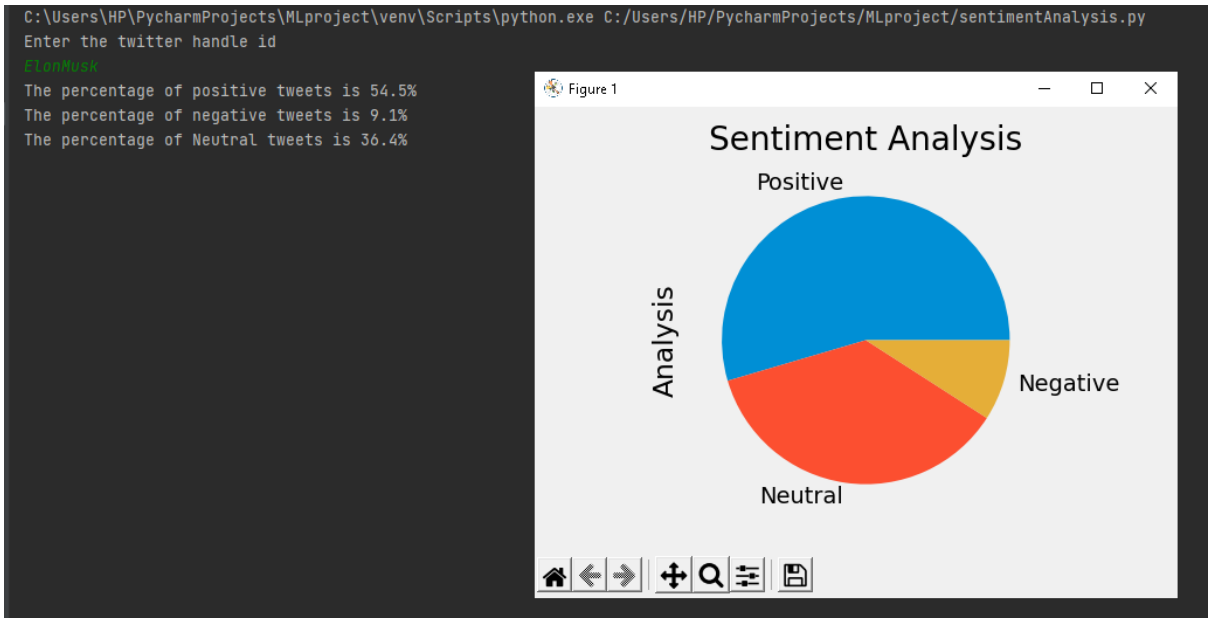
```
    elif score == 0:
        return 'Neutral'
    else:
        return 'Positive'


def display(dataframe, choice):
    if choice == 1: # Positive
        d = dataframe.loc[dataframe['Analysis'] == 'Positive']
        print('Total number of Positive tweets = ', str(d.shape[0]), '\n')
    elif choice == 2: # Negative
        d = dataframe.loc[dataframe['Analysis'] == 'Negative']
        print('Total number of Negative tweets = ', str(d.shape[0]), '\n')
    elif choice == 3: # Neutral
        d = dataframe.loc[dataframe['Analysis'] == 'Neutral']
        print('Total number of Neutral tweets = ', str(d.shape[0]), '\n')
    return d


df['Analysis'] = df['Polarity'].apply(getAnalysis)
# Printing all the positive tweets
#j=1
sortedDF = df.sort_values(by=['Polarity'])

# getting the % of positive, negative & neutral tweets
ptweets = df[df.Analysis == 'Positive']
ptweets = ptweets['tweets']
percentage = round((ptweets.shape[0] / df.shape[0]) * 100, 1)
print("The percentage of positive tweets is " + str(percentage) + '%')

ptweets = df[df.Analysis == 'Negative']
ptweets = ptweets['tweets']
percentage = round( (ptweets.shape[0] / df.shape[0]) * 100, 1)
print("The percentage of negative tweets is " + str(percentage) + '%')

ptweets = df[df.Analysis == 'Neutral']
ptweets = ptweets['tweets']
percentage = round( (ptweets.shape[0] / df.shape[0]) * 100, 1)
print("The percentage of Neutral tweets is " + str(percentage) + '%')
# showing the value counts
df['Analysis'].value_counts()
# plot and visualize the counts
plt.title('Sentiment Analysis')
df['Analysis'].value_counts().plot(kind='pie')
plt.show()
inp = 1
while inp != 4:
    inp = int(input('Enter 1 for positive, 2 for Negative, 3 for Neutral tweets and 4 for exit\n'))
```

```
92.       if inp == 4:
93.           break
94.       elif inp < 4:
95.           df1 = display(sortedDF, inp)
96.           for i, tw in enumerate(df1['tweets']):
97.               print(str(i+1) + ' - ' + tw)
98.    df.to_csv('my_new_file.csv', index = False)
99.  pd.read_csv('my_new_file.csv')
```

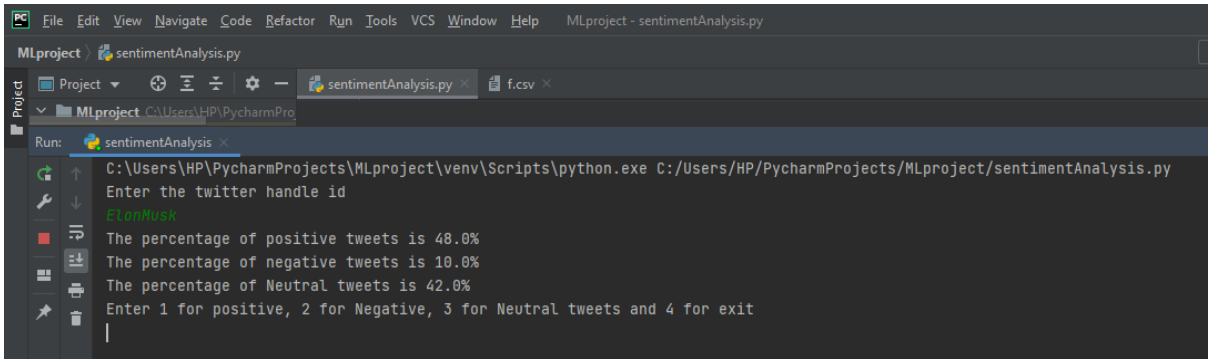
Sample Input/output Screens



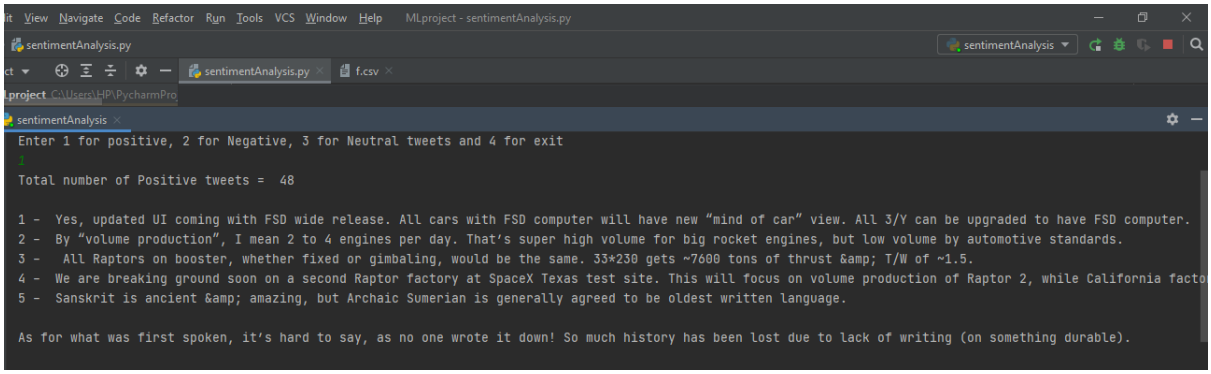
The python model is has a command line interface. That is the program will be run on command prompt. The user will have to enter the twitter user id of whose they want to check the sentiment.



The Id searched was Elon Musk it shows the percentage of positive, negative and neutral tweets and also show a pie chart.



It will also show a menu where you can see the positive, negative and neutral tweets.



The above snippet is an example of the positive tweets done through the user Id. It shows there were 48 positive tweets and all the tweets are listed below.