TWITTER SENTIMENT ANALYSIS
DEMONETISATION IN INDIA

15IT322E- PYTHON PROGRAMMING PROJECT REPORT

Submitted by

NIHARIKA KRISHNAN – RA1511008010126

PRASHANTH VAIDYA – RA1511008010148

ANIMESH RANJAN – RA1511003010335

for the assessment of Semester V Minor Project

SRM UNIVERSITY

KATTANKULATHUR

NOV 2017

DECLARATION

We

NIHARIKA KRISHNAN	RA1511008010126	Information Technology
PRASHANTH VAIDYA	RA1511008010148	Information Technology
ANIMESH RANJAN	RA1511003010335	Computer Science

studying in III year B.Tech at, SRM University, Kattankulathur, Chennai, hereby declare that this project is an original work of mine and we have not verbatim copied / duplicated any material from sources like internet or from print media, excepting some vital company information / statistics and data that is provided by the company itself.

Signature of the Students

Date:

Place:

ACKNOWLEDGEMENT

We express our deep sense of gratitude to our respected faculty Ms Monica Catherine for her valuable help and guidance. We are thankful to her for the motivation and support she has given throughout the project.

We are also grateful to Dr. G. Vadivu, Head of Department, Information Technology for giving us the opportunity to implement this Python project under the Department.

Lastly, we would like to express our deep appreciation towards our classmates and our indebtedness to our parents for providing us the moral support and encouragement.

NIHARIKA KRISHNAN – RA1511008010126

PRASHANTH VAIDYA – RA1511008010148

ANIMESH RANJAN – RA1511003010335

TABLE OF CONTENTS

S.NO.	TOPIC	PAGE
1.	Abstract	5
2.	Introduction and Overview	6
	2.1 Introduction	6
	2.2 Context	7
	2.3 Twitter Data	7
	2.4 Sentiment Analysis	8
3.	Requirement Analysis	9
	3.1 Data Acquisition	9
	3.2 Labeling	11
	3.3 Analysis	12
	3.3.1 TextBlob	12
	3.3.2 Matplotlib	12
5.	Project Code	13
	5.1 twitterAPI.py	13
	5.2 analyseTweets.py	14
	5.3 frontEnd.py	19
6.	Output	21
7.	Conclusion	24
8.	References	25

ABSTRACT

This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users - out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analyzing the sentiments expressed in the tweets.

Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. Predicting the opinions of the citizens of the country on the initiatives taken by the Government will help the government to improve their decision making body and work for the betterment of its citizens.

One such study is the Demonetization in India. Through this project, it helped predict the sentiment of the citizens of our country and concluded that twitter is a good reflection of offline sentiment.

INTRODUCTION AND OVERVIEW

2.1 INTRODUCTION:

This project of analyzing sentiments of tweets comes under the domain of "Pattern Classification" and "Data Mining". Both of these terms are very closely related and intertwined, and they can be formally defined as the process of discovering "useful" patterns in large set of data, either automatically (unsupervised) or semi automatically (supervised).

The project would heavily rely on techniques of "Natural Language Processing" in extracting significant patterns and features from the large data set of tweets. The features that can be used for modeling patterns and classification can be divided into two main groups: formal language based and informal blogging based. Language based features are those that deal with formal linguistics and include prior sentiment polarity of individual words and phrases, and parts of speech tagging of the sentence. Prior sentiment polarity means that some words and phrases have a natural innate tendency for expressing particular and specific sentiments in general.

For example the word "excellent" has a strong positive connotation while the word "evil" possesses a strong negative connotation. So whenever a word with positive connotation is used in a sentence, chances are that the entire sentence would be expressing a positive sentiment.

2.2 CONTEXT:

The demonetization of ₹500 and ₹1000 banknotes was a step taken by the Government of India on 8 November 2016, ceasing the usage of all ₹500 and ₹1000 banknotes of the Mahatma Gandhi Series as a form of legal tender in India from 9 November 2016.

The announcement was made by the Prime Minister of India Narendra Modi in an unscheduled live televised address to the nation at 20:15 Indian Standard Time (IST) the same day. In the announcement, Modi declared circulation of all ₹500 and ₹1000 banknotes of the Mahatma Gandhi Series as invalid and announced the issuance of new ₹500 and ₹2000 banknotes of the Mahatma Gandhi New Series in exchange for the old banknotes.

2.3 TWITTER DATA:

Twitter based features are more informal and relate with how people express themselves on online social platforms and compress their sentiments in the limited space of 140 characters offered by twitter. They include twitter hashtags, retweets, word capitalization, word, question marks, presence of url in tweets, exclamation marks, internet emoticons and internet shorthand/slangs. Classification techniques can also be divided into a two categories: Supervised vs. unsupervised and non-adaptive vs. adaptive/reinforcement techniques.

2.3 SENTIMENT ANALYSIS:

Supervised approach is when we have pre-labeled data samples available and we use them to train our classifier. Training the classifier means to use the pre-labeled to extract features that best model the patterns and differences between each of the individual classes, and then classifying an unlabeled data sample according to whichever pattern best describes it.

For example if we come up with a highly simplified model that neutral tweets contain 0.3 exclamation marks per tweet on average while sentiment-bearing tweets contain 0.8, and if the tweet we have to classify does contain 1 exclamation mark then (ignoring all other possible features) the tweet would be classified as subjective, since 1 exclamation mark is closer to the model of 0.8 exclamation marks. Unsupervised classification is when we do not have any labeled data for training. In addition to this adaptive classification techniques deal with feedback from the environment.

In our case feedback from the environment can be in form of a human telling the classifier whether it has done a good or poor job in classifying a particular tweet and the classifier needs to learn from this feedback. There are two further types of adaptive techniques: Passive and active. Passive techniques are the ones which use the feedback only to learn about the environment (in this case this could mean improving our models for tweets belonging to each of the three classes) but not using this improved learning in our current classification algorithm, while the active approach continuously keeps changing its classification algorithm according to what it learns at real-time. There are several metrics proposed for computing and comparing the results of our experiments.

REQUIREMENT ANALYSIS

The requirements of the projects are as follows:

I. Data Acquisition

II. Labeling

III. Analysis

3.1 DATA ACQUISITION:

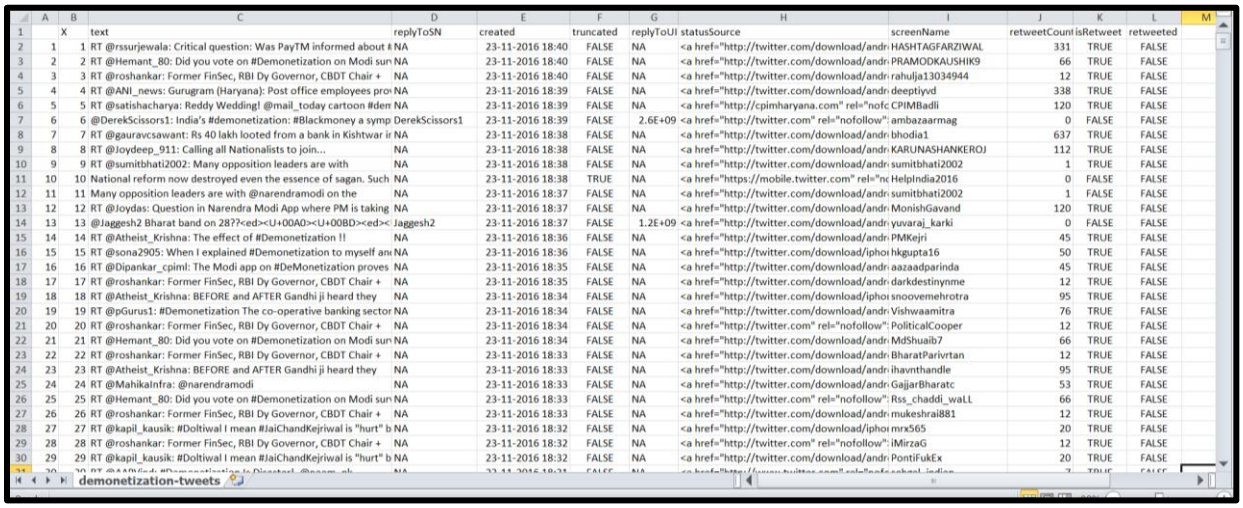
Data in the form of raw tweets is acquired by using the python library "tweestream" which provides a package for simple twitter streaming API. This API allows two modes of accessing tweets: SampleStream and FilterStream. SampleStream simply delivers a small, random sample of all the tweets streaming at a real time. FilterStream delivers tweet which match a certain criteria. It can filter the delivered tweets according to three criteria:

- Specific keyword(s) to track/search for in the tweets
- Specific Twitter user(s) according to their user-id's
- Tweets originating from specific location(s) (only for geo-tagged tweets).

A programmer can specify any single one of these filtering criteria or a multiple combination of these. But for our purpose we have no such restriction and will thus stick to the SampleStream mode. Since we wanted to increase the generality of our data, we acquired it in portions at different points of time instead of acquiring all of it at one go. If we used the latter approach then the generality of the tweets might have been compromised since a significant portion of the tweets would be referring to some certain trending topic and would thus have more or less of the same general mood or sentiment. A tweet acquired by this method has a lot of raw

information in it which we may or may not find useful for our particular application. It comes in the form of the python “dictionary” data type with various key-value pairs. A list of some key-value pairs are given below:

- Whether a tweet has been favourited
- User ID - Screen name of the user
- Original Text of the tweet
- Presence of hashtags
- Whether it is a re-tweet
- Language under which the twitter user has registered their account
- Geo-tag location of the tweet
- Date and time when the tweet was created



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		X	text	replyToSN	created	truncated	replyToUI	statusSource	screenName	retweetCount	isRetweet	retweeted	
2	1	1	RT @rssurjewala: Critical question: Was PayTM informed about t	NA	23-11-2016 18:40	FALSE	NA	<a href="http://twitter.com/download/andr	HASHTAGFARZIWAL	331	TRUE	FALSE	
3	2	2	RT @Hemant_80: Did you vote on #Demonetization on Modi sun	NA	23-11-2016 18:40	FALSE	NA	<a href="http://twitter.com/download/andr	PRAMODKAUSHIK9	66	TRUE	FALSE	
4	3	3	RT @roshankar: Former FinSec, RBI Dy Governor, CBDT Chair +	NA	23-11-2016 18:40	FALSE	NA	<a href="http://twitter.com/download/andr	rahulja13034944	12	TRUE	FALSE	
5	4	4	RT @ANI_news: Gurugram (Haryana): Post office employees pro	NA	23-11-2016 18:39	FALSE	NA	<a href="http://twitter.com/download/andr	deeptiyvd	338	TRUE	FALSE	
6	5	5	RT @satishacharya: Reddy Wedding! @mail_today cartoon #den	NA	23-11-2016 18:39	FALSE	NA	<a href="http://cpimharyana.com" rel="nofc	CPIMBadli	120	TRUE	FALSE	
7	6	6	@DerekScissors1: India's #demonetization: #Blackmoney a symp	DerekScissors1	23-11-2016 18:39	FALSE	2.6E+09	<a href="http://twitter.com" rel="nofollow":	ambazaarmag	0	FALSE	FALSE	
8	7	7	RT @gauravcsawant: Rs 40 lakh looted from a bank in Kishtwar i	NA	23-11-2016 18:38	FALSE	NA	<a href="http://twitter.com/download/andr	bhodia1	637	TRUE	FALSE	
9	8	8	RT @Joydeep_911: Calling all Nationalists to join...	NA	23-11-2016 18:38	FALSE	NA	<a href="http://twitter.com/download/andr	KARUNASHANKEROJ	112	TRUE	FALSE	
10	9	9	RT @sumitbhati2002: Many opposition leaders are with	NA	23-11-2016 18:38	FALSE	NA	<a href="http://twitter.com/download/andr	sumitbhati2002	1	TRUE	FALSE	
11	10	10	National reform now destroyed even the essence of sagan. Such	NA	23-11-2016 18:38	TRUE	NA	<a href="https://mobile.twitter.com" rel="nc	HelpIndia2016	0	FALSE	FALSE	
12	11	11	Many opposition leaders are with @narendramodi on the	NA	23-11-2016 18:37	FALSE	NA	<a href="http://twitter.com/download/andr	sumitbhati2002	1	FALSE	FALSE	
13	12	12	RT @Joydas: Question in Narendra Modi App where PM is taking	NA	23-11-2016 18:37	FALSE	NA	<a href="http://twitter.com/download/andr	MonishGavand	120	TRUE	FALSE	
14	13	13	@Jaggesh2 Bharat band on 28??<ed><U+00A0><U+00BD><ed><	Jaggesh2	23-11-2016 18:37	FALSE	1.2E+09	<a href="http://twitter.com/download/andr	yuvaraj_karki	0	FALSE	FALSE	
15	14	14	RT @Atheist_Krishna: The effect of #Demonetization !!	NA	23-11-2016 18:36	FALSE	NA	<a href="http://twitter.com/download/andr	PMKejri	45	TRUE	FALSE	
16	15	15	RT @sona2905: When I explained #Demonetization to myself an	NA	23-11-2016 18:36	FALSE	NA	<a href="http://twitter.com/download/iphor	hkgupta16	50	TRUE	FALSE	
17	16	16	RT @Dipankar_cpiml: The Modi app on #DeMonetization proves	NA	23-11-2016 18:35	FALSE	NA	<a href="http://twitter.com/download/andr	aazaadparinda	45	TRUE	FALSE	
18	17	17	RT @roshankar: Former FinSec, RBI Dy Governor, CBDT Chair +	NA	23-11-2016 18:35	FALSE	NA	<a href="http://twitter.com/download/andr	darkdestinynme	12	TRUE	FALSE	
19	18	18	RT @Atheist_Krishna: BEFORE and AFTER Gandhi ji heard they	NA	23-11-2016 18:34	FALSE	NA	<a href="http://twitter.com/download/iphor	snoovemehrotra	95	TRUE	FALSE	
20	19	19	RT @pGurus1: #Demonetization The co-operative banking sector	NA	23-11-2016 18:34	FALSE	NA	<a href="http://twitter.com/download/andr	Vishwaamitra	76	TRUE	FALSE	
21	20	20	RT @roshankar: Former FinSec, RBI Dy Governor, CBDT Chair +	NA	23-11-2016 18:34	FALSE	NA	<a href="http://twitter.com" rel="nofollow":	PoliticalCooper	12	TRUE	FALSE	
22	21	21	RT @Hemant_80: Did you vote on #Demonetization on Modi sun	NA	23-11-2016 18:34	FALSE	NA	<a href="http://twitter.com/download/andr	MdShuaib7	66	TRUE	FALSE	
23	22	22	RT @roshankar: Former FinSec, RBI Dy Governor, CBDT Chair +	NA	23-11-2016 18:33	FALSE	NA	<a href="http://twitter.com/download/andr	BharatParivrtan	12	TRUE	FALSE	
24	23	23	RT @Atheist_Krishna: BEFORE and AFTER Gandhi ji heard they	NA	23-11-2016 18:33	FALSE	NA	<a href="http://twitter.com/download/andr	ihavnthandle	95	TRUE	FALSE	
25	24	24	RT @MahikaInfra: @narendramodi	NA	23-11-2016 18:33	FALSE	NA	<a href="http://twitter.com/download/andr	GajjarBharatc	53	TRUE	FALSE	
26	25	25	RT @Hemant_80: Did you vote on #Demonetization on Modi sun	NA	23-11-2016 18:33	FALSE	NA	<a href="http://twitter.com" rel="nofollow":	Rss_chaddi_waLL	66	TRUE	FALSE	
27	26	26	RT @roshankar: Former FinSec, RBI Dy Governor, CBDT Chair +	NA	23-11-2016 18:33	FALSE	NA	<a href="http://twitter.com/download/andr	mukeshrai881	12	TRUE	FALSE	
28	27	27	RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is "hurt" b	NA	23-11-2016 18:32	FALSE	NA	<a href="http://twitter.com/download/iphor	mrx565	20	TRUE	FALSE	
29	28	28	RT @roshankar: Former FinSec, RBI Dy Governor, CBDT Chair +	NA	23-11-2016 18:32	FALSE	NA	<a href="http://twitter.com" rel="nofollow":	iMirzaG	12	TRUE	FALSE	
30	29	29	RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is "hurt" b	NA	23-11-2016 18:32	FALSE	NA	<a href="http://twitter.com/download/andr	PontiFukEx	20	TRUE	FALSE	

demonetization-tweets

Since this is a lot of information we only filter out the information that we need and discard the rest. For our particular application we iterate through all the tweets in our sample and save the actual text content of the tweets in a separate csv file given language of the twitter is user’s account is specified to be English.

3.2 LABELING:

We labelled the tweets in four classes according to sentiments expressed/observed in the tweets: positive, negative, neutral/objective and ambiguous. We gave the following guidelines to our labellers to help them in the labeling process:

- **Positive:**

If the entire tweet has a positive/happy/excited/joyful attitude or if something is mentioned with positive connotations. Also if more than one sentiment is expressed in the tweet but the positive sentiment is more dominant. Example: “4 more years of being in shithole Australia then I move to the USA! :D”.

- **Negative:**

If the entire tweet has a negative/sad/displeased attitude or if something is mentioned with negative connotations. Also if more than one sentiment is expressed in the tweet but the negative sentiment is more dominant. Example: “I want an android now this iPhone is boring :S”.

- **Neutral/Objective:**

If the creator of tweet expresses no personal sentiment/opinion in the tweet and merely transmits information. Advertisements of different products would be labelled under this category. Example: “US House Speaker vows to stop Obama contraceptive rule.

After Analysis, the final results were as follows:

- Total: 14941 tweets
- Positive: 5989 tweets
- Negative: 6723 tweets
- Neutral: 8218 tweets

3.3 ANALYSIS:

3.3.1.1 TEXTBLOB:

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

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- n-grams

3.3.2 MATPLOTLIB:

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code.

PROJECT CODE

4. 1 twitterAPI.py

```
import tweepy  #Library for accessing Twitter API
from textblob import TextBlob  #Library for Text Processing

#Setting up API credentials
#The credentials can be obtained by setting up your account at https://apps.twitter.com/
consumer_key=#insert consumer key here!
consumer_secret=#insert consumer secret here!
access_token=#insert access token here!
access_token_secret=#insert access token secret here

#Twitter API authentication
auth=tweepy.OAuthHandler(consumer_key,consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api=tweepy.API(auth) #API instance

#Printing random tweets and using them for some purpose!
public_tweets=api.search('#Demonetisation')

for tweet in public_tweets:
        text=tweet.text
        analysis=TextBlob(text)
        print(text)
        print(analysis.sentiment)
```

4. 2 analyseTweets.py

```
import csv
import random
import re
import codecs #provides transparent encoding/decoding
from textblob import TextBlob  #Library for Text Processing
import time
from collections import Counter

#Plotting dependecies
import matplotlib.pyplot as plt; plt.rcdefaults()
import numpy as np
import matplotlib.pyplot as plt

#Sentiment values
positive=0
negative=0
neutral=0
total=0

hashtags = []

#Loading.....
print("Performing Sentiment Analysis",end="")
for i in range(5):
        print(".",end="")
        time.sleep(1)
```

```
# reading csv file
filepath="demonetization-tweets.csv"
with codecs.open(filepath, "r",encoding='utf-8', errors='ignore') as csvfile:
        reader = csv.reader(csvfile)
        tweetsList=[]
        cleanTweetsList=[]
        for row in reader:
                tweet=row[2].strip() #contains tweet
                cleanTweet=" ".join(re.findall("[a-zA-Z]+",tweet))
                analysis=TextBlob(cleanTweet)

                #appending tweet to list
                tweetsList.append(tweet)
                cleanTweetsList.append(cleanTweet)



                #Assigning polarity and calculating count
                total=total+1
                if(analysis.sentiment.polarity>0):
                        positive=positive+1
                if(analysis.sentiment.polarity==0):
                        neutral=neutral+1
                else:
                        negative=negative+1
```

```
#Result
print() #newline
print("Total Tweets: ",total)
print('Positive = ',positive)
print('Neutral= ',neutral)
print('Negative= ',negative)

#Random number generator to pick tweets
randomTweets=[]
randomCleanTweets=[]

randomNumber=random.sample(range(1, 6000), 5)
index=0

#Collecting random tweets
for i in range(5):
        number=random.randint(1,6000) #Picks a random number between 1-6000
        randomTweets.append(tweetsList[randomNumber[index]])      #Stores      a
random tweet from data without repeatition
        randomCleanTweets.append(tweetsList[randomNumber[index]])
        index=index+1

for tweet in randomCleanTweets:
        print()
        print(tweet,end=' ')
        analysis=TextBlob(tweet)
        print(" => ",analysis.sentiment)
```

```
#Writing random tweets to a text file for display
with open('tweets.txt', 'w') as file:
        for tweet in randomTweets:
                file.write(tweet)
                file.write('<br><br>\n')

#Finding the hashtags in all the tweets
finalcount={}
for i in tweetsList:
        hashtags.append(re.findall(r"#(\w+)", i))
hashtagnew = [item for sub in hashtags for item in sub]
counts = Counter(hashtagnew)
counts = dict(counts)
finalcount = dict(sorted(counts.items(), key=lambda kv: kv[1], reverse=True))
countname = list(finalcount.keys())

#Plotting data
#Bar Graph
objects = ('Positive','Neutral','Negative')
y_pos = np.arange(len(objects))
performance = [positive,neutral,negative]

plt.bar(y_pos, performance, align='center', alpha=0.5)
plt.xticks(y_pos, objects)
plt.ylabel('# of tweets')
plt.title('Twitter Sentiment Analysis- Demonetisation (Bar Graph) \n')
plt.show()
```

```
#Pie Graph

colors = ['yellowgreen', 'gold', 'orangered']
explode = (0, 0, 0.1)  # explode last slice

plt.pie(performance, explode=explode, labels=objects, colors=colors,
        autopct='%1.1f%%', shadow=False, startangle=140)

plt.axis('equal')
plt.title('Twitter Sentiment Analysis- Demonetisation (Pie Chart) \n')
plt.show()

# Hashtag Plot
x = np.arange(len(finalcount))
y = list(finalcount.values())
x = x[:15]
y = y[:15]

countname = countname[:15]
plt.bar(x, y)
plt.title('Most Trending Hashtags\n')
plt.xticks(x, countname, rotation='vertical')
plt.ylabel('Number of tweets')
plt.xlabel('#Hashtags')
plt.tight_layout()
plt.show()
```

4. 2 frontEnd.py

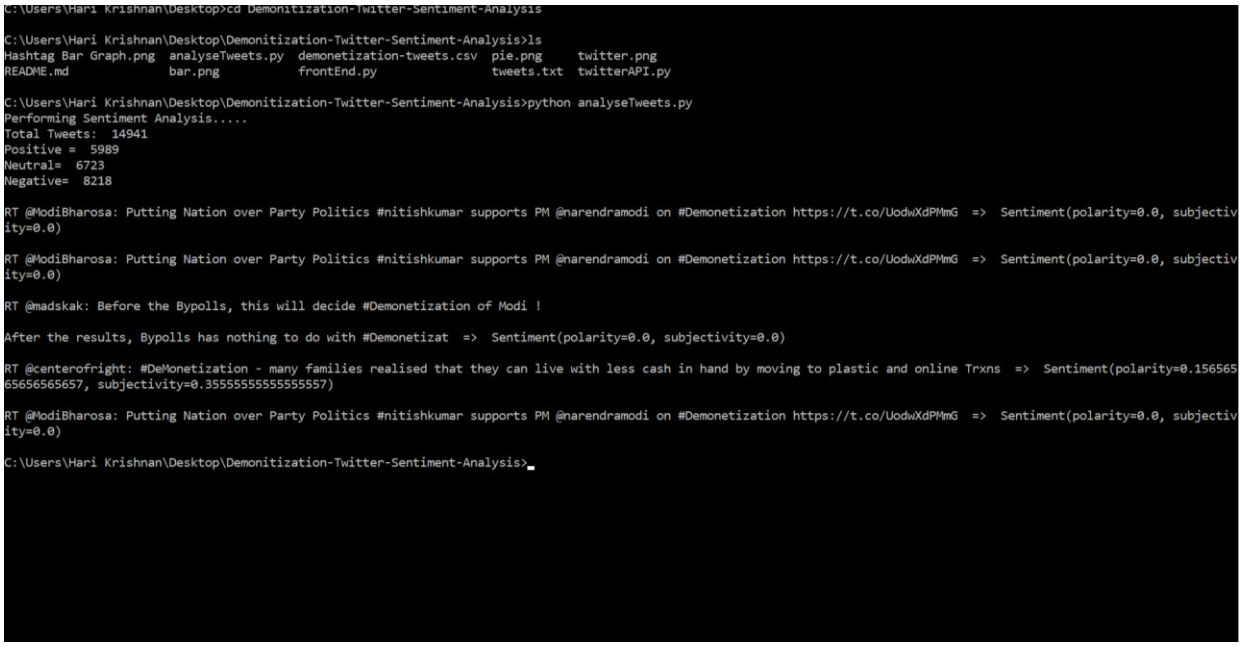
```
#!/usr/bin/python

import cgi, cgitb
tweet=""

file = open('tweets.txt', 'r')
text = file.read()
file.close()

tweet=text
print(text)
print ("Content-type:text/html\r\n\r\n")

print ("""
<!DOCTYPE html>
<html>
<head>
        <title>Sentiment Analysys</title>
</head>

<body style="background-color:#4ab3f4">
<div id="header-image" style="position:absolute;top:0">
<img src="twitter.png">
</div>
<div id="bar-graph" style="position:absolute;top:10vw;left:0.6vw;">
```

```
<img src="bar.png" height=500px>
</div>
<div id="pie-graph" style="position:absolute;top:10vw;right:12vw;">
<img src="pie.png"  height=500px>
</div>
<div id="hashtag" style="position:fixed;top:60%;left:40%; transform: translateX(-50%) translateY(-50%);height: 100px;width: 500px;">
<img src="Hashtag Bar Graph.png"  height=500px>
</div>
<div id="tweet" style="position:fixed;top:25%;left:45%; transform: translateX(-50%) translateY(-50%);height: 100px;width: 500px;">
<strong>
<span style="text-decoration: underline;">Some random tweets: </span>
<br><br>
""")

print(tweet)

print("""</strong></div>
</body>
</html>

""")
```

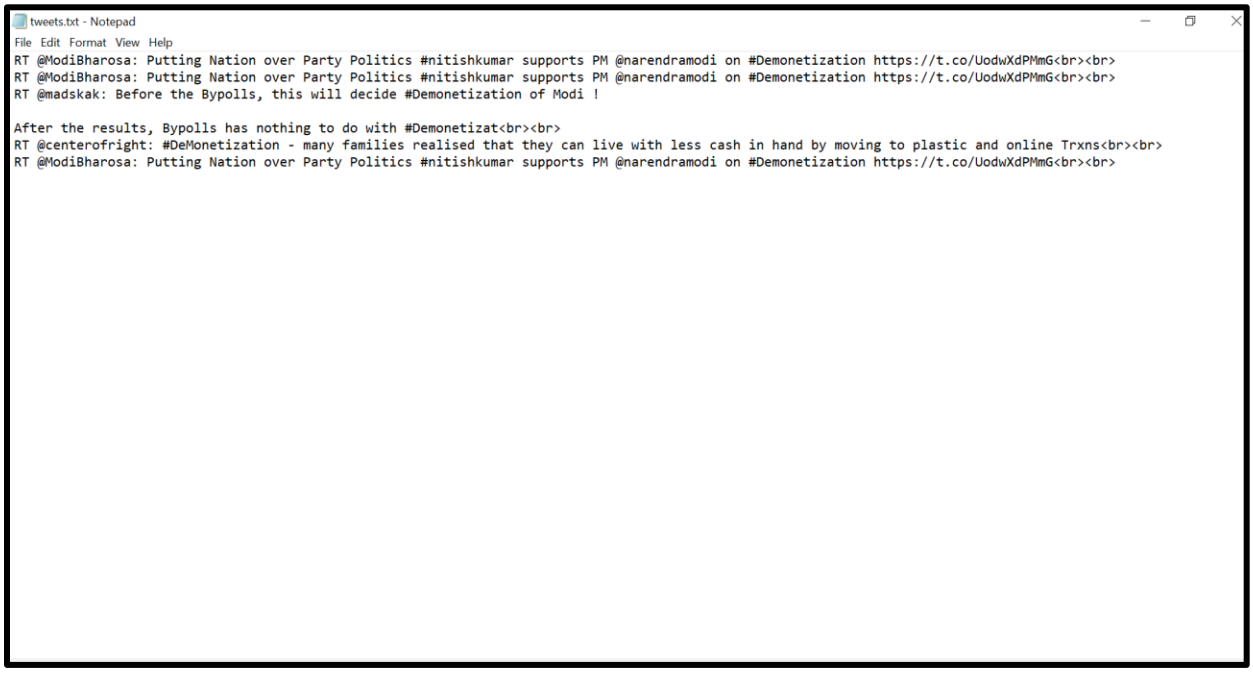
OUTPUT:

The output consists of:

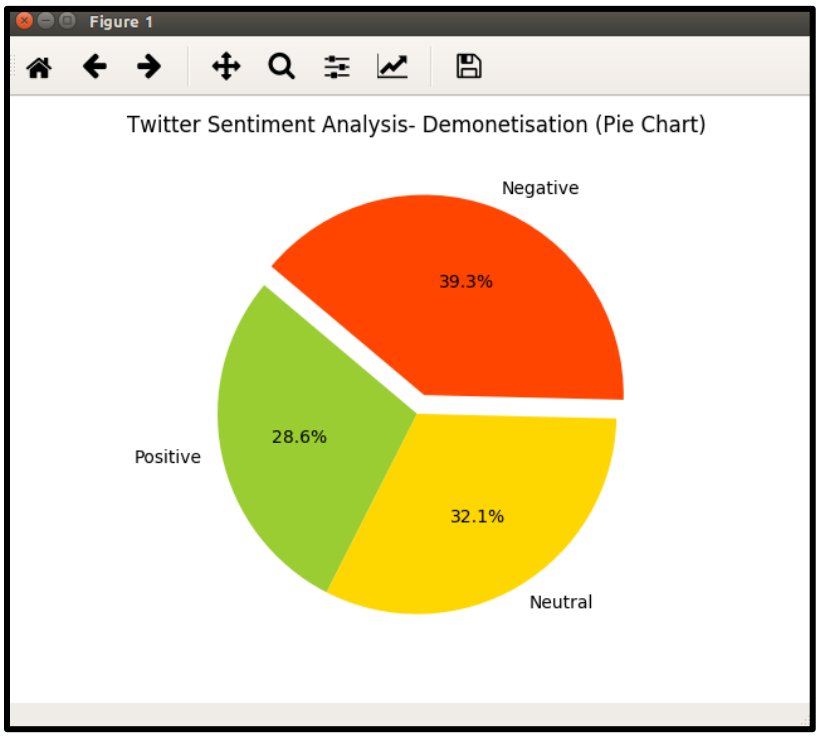
- Total number of tweets
- Number of tweets positive, negative and neutral
- 5 random tweets with their polarity and sentiment



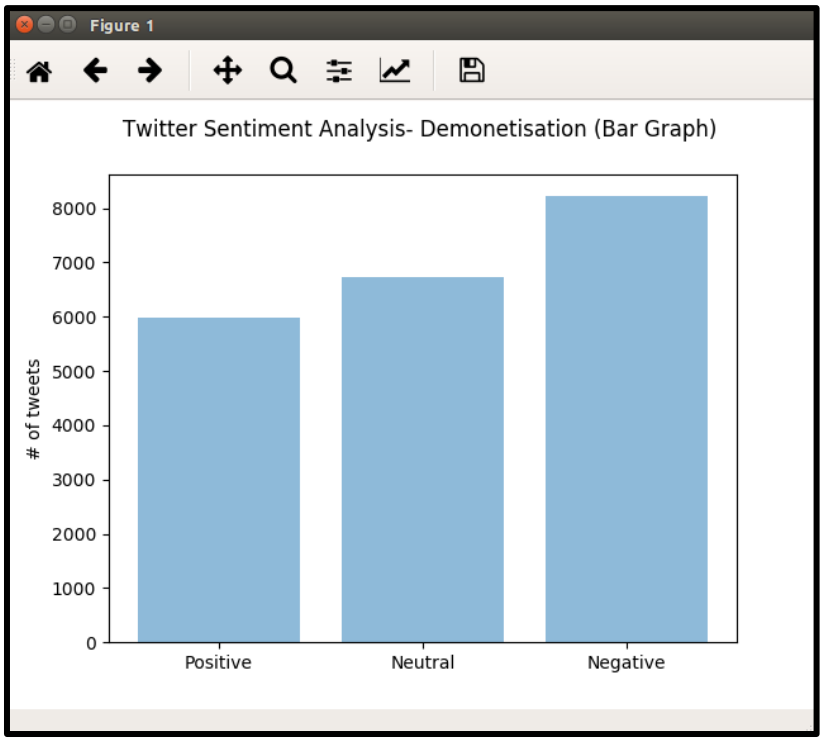
The tweets when written to tweets.txt file:



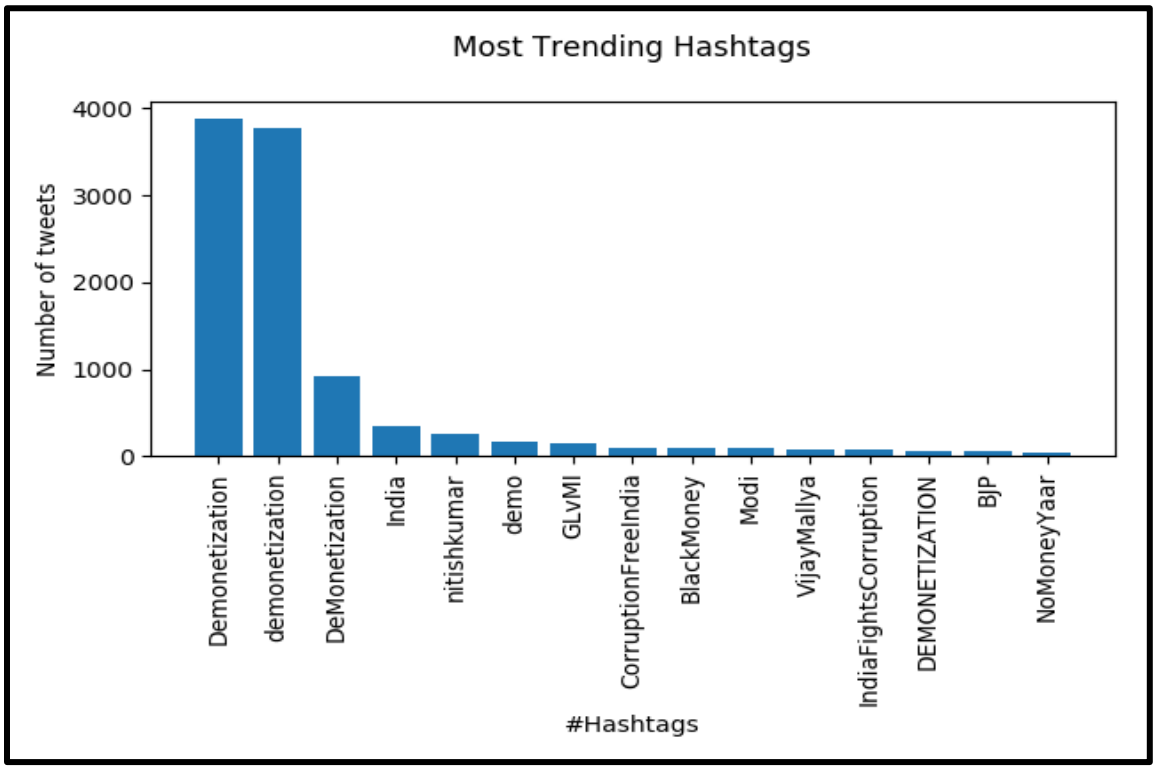
Pie Chart Representation of Sentiments



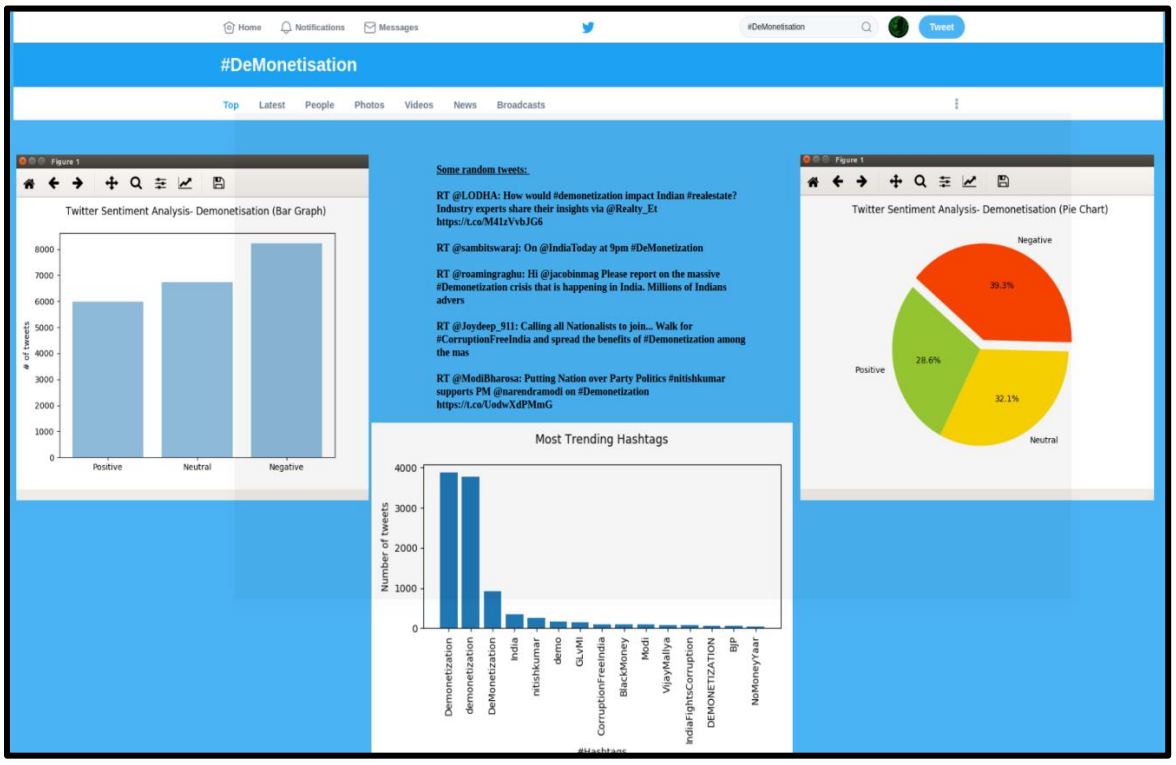
Bar Chart Representation of Sentiments



Bar Chart of Trending Hashtags



Webpage using cgi script



CONCLUSION

Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. Predicting the opinions of the citizens of the country on the initiatives taken by the Government will help the government to improve their decision making body and work for the betterment of its citizens.

Through this project we came to a conclusion that the opinion of the citizens on demonetization is generally neutral or negative.

REFERENCES

- Alexander Pak, Patrick Paroubek. 2010, Twitter as a Corpus for Sentiment Analysis and Opinion Mining.
- Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision.
- Jin Bai, JianYun Nie. Using Language Models for Text Classification.
- http://en.wikipedia.org
- http://help.sentiment140.com/for-students