Journal of Electrical Engineering and Computer Science
ISSN: 2287-106X
Volume 01, Issue 03, October, 2019

Data Analysis with Twitter Data Mining

Mansi Bhavar[1], Rushit Dave[2], Evelyn Sowellz Boone[3]
Electrical & Computer Engineering, NCAT, USA[1], Computer Science, NCAT, USA[2], Computer Systems Technology, NCAT, USA[3]

Abstract— This paper gives a brief overview of Big data and data analysis using twitter application. For running a successful business, we must deal with a large amount of data. This paper explains methods for storing, Designing and examine a large amount of real-time data. For making productive decisions for real-time application, we are fetching the real-time data from the Twitter Application. We are using the Twitter Application for our source of data. Twitter's acceptance in the social media gives the bunch of data assets (that is information) and led towards the research in the different area. Demographic, academic, social science, stock predictions are some of the most famous examples for researchers.

Keywords— Data Analysis, Big Data, twitter, sentimental analysis.

1. Introduction

In this changing era, data can be consumed and develop at a tremendous amount. Moreover, social networking sites are also growing at a fast rate. From all this improvement, Big data comes in the picture. And to store this significant amount of data with safely is also a big challenge. For that many researchers are work on the big data.

Big data is nothing but the broad set of data which can be analyzed arithmetically to release the human behavior, interaction, etc. it gives the new value for the organization and opens the new path towards the innovation. This is due to the microbiology which led the different opinion of people such as various topics, express their emotion, the social message to society, current issue and many more which are used in day to day life. And the companies are analyzing the people's perspective and provide the solution to make consumers experience well. But it's time-consuming. Therefore, the challenge is to find the proper outcome which analyzes the data for more efficient design and better decisions.

In this illustration, with the help of the Twitter application, we analyze and conceptualize the different aspects of microbiology in providing the active service.

2. Definitions

2.1 Big Data

It is the source of a large amount of structured and unstructured data. For business, it helps drive efficiently, quality and services, also improve the level of customer satisfaction.

2.2 Twitter

It is the source of information. It is a social networking application which is very simple, attractive and public. It allows solving the multiple interrogations such as analyze the tweets about a specific word or topic within the last five minutes or target the users' tweets which tells about the number of positive tweets and bad tweets.

The approach in this paper is the use of real data. Bundle of data can be collected from Twitter and it requires to label the tweets of polarity which subjects to a query. There are many social platforms whose

generates the tons of data, but the Twitter is the publicly available dataset also, it is user-friendly. Their tweets are mostly used for sentimental analysis. Twitter Application Programming Interface (API) querying the gathered dataset from Neutral Polar Irrelevant training data.

2.3 Python

Python is a simple programming language. Python libraries are quite simple and useful for data analysis. Even if you are a first-time user, it gives the flexibility in statistical data and in analyzing data. Therefore, most of the data scientists are preferred python to be a perfect fit for data analysis. Here, we are using the python3.7.0 for writing the code and using Spyder with Anacoda3.

3. Methodology

To perform the Twitter API for data mining does not appear to be modifying the subject the following.

3.1 Collecting Data

To access the tweets of Twitter, we first need to create and register the twitter app on twitter website for verification, and then we can access the data by using the Twitter API.

a) Creating the App:

To create the twitter account, first, download the app from http://apps.twitter.com/.

After that create the new account and fill the details and then you will receive the four main credentials.

i. Get the consumer_key
ii. Consumer_secret_key
iii. Access_token
iv. Access_token_secret

This four keys and tokens are very useful tokens in the later sessions and keep these tokens.

b) Data access:

Together the data with python, we need to install the Tweepy. It is an open source python library. It gives access to connect with the Twitter platform. It fetches the real-time data. In Tweepy, it works with two important API.

i. Streaming API: it provides real-time access to data. And the output is in JSON format.
ii. Reset API: it gives the user information and posts all the tweets.

For installing the Tweepy, use below command in command prompt.

Conda install -c conda-forge tweepy

After this, we need to authenticate the credentials to access the data directly with command.

For that, we are using the OAuth interface to configure the keys to get access to Twitter directly through python. Now, define the variable which gives the entry point from where we can call the twitter.

```
6
7 auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
8 auth.set_access_token(OAUTH_TOKEN, OAUTH_TOKEN_SECRET)
9
10 api = tweepy.API(auth)
11
```

Figure 1 Initializing the keys to access Twitter

c) Data storage:

After getting access to all tweets data from the twitter we need to store that data in JSON format to analyze the data. For that, the tweepy library gives the simple interface to capsulize all tweets and save all the data in. JSON format.

d) Design the data:

To analyzing the twitter data, preprocessing of data is very important in analysis. Preprocessing is the simple method to get the data and design it to give the desired outcome of fulfilling the requirement.

Before we start to analyze data, it's important to know what Tweet is. It is a quick and short message which are express the human emotions, unique thoughts, make the spelling mistake and many more. Users of Twitter use the '@' symbol to relate to the other person on the blog. One popular terminology is Hashtags. It is used for marking the specific matter. For that, they are using the '#' symbol. Some specific character to express their emotions socially.

A tweet contains the more information regarding the person's data, the date of creating the tweets, location, text, thought regarding that tweet, emotions, the personality of users, positive or negative tweets and many more. We are using some of those aspects of analysis.

The useful fields for a tweet are as follows:

The text= text of the tweet.
Language = tweet language such as English, Gujrati, Hindi etc.
Like= how many times users like the tweets
Date= date of creating that date.

From all above entities, we can determine the tweets like how many times user's tweet, like tweets by user, most favorites tweets, hashtags of tweets and many more. Also, we can do the sentimental analysis.

e) Analysis the data:

After tokenizing the data, we can now perform the different analysis of that data.

- Person's tweet:

The most common analysis with Twitter data mining is extracting the tweeter's information. The information is like Names of that person, followers of that person, most recent 5 tweets which were uploaded by that user. Another data can be mine which is user's personal information that is id, source, most liked tweet counts, retweets counts so on and so forth.
Here, in this paper, we mentioned the 'Barack Obama' who was the President of USA. Analyze their tweet. From this we can use this kind of information to know their view on some topic, also we can find the person's nature how they behave in a certain situation.

Following shows the code to analyze the user information.

```
api = tweepy.API(auth)

#twitter.api.GetFollowers
user = api.get_user('BarackObama')
print("SCreen Name:", user.screen_name)
print("Number of followers:", user.followers_count)

for friends in user.friends():
    print("Screen Name:", friends.screen_name)

# creating a tweet list from the obama
tweets = api.user_timeline(screen_name="BarackObama", count=200)
print("Number of tweets extracted: {}.\n".format(len(tweets)))

```

Figure 1 python code for fetching the user information

- Commonly used words:

The above term means that the word which is famous for tweeter users. How many times people are using that word for expressing their emotions? It is generally used for analysis to see the person's thoughts and whether it is positive or negative.

- Most likely tweet:

Here, in this paper, we are illustrating the favorite tweets which are having the highest possible on twitter.

- Upload the tweet:

In this analysis, we are uploading the tweet which user/developer wants to upload on their timeline. It shows that with the help of Tweepy in a python we can tweet directly without opening the Twitter app.

It can be manageable from the user's end. There are also further doing the survey about the tweet is positive or negative with respect to the reference polarity. If your tweet is highly negativity it will not upload your tweet, it shows the pop up which defines the tweet is negative.

In this paper, we are tries to build a message box which prompt to ask what tweet you need to upload.

The message box looks like

```
# printing most 5 recent tweets by obama
print("5 recent tweets:\n")
for tweet in tweets[:5]:
    print(tweet.text)
    print()

public_tweets = api.user_timeline()
for tweet in public_tweets:
    print(tweet.text)

# showing the information regarding the tweets
print(tweets[0].id)
print(tweets[0].created_at)
print(tweets[0].source)
print(tweets[0].favorite_count)
print(tweets[0].retweet_count)
```

Figure 2 python code printing the user information

Journal of Electrical Engineering and Computer Science
ISSN: 2287-106X
Volume 01, Issue 03, October, 2019

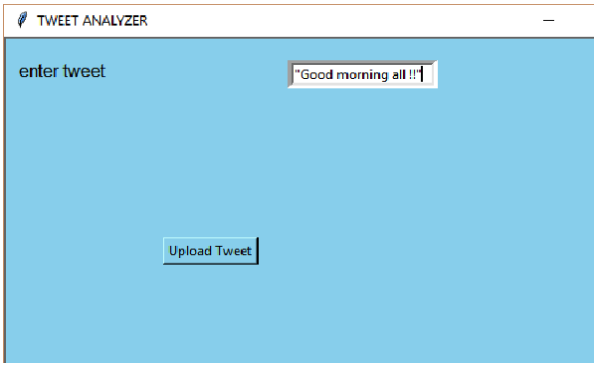

Figure 3 Input message box for uploading tweet

Below shows the python code on uploading the Tweet with sentimental analysis.

```
13
14    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
15    auth.set_access_token(access_token, access_token_secret)
16
17    tweepyapi = tweepy.API(auth)       # defining input
18    s1=name.get()
19
20    sub=TextBlob(s1)
21    p,s=sub.sentiment
22    print("polarity:",p)
23
24    print("sub",s)
25    if p<=-.2:
26
27        print("opzzz ur tweet contain negetivity cant upload.")
28
29
30    else:
31        tweepyapi.update status(name.get())
```

Figure 4 python code for authentication & uploading the tweet

In this literature, we define the function which also shows the polarity and subjectivity of that tweet. If the tweet is with negative polarity, then it shows the message and it doesn't allow the person to upload the tweet.

- Sentimental analysis:

It is the process of review from the different data. It is done by the collection of data from social media. It determines whether the sentence express positive, negative or neutral emotion towards the tweet. It is also known as opinion base which gives the opinion, thought and attitude of the speaker from their tweet. A sentimental analysis is useful for business as they develop their tricks to know the consumer's emotions towards their company product. It is also useful in politics as they keep the track regarding the political view.

In this paper, we are showing the sentimental analysis with positive, negative and neutral tweets. Also, differentiate positive and negative tweets. If it is negative, then it will show the message like 'Oops your tweet contains the negativity can't upload'. It also analyzes the 'Obama vs Trump' tweet what percentage the tweet should positive, negative and neutral with a percentage. Also, it shows the line graph which shows the subjectivity vs polarity graph. Subjectivity tends to show the client's own feelings, emotions, viewpoint, experiences and many more. Whereas polarity tends to a correlation of subjects with a given polarity of that sentiment. With this type of analysis helps to determine the tweet's sentiment.

4. Results

For uploading the tweets:

1. Upload the tweets of "Good Morning all!!".

Output:

polarity: 1.0

sub 0.6000000000000001

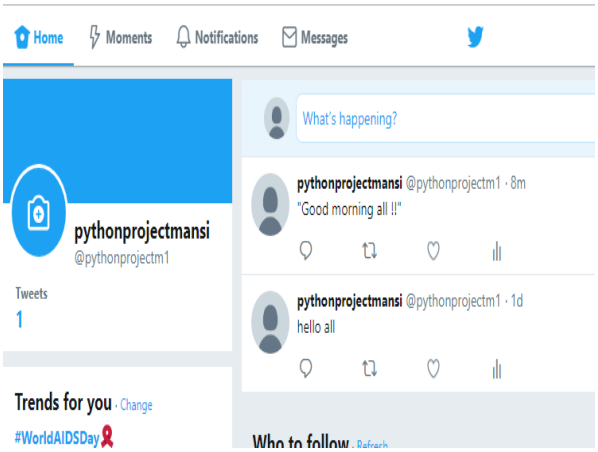

Figure 5 output of uploading the tweet

2. When it is a negative tweet like **"feels bad"**
It will show the message me like below,

polarity: -0.6999999999999998
sub 0.6666666666666666
"Opzzz your tweet contains negativity can't upload."

For detailed information, it will see the output below:

```
In [2]: runfile('C:/Users/Mansi Bhavsar/Deskt
Bhavsar/Desktop')
SCreen Name: BarackObama
Number of followers: 103450315
Screen Name: oddwun
Screen Name: Jeanette_Vea
Screen Name: mindcorners
Screen Name: Sean_007
Screen Name: jenmidori
Screen Name: MartiniGuy
Screen Name: BitchesBinders
Screen Name: atouilcanada
Screen Name: SusanWesthof
Screen Name: Brenmarie
Screen Name: somaris1
Screen Name: danastafford
Screen Name: jakstrong
Screen Name: robertwolf32
Screen Name: davidplouffe
Screen Name: DevalPatrick
Screen Name: johndoerr
Screen Name: AndreLohse
Screen Name: themjeans
Screen Name: MsTurnageYoung
Number of tweets extracted: 200.
```

Figure 6 output of analyzing the tweet of Obama

5 recent tweets:

America has lost a patriot and humble servant in George Herbert Walker Bush. While our hearts are heavy today, they… https://t.co/2Gn7sGTG90

I am grateful for the next generation of leaders -- the young people who are tolerant, creative, idealistic and do… https://t.co/Mnydz2BMZX

Thanks to the Chicago @FoodDepository team for all you do and to the volunteers who are doing great work and let me… https://t.co/noTTWJDrlv

RT @ObamaFoundation: The Obama Presidential Center represents a historic opportunity to build a world-class museum and public gathering spa…When someone shares their story, we see the world through their eyes. I'm looking forward to hearing a few from lea… https://t.co/4ZM6S85yuU

"Good morning all !!"

hello all

1068740570921738241

2018-12-01 05:36:23

Twitter Web Client

360090

68116

For sentimental analysis:

The result shows the what word of tweet we should analyze.

Also, it gives the option such as Random word analysis and trump vs Obama graph.

Word analysis is a pie-chart which shows the what percentage of a word is positive, neutral and negative in their tweets. Whereas line graph shows the polarity vs subjectivity in their tweets over the 250 tweets.

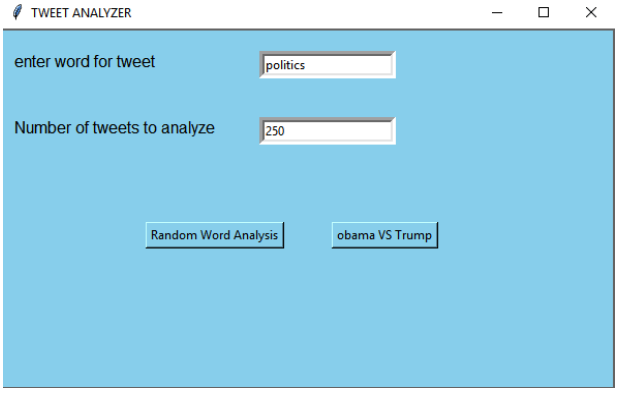

Figure 7 Tweet analyzer message box

PIE CHART:

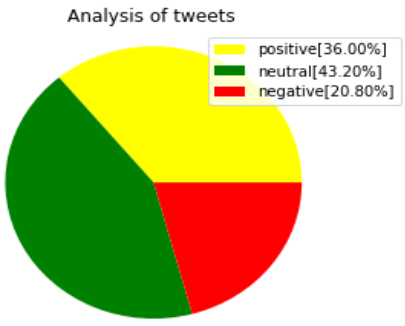

Figure 8 output of analysis of tweet

It shows the polarity of x and y axis which are as follows
x ['-0.20', '0.20', '0.50', '-0.03', '-0.20', '0.17', '0.32', '0.20', '0.00', '0.00', '0.50', '0.22', '-0.10', '0.05', '0.10', '0.45', '0.00', '0.00', '0.00', '-0.25']

y ['0.20', '0.00', '0.13', '0.44', '0.12', '0.00', '0.00', '0.00', '0.50', '1.00', '0.71', '0.17', '0.45', '0.00', '0.00', '0.40', '0.44', '0.00', '0.00', '0.00']

LINE GRAPH:

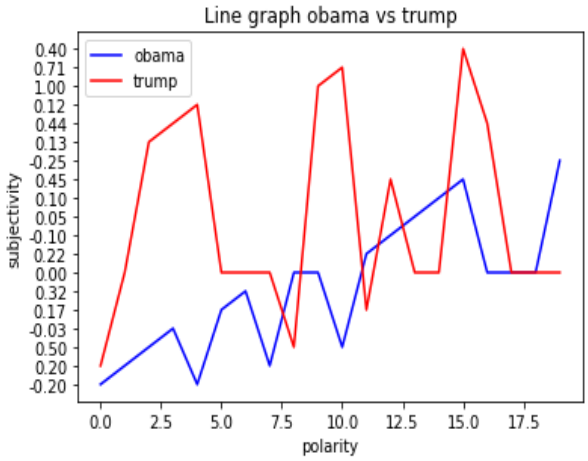

Figure 9 analysis of tweet of Obama vs trump

5. Literature survey

[1] from the lecture notes we refer the machine learning, python and other relevant libraries which uses throughout the paper.

[2] An introduction with twitter data analysis gives the brief introduction of how to create the twitter application, also provides with some of the examples which gives the complete overview of how to deal with the twitter data.

[3] Paramjit Singh, March 2014 paper gives the detailed discussion on how polarity and subjectivity matter for analyzing the tweet. They took the simple example on tweeter statistics and explain it with simple graphical presentation.

[4] Scott (2011) describes the importance of social network analysis with their new development and viewpoint with respect to the analysis. They briefly give the idea of data mining and network analysis.

[5] DEV, from this link we can easily understand the python code in some partial way. He describes how to extract the hashtags, followers, their location and many more with short examples, which helps a lot to reach our aim.

[6] with the Documentation it clearly provides the all documentation and library related to twitter. Also, it specifies some examples which will help to which library we can use for our analysis and how it works.

6. Conclusion

In this paper, we describe the very simple twitter analysis. We explained how OAuth and Tweepy should

authenticate to use Twitter. With collecting data, we should preprocess the data using tokenizers.

In the following steps, we attempt to analyze the stored data. Which also gives the username, their source, friends name, followers count and many more. We then move towards to how to upload the tweet. It just looks simply but here we develop the function which defines whether your tweet is positive or negative and if it is negative then it will not allow you to upload the tweet.

With line graph, it gives the idea about the polarity and subjectivity of tweets which compares the tweet between Obama and Trump tweet with which gives the more positive and polar tweets.

In this paper, we just want to present the simple data analysis technique for that it is quite simple in nature but also it provides some good knowledge on Sentimental analysis with some good data analysis of Twitter analysis.

In the last step, we define the sentimental analysis and tried to analyze the tweet with over 250 tweets of Obama and Trump tweets and define the analysis with a pie chart to describe how many tweets are positive, neutral and negative with their percentage.

7. References

[1] Lecture notes.

[2] An introduction to Twitter data analysis in python (DOI: 10.13140/RG.2.2.12803.30243) published in September 2016.

[3] Polarity classification using Twitter data by Paramjit Singh from IJCSMC (ISSN 2320-088X) published on March 2014.

[4] Scott, J. (2011). Social network analysis: developments, advances, and prospects. Social network analysis and mining, 1(1), 21-26.

[5] DEV, Sentimental analysis on Trump's tweet using python https://dev.to/rodolfoferro/sentiment-analysis-on-trumpss-tweets-using-python-

[6] Python-twitter google-groups, Python-twitter documentation, Release 3.4.1 published on March 2018.