SENTIMENT ANALYSIS USING PYTHON

Cluster Innovation Centre

University of Delhi

MOHAMMAD ABDULLAH

Summer Internship Project submitted for the Topic

Sentiment Analysis using python

Mentored by

Shashank Shekhar

Software Engineer

(Global Logic India)

Certificate of Originality

The work embodied in this report entitled **"Sentiment Analysis Using PYTHON"** has been carried out by **Mohammad Abdullah** for the "**Semester Long Internship project "**. We declare that the work and language included in this project report is free from any kind of plagiarism.

(**Mohammad Abdullah)**

Acknowledgement

The success and outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along with the completion of our project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I owe my deep gratitude to my project supervisor Mr.Shashank Shekhar, who took a keen interest in my project work and guided me all along, till the completion of my project work by providing all the necessary information for developing a good system.

We respect and thank our B.Tech Program Coordinator Prof.Shobha Bagai, for providing us an opportunity to do the project and giving us all the support and guidance, which made us complete the project duly. We are extremely thankful to her for providing such a nice support and guidance, although she has a busy schedule throughout.

Thank you to all once again

Abstract

Social media websites have emerged as one of the platforms to raise users' opinions and influence the way any business is commercialized. Opinion of people matters a lot to analyze how the propagation of information impacts the lives in a large-scale network like Social media. Sentiment analysis of social media data determines the polarity and inclination of a vast population towards a specific topic, item or entity. These days, the applications of such analysis can be easily observed during public elections, movie promotions, brand endorsements and many other fields. The primary aim is to provide a method for analyzing sentiment score in noisy social media streams. Results classify user's perception via tweets into positive and negative.

Acronyms

SVM	Support Vector Machine
NB	Naïve Bayes Classifier
NLTK	Natural Language Toolkit
POS	Part of Speech
SA	Sentiment Analysis
PT	Partial Tree

Symbols

b	Bias vector
w	Weight vector
φ()	Non Linear Mapping function X Feature vector

Chapter 1: Introduction

As the internet is growing bigger, its horizons are becoming wider. Social Media and Micro blogging platforms like Facebook, Twitter, Tumblr dominate in spreading encapsulated news and trending topics across the globe at a rapid pace. A topic becomes trending if more and more users are contributing their opinions and judgements, thereby making it a valuable source of online perception. These topics are generally intended to spread awareness or to promote public figures, political campaigns during elections, product endorsements and entertainment like movies, award shows. Large organizations and firms take advantage of people's feedback to improve their products and services. Sentiment Analysis in social media data is quite difficult due to its short length. Presence of emoticons, slang words and misspellings in tweets forced to have a preprocessing step before feature extraction. There are different feature extraction methods for collecting relevant features from text which can be applied to tweets also. But the feature extraction is to be done in two phases to extract relevant features. In the first phase, twitter specific features are extracted. Then these features are removed from the data to create normal text. After that, again feature extraction is done to get more features. This is the idea used in this project to generate an efficient feature vector for analyzing data sentiment. Since no standard dataset is available for social media posts of electronic devices, we created a dataset by collecting data for a certain period of time from twitter using tweepy. Sentiment analysis enlightens user whether the information concerning the product is satisfactory or not before they get it. Marketers and firms utilize this analysis data to comprehend about their products or services in a manner that it can be offered according to the user's requirements. Textual Information retrieval techniques primarily concentrate on processing, searching or analyzing the factual data show. Actualities have an objective component yet, there are some other textual contents which express subjective characteristics. These contents are for the most part opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis. It offers numerous challenging opportunities to develop new applications, for the most part because of the immense development of available information on online sources like blogs

and social systems. For instance, recommendations of items proposed by a suggestion system can be predicted by considering considerations, for example, positive or negative opinions about those items by making utilization of Sentiment Analysis. The automated process which helps in extracting the attitudes, opinions and other emotions present within the various types of information generated by the users in the form of text, speech or tweets is known as the sentimentanalysis process. There are various opinions present within the data which can be categorized into three broader categories namely positive, negative and neutral. There is a difference between the words utilized in some aspects instead of sentiment such as views, beliefs, opinions and so on.

1.1 Motivation

This project deals with the social media data for sentiment analysis. Social Media Platforms are becoming wider with every passing day. There is a lot of Data which we can retrieve from these platforms. The data is usually the opinion of the people regarding upcoming elections or an upcoming technology, etc. If we have an opinion or reviews of a very large audience, we can improve the product,etc. We can even predict using the opinion of the people. For Example we can predict the result of the upcoming elections if we get to know the opinion of the voters beforehand. Not only this we can expand the use of sentiment analysis to judge institutions by the reviews of the students studying there or the teachers teaching there. Sentiment analysis has vast scopes which are unlimited and never seems to get obsolete.

1.2 Objectives and Scope

The objective is to build an application which can extract data from social media, regarding a particular issue or a topic, and then classify it as positive, negative or neutral.

This would help in classifying the opinions and reviews of the users, voters ,etc.

Chapter 2: Literature Review

As the available social platforms are shooting up, the information is becoming vast and can be extracted to turn into business objectives, social campaigns, marketing and other promotional strategies. The benefit of social media is to know public opinions and extract their emotions, for example twitter gives advantage during elections. Further, the concept of the hashtag is used for text classification as it conveys emotion in a few words.

Another significant effort for sentiment classification on Twitter data is by Barbosa and Feng. They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like retweets, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. I will extend their approach by using real valued prior polarity, and by combining prior polarity with POS. Our results show that the features that enhance the performance of our classifiers the most are features that combine prior polarity of words with their parts of speech. The tweet syntax features help but only marginally. Gamon performs sentiment analysis on feedback data from Global Support Services survey. One aim of their paper is to analyze the role of linguistic features like POS tags. They perform extensive feature analysis and feature selection and demonstrate that abstract linguistic analysis features contribute to the classifier accuracy. In this paper we perform extensive feature analysis and show that the use of only 100 abstract linguistic features performs as well as a hard unigram baseline.

Let's talk a bit about the data description, as it plays a very important role. Twitter is a social networking and microblogging service that allows users to post real time messages called tweets. Tweets are short messages, restricted to 140 characters in length. Due to the nature of this microblogging service people use acronyms, make spelling mistakes, use emoticons and other characters that express special meanings. Following is a brief terminology associated with tweets.

Emoticons: These are facial expressions pictorially represented using punctuation and letters; they express the user's mood.

Target: Users of Twitter use the "@" symbol to refer to other users on the microblog. Referring to other users in this manner automatically alerts them.

Hashtags: Users usually use hashtags to mark topics.

This is primarily done to increase the visibility of their tweets. I have acquired data from freely available sources. These free sources collect the data by archiving the real-time stream. No language, location or any other kind of restriction was made during the streaming process. In fact, their collection consisted of tweets in foreign languages. They use Google translate to convert it into English before the annotation process. Each tweet is labeled by a human annotator as positive, negative, neutral or junk. The "junk" label means that the tweet cannot be understood by a human annotator. A manual analysis of a random sample of tweets labeled as "junk" suggested that many of these tweets were those that were not translated well using Google translate. I eliminated the tweets with junk labels for experiments.

It also becomes important to throw some light on the Design of the tree kernel. We design a tree representation of tweets to combine many categories of features in one succinct convenient representation. For calculating the similarity between two trees we use a Partial Tree kernel. A PT kernel calculates the similarity between two trees by comparing all

possible sub-trees. This tree kernel is an instance of a general class of convolution kernels. Convolution Kernels, first introduced by Haussler , can be used to compare abstract objects, like strings, instead of feature vectors. This is because these kernels involve a recursive calculation over the "parts" of abstract objects. This calculation is made computationally efficient by using Dynamic Programming techniques. By considering all possible combinations of fragments, tree kernels capture any possible correlation between features and categories of features. This tree is for a synthesized tweet: @Fernando this isn't a great day for playing the HARP! :). We use the following procedure to convert a tweet into a tree representation: Initialize the main tree to be "ROOT". Then tokenize each tweet and for each token: a) if the token is a target, emoticon, exclamation mark, other punctuation mark, or a negation word, add a leaf node to the "ROOT" with the corresponding tag. For example, in the tree in Figure 1 we add tag ||T|| (target) for "@Fernando", add tag "NOT" for the token "n't", add tag "EXC" for the exclamation mark at the end of the sentence and add ||P|| for the emoticon representing a positive mood. b) if the token is a stop word, we simply add the subtree " (STOP ('stop-word'))" to "ROOT". For instance, we add a subtree corresponding to each of the stop words: this is, and for. c) if the token is an English language word, we map the word to its part-of-speech tag, calculate the prior polarity of the word using the procedure described in section 5 and add the subtree (EW ('POS' 'word' 'prior polarity')) to the "ROOT". For example, we add the subtree (EW (JJ great POS)) for the word great.

Chapter 3: Phases of Sentiment Analysis.

Firstly, we made a twitter developer account .It took nearly a week for them to authorize our account,after that we applied for the access key and token .These keys and tokens would be used by the twitter authorities to authorize the person who is fetching data from them.

Then comes the phase of Pre-processing of the datasets There is a lot of data expressed in different manners by the various users within the tweets. There are two classes in which the complete dataset of the tweets utilized is divided within this study. They are the negative and the positive polarity. Due to this categorization of the data, it becomes very easy to observe the impact of the features present within the overall data through this method. There is a huge susceptibility related to the inconsistency and redundancy related to the polarity of raw data available here. There are various key points followed throughout this process which is given below:

• The elimination of all the URLs, hashtags as well as targets are done. It is to be ensured that there are no spelling mistakes and the sequence of the repeated characters is also to be taken care of.

• The emoticons present within the data are to be replaced with the relative sentiments.

• The various punctuation, symbols as well as numbers are to be eliminated. • The stop words present within the data are to be removed.

• All acronyms are to be expanded.

• The non-English tweets are to be eliminated.

The feature extraction process is utilized to extract these properties from the dataset. The positive or negative polarity of a particular sentence is also done with the help of these extractions achieved. This process helps in determining the opinions of the various

individuals by also using the different models such as unigram, bigram within the process. For the purpose of processing text or documents, the representation of various key features is done within the machine learning processes.

Within the classification tasks, these features are utilized as feature vectors. Some of them are enlisted below:

1. **Words and Their Frequencies**: As per the frequency counts of unigrams, bigrams and n-gram modes; these models are provided are features within the process. For identifying the feature in a better way, there has been more study proposed in the utilization of the word as compared to its frequency. This method has provided better results.

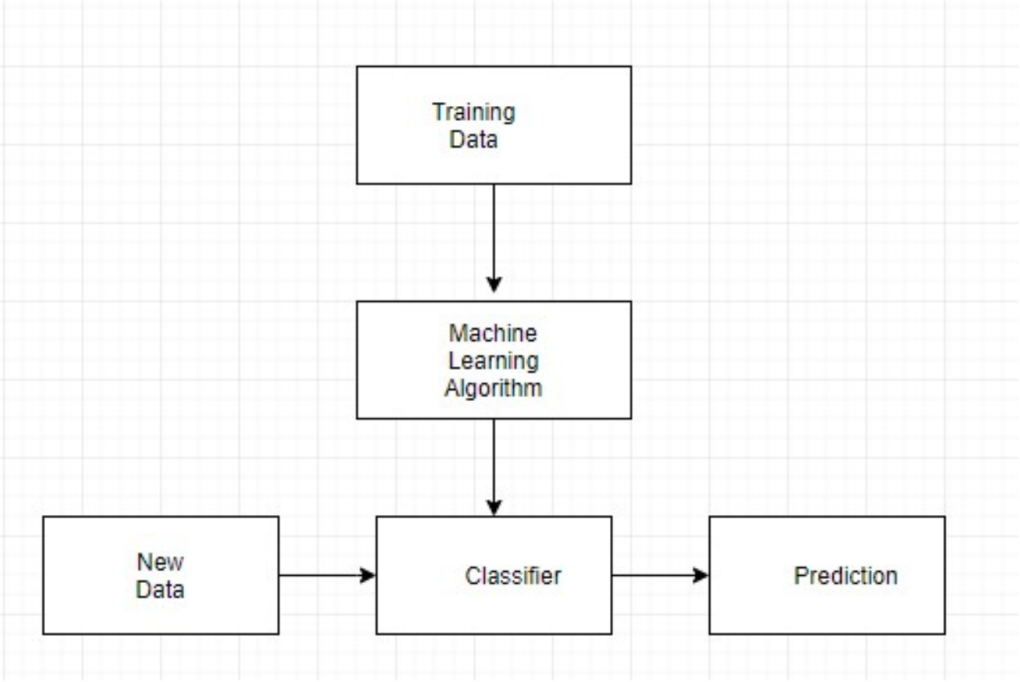

Figure 1: System workflow Diagram

2. **Parts of Speech Tags**: The subjectivity and sentiment within a sentence can be easily analyzed through the various parts of speech such as the adjective, adverbs and groups of verbs or nouns present in it. With the help of parsing or independent trees the syntactic dependency patterns are created here.

3. **Opinion Words and Phrases**: There can be few phrases and idioms involved within the sentences which can be helpful in determining the sentiments by considering them as important features [9].

4. **Position of Terms**: The presence of a term within a certain part of the sentence is a very important factor as it can change the complete meaning of the sentence and provide difference in the sentiment as well.

5. **Negation**: The polarity of the sentence is affected a lot due to the presence of negation within a sentence and so it is a very important and complicated feature.

6. **Syntax**: For learning the subjectivity patterns the various syntactic patterns such as collocations are used for determining the sentiments here.

Training The various classification issues can be solved with the help of supervised learning methods. The future predictions for an unknown data are not required if the classifier is trained well.

In [1]:

```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

#Variables that contains the user credentials to access Twitter API
access_token = "1009323315779944448-0Fcp5egiTp1xeHGRguBdq82mIgV6iK"
access_token_secret = "Eja23RrzKmwNfbLjQ2bGpCbgQowNN6TquekNqrVNI9dfX"
consumer_key = "lnLzHXLDvyxgLFSTpUGERWxNx"
consumer_secret = "L1In42TVUnBZx0oOr4yncAwgZxQ2CIuCdNcMSpXyYwwqfdz16X"


#This is a basic listener that just prints received tweets to stdout.
class StdOutListener(StreamListener):

    def on_data(self, data):
        print (data)
        return True

    def on_error(self, status):
        print (status)


if __name__ == '__main__':
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)
    stream.filter(track=['SACHIN'])
```

Figure 2:Snap of Code.

```
if __name__ == '__main__':
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)
    stream.filter(track=['SACHIN'])
```


```
{"created_at":"Tue Sep 24 15:40:46 +0000 2019","id":1176521867244392448,"id_str":"1176521867244392448","text":"RT @vjkeerthana: Sachin \ud83e\udd70 \n\n#\u0b8e\u0ba9\u0bcd\u0ba9\u0bcb\u0b9f_\u0ba4\u0bb3\u0baa\u0ba4\u0bbf https:\/\/t.co\/ykNjdrDpoP","source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":410496246,"id_str":"410496246","name":"SivaPerumal","screen_name":"SivaCivilBE","location":null,"url":null,"description":"\ud83d\ude18THALAPATHY FAN\ud83d\ude0d","translator_type":"none","protected":false,"verified":false,"followers_count":250,"friends_count":1135,"listed_count":1,"favourites_count":38647,"statuses_count":9362,"created_at":"Sat Nov 12 05:25:12 +0000 2011","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":null,"contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":true,"profile_link_color":"1B95E0","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/1063360981919653889\/HWKBwqVN_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/1063360981919653889\/HWKBwqVN_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/410496246\/1529606013","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"Tue Sep 24 15:36:48 +0000 2019","id":1176520870543577088,"id_str":"1176520870543577088","text":"Sachin \ud83e\udd70 \n\n#\u0b8e\u0ba9\u0bcd\u0ba9\u0bcb\u0b9f_\u0ba4\u0bb3\u0baa\u0ba4\u0bbf https:\/\/t.co\/ykNjdrDpoP","display_text_range":[0,24],"source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_us
```


In [ ]:

Figure 3: Snap of all tweets fetched

CHAPTER 4: DESKTOP APPLICATION

Last phase consists of the development of a desktop application for this project. For GUI I have used Pyqt for developing my Desktop Application PyQt is a set of Python v2 and v3 bindings for The Qt Company's Qt application framework and runs on all platforms supported by Qt including Windows, OS X, Linux, iOS and Android. PyQt5 supports Qt v5. PyQt4 supports Qt v4 and will build against Qt v5. The bindings are implemented as a set of Python modules and contain over 1,000 classes.

PyQt4 and Qt v4 are no longer supported and no new releases will be made. PyQt5 and Qt v5 are strongly recommended for all new development.

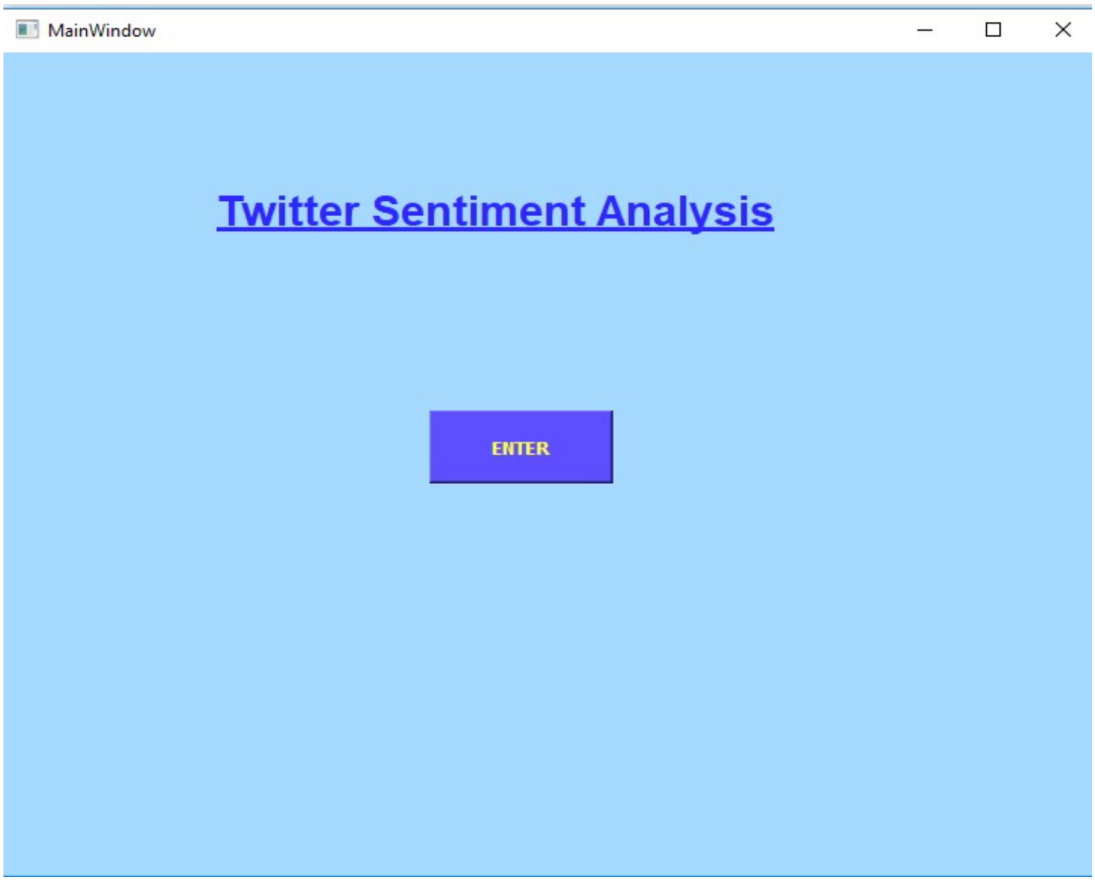

Fig4- This is the homescreen of my desktop application

On pressing the enter another page erupts and asks for the keywords for which we need to fetch

the tweets and do sentiment analysis for the time being we are fetching only about 50 tweets as a lot of tweets would be a tad bit difficult to train on my low specs system, another page looks like.

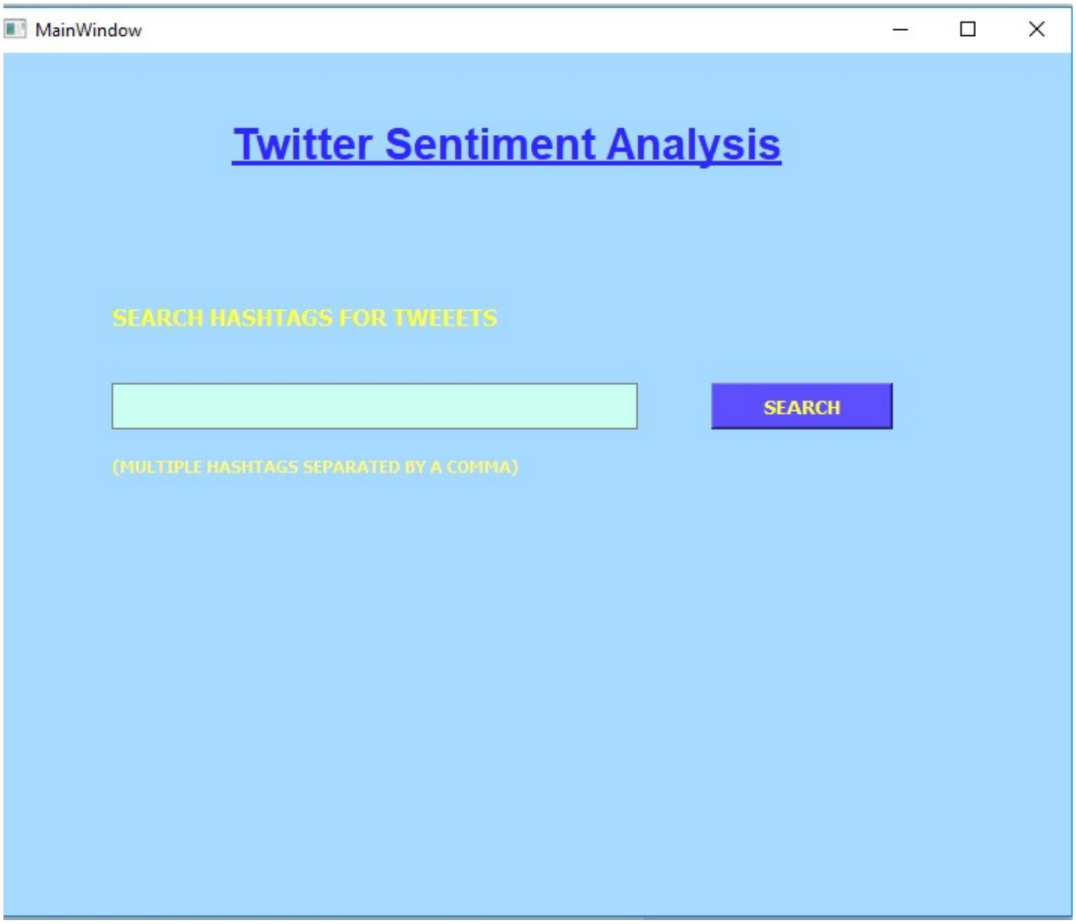

Fig 5 The page which asks for the keywords or hashtags related to which tweets are to be fetched.

On pressing search the application starts collecting the tweets related to entered keywords. The fetched tweets are stored which looks like.

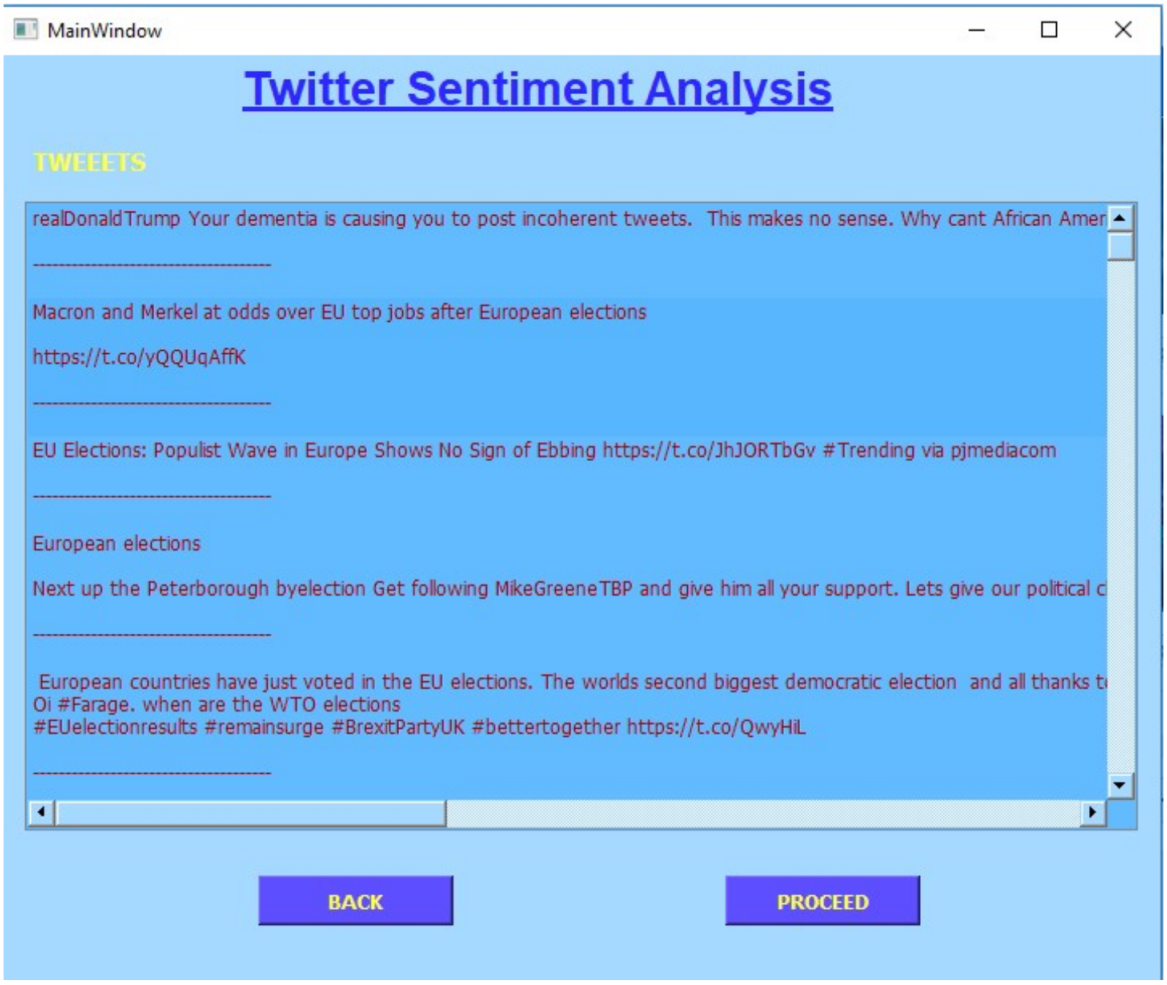

Fig 6 –Tweets being displayed

When the user proceeds the application asks the user whether the user wants to use the default training set or the user wants to upload his own training set in csv format.

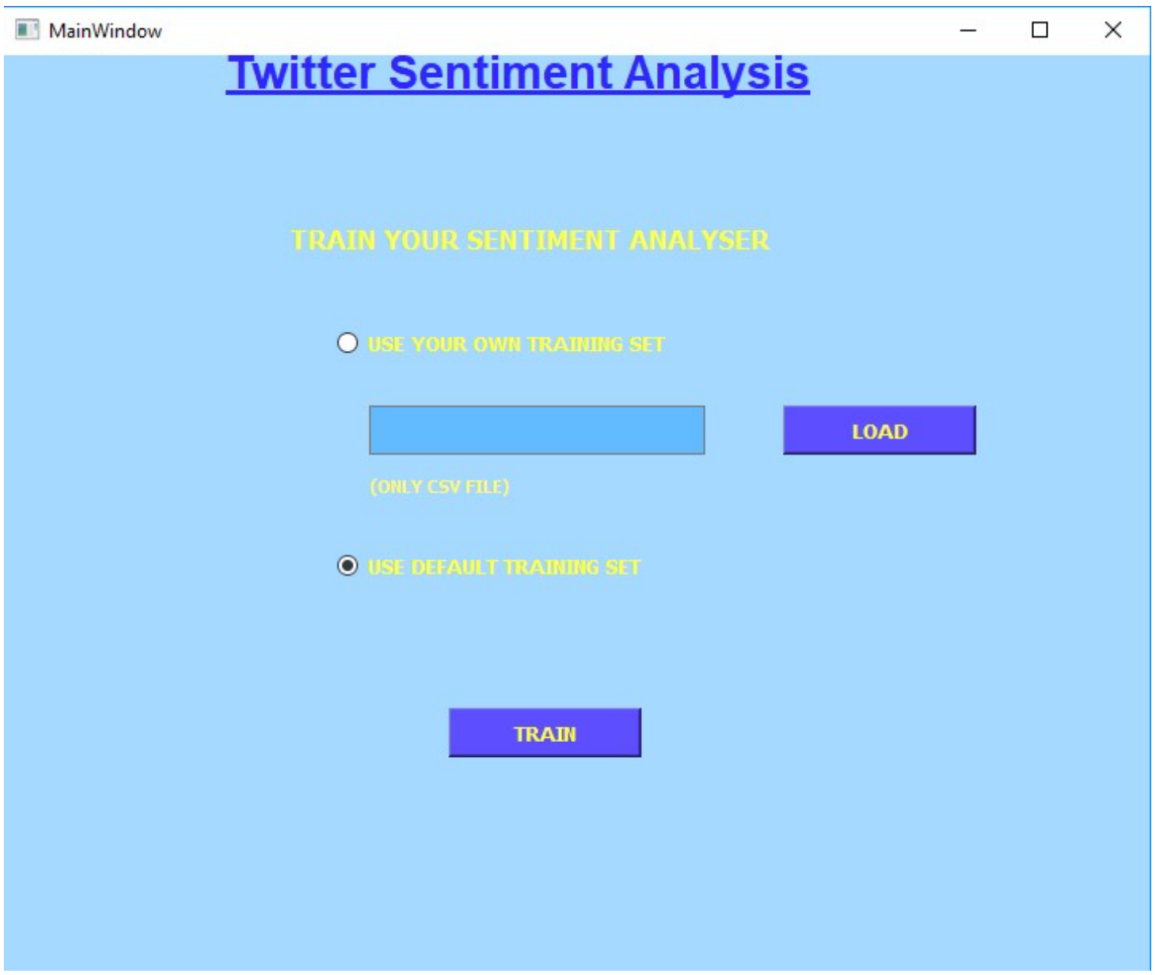

Fig 7-Train using dataset

The application asks the user to train using the default dataset or they want to enter a dataset of their choice.

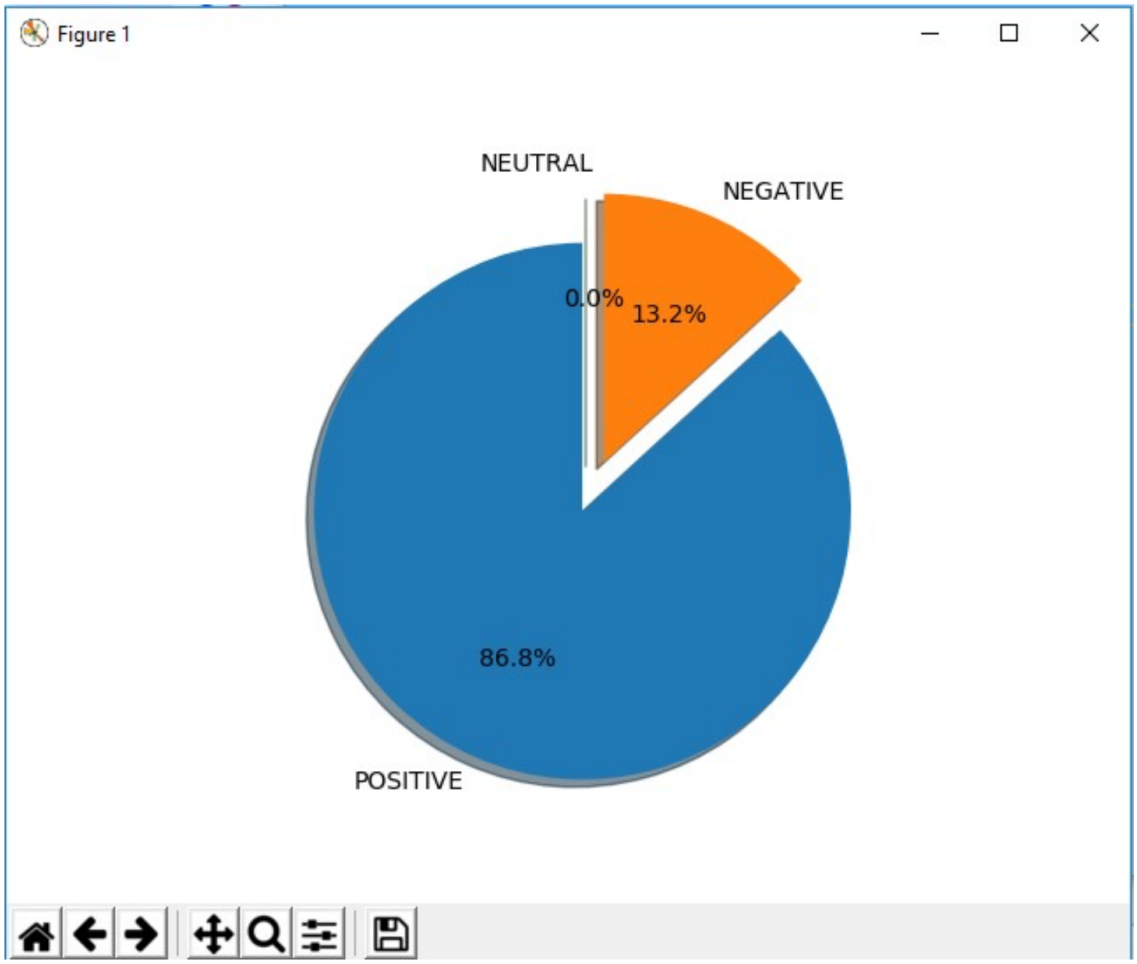

Fig 8 –Result displayed in form of a pie chart

The result is shown in the form of a pie chart , which clearly states the percentage of positive and negative tweets.

Chapter 5: Classification Techniques

There are different types of classifiers that are generally used for text classification which can be also used for social media data sentiment classification.

5.1 NB Classifier

Naive Bayes Classifier makes use of all the features in the feature vector and analyzes them individually as they are equally independent of each other.

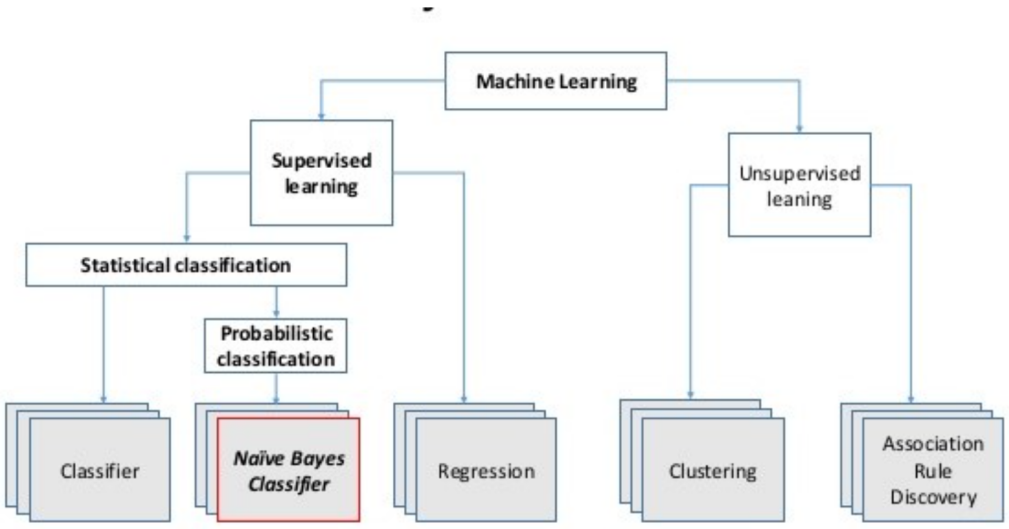

Fig 9 NB Classifier

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

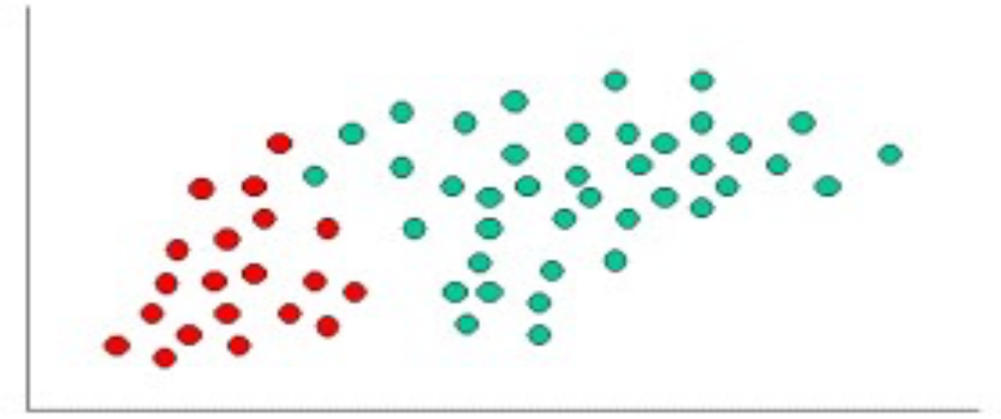

Fig 10 What NB Classifier does.

Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen.

5.2 SVM Classifier

SVM Classifier uses large margin for classification. It separates the tweets using a hyperplane. SVM uses a discriminative function defined as

$$g(X)=wT\varphi(X)+b(2)$$

5.3 Ensemble classifier

Ensemble classifiers can be of different types. They try to make use of the features of all the base classifiers to do the best classification. The base classifiers used here are Naive Bayes, Maximum entropy and SVM. Here an ensemble classifier is generated by voting rule. The classifier will classify based on the output of the majority of classifiers.

Chapter 6 : Conclusions and Future Work

6.1 Conclusion

Social Media is a source of vast unstructured and noisy data sets that can be processed to locate interesting patterns and trends. This application would fetch data from various social platforms and then classify it, which could be used for commercial benefit, educational purposes ,etc. The analysis of data makes it possible for business organizations to keep track of their services and generates opportunities to promote, advertise and improve from time to time.

6.2 Future Work

We can make use of different techniques for classification , which can result in better results. We can also develop a system which can make use of real time data.

We can explore even richer linguistic analysis, for example, parsing, semantic analysis and topic modeling.

REFERENCES

[1] Erik, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. "Statistical approaches to concept-level sentiment analysis." IEEE Intelligent Systems 3 (2013): 6-9

[2]Cheng, Hong, et al. "Discriminative frequent pattern analysis for effective classification." Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.

[3]Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008): 1-135.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892.

[4.]Tan, Songbo, and Jin Zhang. "An empirical study of sentiment analysis for chinese documents.Expert Systems with Applications.