"Design and Implementation of a Generalized Real Time Sentiment Analysis System on Tweeter Data Using Bag of words method through Natural language Processing"

Ashish Soni , Shweta Shah, Harsha Solanki
Malwa Institute of Science and Technology,Indore

Abstract:

Sentiment analysis is the task to classify a short paragraph of text is being positive or negative. Sentiment analysis is a task that becoming recently important for many companies. Because the merchant subscriptions on social media like twitter, facebook and other side get their product reviews. If the company wants to track tweets about their brand to command over the impact on time or many website analyze the comments on their articles. This will helps them to track comments and impact. So the sentiment analysis is an automated system that collects and analyzes the content and generates the desire results. This paper proposed a sentiment analysis system for twitter posts. Proposed system will work on real time tweets. System is also designed in such a way that this can analyze data related to any topic. Python programming language is used to extract tweets form twitter feeds. Proposed system also calculates the level of sentiments. That how much negative or positive are tweets. This paper also presented some real time results analysis.

Keywords: Sentiment analysis, python, tweepy, API, machine learning. Lexicon and NLP.

I. INTRODUCTION

An important part of our information gathering has always been to find out what other people think about any person, incident or issue. More availability of opining resources including online review sites and personal blogs mean one can use technologies to seek out and understand the opinions of others and put it to use. Sentiment analysis deals with the computational treatments of opinion, sentiments and subjectivity in text. Sentiment analysis is classification of the polarity of a given text in the documents, sentence or phrase. Contents of every text documents can be categories weather content are positive, negative or neutral. If text contain has possible thought word with positive meaning sentence that means sentence is positive polarities. Like a sentence "you are such a special human being and I am so honored to be here on this earth with you". Similarly some sentences may give negative impression and some may give neutral. So the computing based sentiment analysis system can calculate these sentiments by computing text through algorithm and generate some short of sentiment value. This can give a clear impression of text. Sentiment analysis system has bunch of applications platform like in stock market [1, 2], political issues [3] or news article [4].

1.1. Objectives

Proposed implementation is based on following objectives.

(i) To design a sentiment analysis system based on machine learning concept.

(ii) To implement a generalize sentiment analysis system for twitter.

(iii) To calculate sentiments on different levels.

2. Background

Sentiment analysis also known as opinion mining, however some researchers mentioned few differences in both [5]. Sentiment analysis or opinion mining can be categorized form different aspects they are as follows:

(i) Technique based
(ii) Text view based
(iii) Analysis level based
(iv) Rating level based

On the view of technical aspects, sentiment analysis can be done through machine learning, lexicon based, rule-based and some time hybrid approach is used. The machine learning approach works on natural language processing (NLP) concept [6,8]. This approach used training data set to calculate sentiment values. Lexicon based technique used frequency of words to calculate sentiment values in text. This technique also called "Bag of words" model. An another simple approach called rule based mining, in this simple positive and negative words are count to calculate the polarities of statements. Some researchers also proposed some hybrid approach in which they proposed mixed approach of machine learning and lexicon based technique.

Another level of categorization can be defined at text view. In this classification sentiment analysis can be done on the structure of the text. It can be either at document level, sentence level or word level. In document level classification sentiment analysis can be implemented document wise. Sentiment polarities are calculated for whole document in a single iteration. Whereas in sentence level, polarities are calculated on the basis of sentence and in word level polarities of each word is calculated separately. Among all three techniques it has been observed that document level classification is mostly used.

There is one more parameter to classify the sentiment analysis that is rating level. According to rating level, it can be classify into global rating and aspect rating. Global rating is overall rating of any product and aspect rating deals in any specific feature of product to rate it.

3. Sentiment Analysis on Social Networking Content

Social sentiment analysis is the use of social media like Twitter and Facebook to understand the wisdom of the crowd [7]. So analyst take Twitter firehouse and thus put certain key words about a certain subject matter and they try and understand from NLP, that what the people are saying about that specific topic. A lots of work analyst try to do is to understand, that how to filter out sarcasm, what is positive and what is negative. Also understand hundreds of mode of icons used by people on Twitter. This is very useful for marketer to understand what the pulses about the product. It also helps to understand where company putting their marketing amount. What is the reaction to advertising campaign? It is ability to look at millions of tweets in couple of time about a given subject. It is like having millions of people focus group. So that sentiment analysis of social media is much more representative and useful.

4. Implementation of Twitter Sentiment Analysis System

Twitter is trouser of sentiments. This section explores the technique through which sentiment can be calculated through Tweets. To do this there are four step procedure.

i. Get the Twitter credential
ii. Access the Twitter data

iii. Calculate Sentiments

iv. Represent results in Desire Format

To run social media campaigns on Twitter with social moves, API key and access token needs to create for each Twitter profile, which is used. API stands for Application Program Interface. API keys are series of code produced by Twitter to allow Twitter profile to poll Twitter feeds. For that a Twitter application needs to be created by going to the http://apps.twitter.com and signing in we have to create a application and need to five information as requested by the dialog boxes. Once after giving all the basic data Twitter application is created. That allow user to move social profile to function. Before this entire process, it is mandatory to have a valid mobile number associated with profile, which is going to be used. To check this we can visit mobile tab in settings of Twitter account. Twitter required mobile number to validate the Twitter profile.

After creating of application Twitter, application manager area will be there. Here tab "Key and Access Token" gives all credential required for Twitter application development. Four main information are required to develop Twitter app development and they are Consumer Key (API Key), Consumer Secret (API Secret), Access Token and Access Token Secret. We can also mange access level through permission tab weather app required read only permission or read/write or read/write and access direct message permission.

After getting this Twitter credential, Twitter data needs to be retrieve. As in this implementation Python programming language is used. Python provides a API called "tweepy" to provide access functionality on Twitter account. So first of all tweepy API must be installing on system. After installation three main classes of tweepy needs to be import in program. Those are stream, OAuthHandler and StreamListener. In this proposed implementation "Sentdex" API is used for sentiment analysis. Instead of Sentdex API one can use any other API like TextBlob or any one can create their own custom function to calculate sentiments of Tweets. Since this implementation is using Sentdes API, implementation required to import urllib. Sentdex API gives different level of sentiments through integer numbers. If number is positive means Tweets contain positive thought, if negative means negative thoughts and if number is zero it means tweet is neutral. Finaly these generated results can be represented in graphical form to understand the overall sentiments of public opinion Figure 1 shows the overall data flow of implementation of sentiment analysis system for Twitter data.

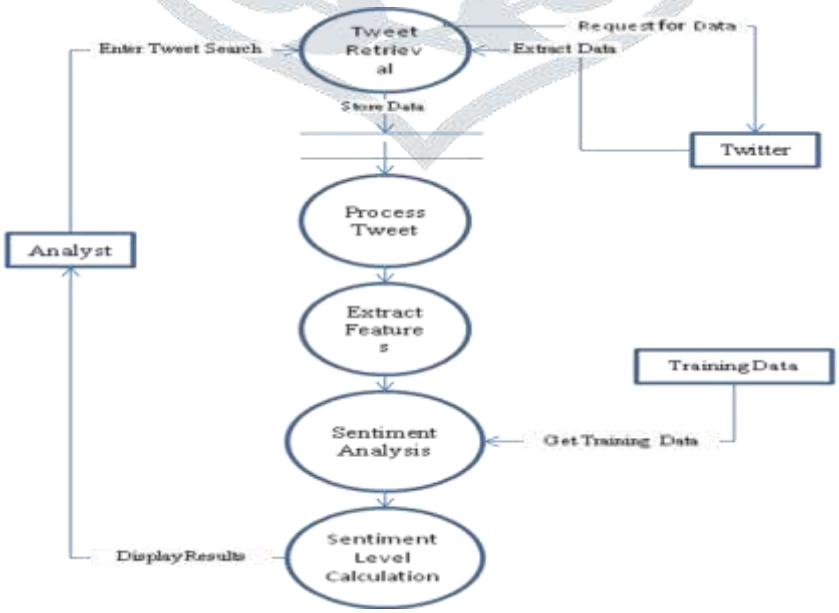

Figure 1: Data Flow Diagram of Twitter Sentiment Analysis System

5. Result Analysis

This section represents some experimental results and their analysis drowns by implemented application. Broadly proposed implementation work can be divided into three major steps.

i. Data extraction from Twitter
ii. Sentiment value calculation
iii. Result display

In this implementation first two steps are done using Python programming language and last step is using C# (.net). It is not mandatory to use same technology and programming language to develop sentiment analysis system. It can be implemented in all major programming languages.

As discussed in previous section, "tweepy" API is used to extract the data form Twitter. Stream Listener of tweepy is used for this. Stream listener is the Twitter streaming API that is used to download Twitter messages in real time. It is also useful for extracting a high volume of tweets, or for creating a live feed using a site stream or user stream. A predefine method called "on_data()" of StreamListener is invoked every time whenever live tweets are received by listener. This method need to override for getting tweet information. This method receives all messages and calls function according to the message type. The default StreamListener can classify most common twitter feeds and send them to appropriately named method. There is a three step process of using streaming API. First, create a class inheriting from StreamListener. Second, using StramListener class creates a stream object. And third, connect to the twitter API using the stream object. A filter on tweets can be apply using "filter" method to stream object. By using this "filter" method, proposed implementation is designed as a generalized sentiment analysis system for any type of subject matter. In the second step sentiment values are calculated by creating a "SentimentAnalysis" method. The "urllib" module of Python is used for sentiment analysis. Proposed implementation used "sentdex" API for sentiment calculation. The "urllib" module allows accessing any website or API through url in user program. This open up as many doors for programs as the internet opens up. Through "urllib" program can access websites, download data, parse data, modify header and do any GET or POST requests. The "sentdex" API calculate sentiments using machine learning and give different level of sentiments in integer numbers. That can be positive, negative or zero according to sentiment calculation.

After getting sentiments value, results has been presented in graphical form to get quick idea about overall sentiment analysis of any subject matter. Implemented program saves all the sentient value in a csv file, as shown in figure 2. This file is used to represent consolidated results in the form of graphs, like pie chart and bar charts. Figure 3 shows the bar char of result extracted and figure 4 shows the pie chart to get clear idea about sentiment analysis.

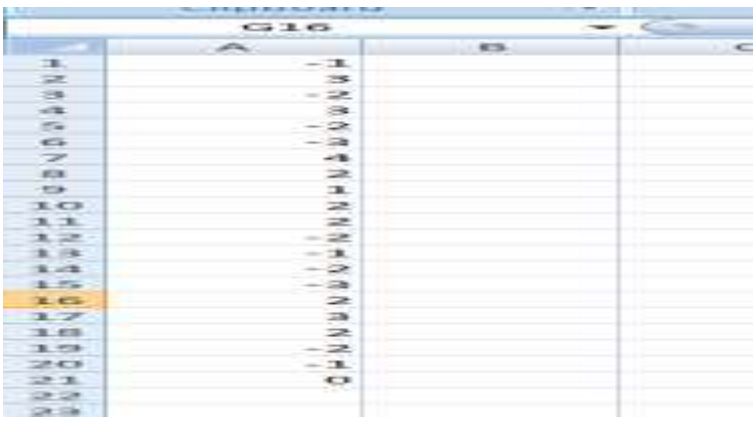

Figure 2: csv File of Sentiment Values

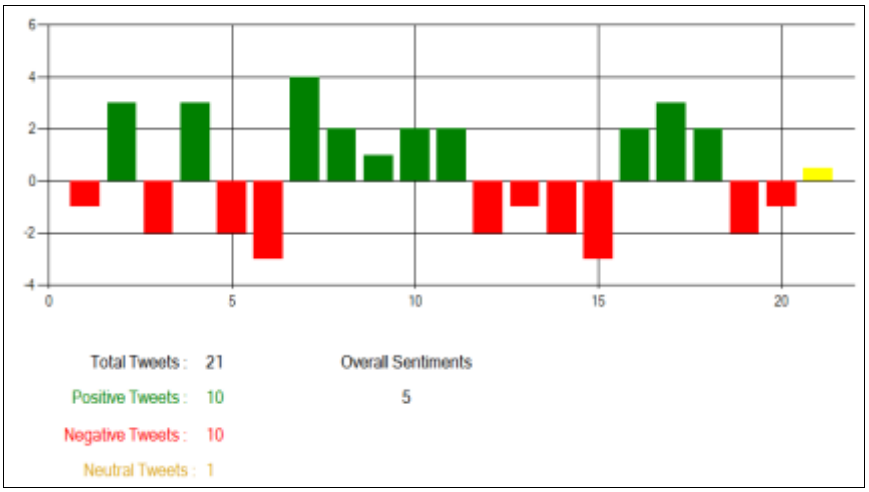

Figure 3: Bar Chart

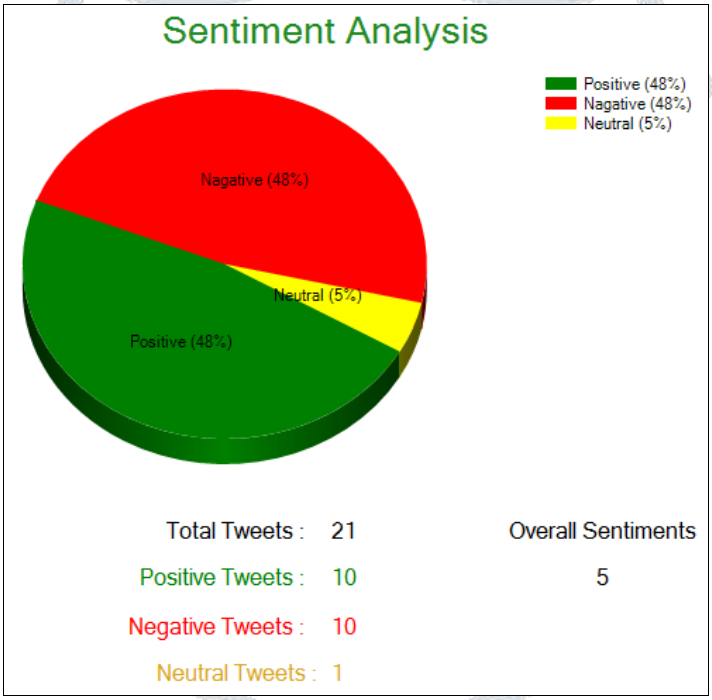

Figure 4: Pie Char

This paper also included the comparison of proposed implementation with Deep learning sentiment analysis system proposed by Oscar Araque et al. in 2017 [7] and Surface classifier which is based on bag or word model.

Table 1 shows the comparison of sentiment analysis system with proposed system. Six parameters are taken as analysis of system. Parameters are as follows.

i. Bag of word model: - This is a base model of sentiment analysis. In which sentiments are calculated based on count of positive and negative words.
ii. Machine learning: - This is also a base model of sentiment analysis system. In this natural language processing concepts is being implemented to develop system.
iii. Real time analysis: - This parameter shows that weather application is working on real time data or not.

iv. Level of Sentiments: - This shows that sentiment analysis system calculating level of sentiments or giving results in positive and negative sentiments only.

v. Stored data analysis: - This shows that weather algorithm are able to calculate sentiments of stored data or not.

S.No.	Parameter	Surface Method Based SA	Deep Learning Based SA	Proposed SA
1	Bag of word model	✓	✗	✗
2	Machine Learning	✗	✓	✓
3	Real time analysis	✗	✗	✓
4	Level of Sentiments	✗	✗	✓
5	Stored data analysis	✓	✓	✗

6. Conclusion and Future Scope

Social networking sites are the open platform to express sentiments and people all over the word doing this every moment of every day. Proposed system is a solution to get abstract sentiments of any topic or subject on twitter. In this machine learning approach for sentiment analysis is used. Proposed implementation gives real time sentiment analysis and also generates level of sentiments. It can be used for any marketing campaign, recommendation system etc.

Proposed implementation is working on real time data only. In future it can be implemented for both real time and stored data. This implementation is focused on twitter data, on the same way this concept can be implement for other social media like facebook or instagram. Proposed work calculates sentiments form text data of tweets. It has been observed that lots of user of on social networking sites uses images and symbols to express their opinion. In future images and symbols can be used with text to do sentiment analysis.

References

[1] Yu Liang-Chih, Wu Jheng-Long, Chang Pei-Chann, Chu Hsuan-Shou, “Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news.” Knowl-Based Syst 2013;41:89–97.

[2] Michael Hagenau, Michael Liebmann, Dirk Neumann, “Automated news reading: stock price prediction based on financial news using context-capturing features.” Decis Supp Syst; 2013.

[3] Maks Isa, Vossen Piek, “A lexicon model for deep sentiment analysis and opinion mining applications.” Decis Support Syst 2012;53:680–8.

[4] Tao Xu, Peng Qinke, Cheng Yinzhao, “Identifying the semantic orientation of terms using S-HAL for sentiment analysis.” Knowl- Based Syst 2012;35:279–89.

[5] Tsytsarau Mikalai, Palpanas Themis. “Survey on mining subjective data on the web”, Data Min Knowl Discov 2012;24:478–514.

[6] Min Hye-Jin, Park Jong C, “Identifying helpful reviews based on customer’s mentions about experiences”, Expert Syst Appl 2012;39:11830–8.

[7] Oscar Araque , Ignacio Corcuera-Platas, J.Fernando Sánchez-Rada and CarlosA. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications", Expert Systems with Applications-Elsevir, 77 (2017), 236–246.

[8] Mr. S. M. Vohra, Prof. J. B. Teraiya, "A Comparative Study of Sentiment Analysis Techniques", Journal of Information, Knowledge and Research in Computer Engineering, Volume–02, Issue– 02, 2017, 313-317.

[9] Mr. Ashish Soni, ,"Student Behavior using R language" ISSN: 2249-9555 Vol.6, No4 IJCSITS,2016