International Journal of Advances in Electronics and Computer Science, ISSN(p): 2394-2835 Volume-4, Issue-11, Nov.-2017
http://iraj.in

SOCIAL MEDIA DATA ANALYSIS: TWITTER SENTIMENTAL ANALYSIS USING R LANGUAGE

[1]SONAL SINGH, [2]SHYAM S CHOUDHARY

[1,2]Department of Computer Science, DR. RR & DR. SR Veltech Technical University
E-mail: [1]sonalsingh2606@gmail.com, [2]738932ssc@gmail.com

Abstract - Social Media is one of the biggest platform for information sharing. Social data Analysis is the analysis of people's interaction in social context. The data analyzed here is collected through social networking site Twitter. Sentimental analysis allows us to grab a hint of inclination of people's views in favor or against of any subject. This paper is about the Sentimental Analysis of tweets on topic BARCELONA TERROR ATTACK. The basic motivation to use this topic is to observe, examine and analyze how people criticize a situation either by expressing their aggression against terrorist or supporting the victims, as we all condemn such inhuman activities in our own way. To analyze sentiments of tweets we are using a powerful statistical tool, R programming. This analysis will be based on classifying views of people in eight different categories of emotion (anger, trust, fear, anticipation, disgust, sadness, joy, surprise) and two different sentiments (namely positive and negative) from the emotion lexicon EmoLex.

Index Terms - Barcelona Terror Attack, Data analysis, Emotions, R Programming, Sentimental Analysis, Social Media, Social Networking, Twitter.

I. INTRODUCTION

With the advent of new era, Technology has got its new and higher pace. This growth has led the change in people's way of expressing their views, sentiments and opinions and also the platforms in which they do so. Now the use of social sites, blogs, online forums have come into play and thus they generate huge amount of data which if analyzed can be useful for business, inventions, worldly issues etc. Data analytics is gaining huge momentum across the world. One of its applications is sentiment analysis which is an area currently under research. Sentiment analysis utilizes the power of data analysis to extract emotions expressed through words used by people across various social platforms, comments, reviews etc.

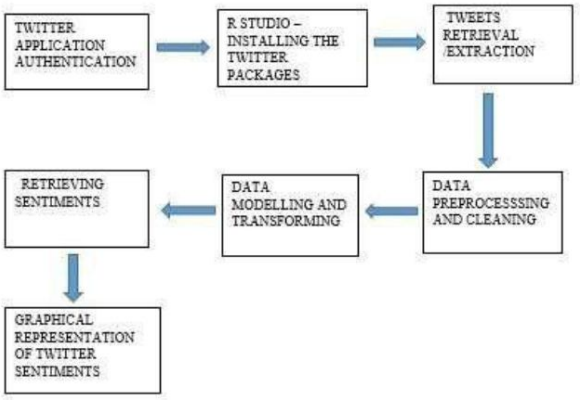

Fig 1 Flow Chart

Twitter is a platform where millions of people express their views and opinions daily on any specific or unspecific topic in the form of tweets. Analyzing unstructured data is in itself a difficult task and extracting useful information from it is a big challenge. For doing so, there is need of powerful tools and technologies which can help to handle millions of tweets and extracting sentiment from them. Although there are various ways possible to do so in this paper R language is used to perform the operation.[1][6]
Sentimental Analysis is a method to explore whether a written text is in positive, negative or neutral state. Basically, it comprises studying the emotions associated with a piece of writing for any topic. [2]

II. SENTIMENTAL ANALYSIS ARCHITECHTURE

The Sentimental Analysis Flow Chart represented in Fig.1 is explained in brief below: -

a) Twitter Application Authentication:
We need to connect to Twitter API using the login credentials of twitter developer application. It is important to authenticate, to connect R Studio with Twitter for extracting the tweets. Once the authentication is completed, proceed to further steps.

b) Installing R Packages:
Installation of required packages is necessary in order to perform the analysis. The package consists of various functions which will be needed in analyzing the sentiments.

c) Extraction of tweets:
It is to collect the data from the tweets on any topic using hash tag "#".

d) Data Preprocessing and Data Cleaning:
The data is cleaned by removing unwanted expressions and words.

e) Data Modeling and Transformations:
After retrieval and cleaning the data is transformed and prepared in a clear structured format to retrieve sentiments.

f) Retrieving Sentiments:
Analysis of sentiments is performed.
g) Graphical Representation:
It is the last step, where the sentiments are plotted and are visualized by graphs and word cloud.

III. MODUS OPERANDI FOR SENTIMENTAL ANALYSIS

R is an open source Programming language and software basically used by statisticians and data miners to study various statistical data's such as poles surveys etc. The complete procedure for analyzing sentiments has been detailed in several series of steps which is described below:

Step 1: Create your account on Twitter Application using Twitter Developers and login to gain access to tweets from R Studio. Download the access "token key" and
"api key" to perform handshake with R console.

Step 2: install and load the packages required for sentimental analysis. Some of the packages used are:
Twitter – provides interface to the web API.
twitter httr – controls request and works with URLs.
ROAuth – used to authenticate to the server of choice.
ggplot2 – for visualization.
wordcloud – to form word clouds and to visualize it
stringr – to use wrappers.
syuzhet – consist of sentiment dictionary and to extract sentiments.
tm – used for text mining.

Step 3: Once the authentication is successfully done tweets are extracted using hash tag.

Step 4: This step involves text mining that is data cleaning, as already mentioned, tweets on Twitter contain many unwanted information, so it is important to clean the text and derive only the useful information.

Step 5: The cleaned data is arranged in data frame and matrix so that desired operation can be performed.

Step 6: The emotions are extracted from the tokenized words and visualization of analysis is done to observe the sentiments.
Thus, the task is done.

IV. THE TWEET

Here is the analysis of tweets on very recent topic BARCELONA TERROR ATTACK. Barcelona is the capital and second most populous city in Catalonia in Spain. More than 120 people killed and 450 injured, the city is facing huge loss. People are expressing their grief and support for Barcelona on Twitter through tweets.

V. DATA EXTRACTION

Data Extraction, in simple words, means collecting data for analysis. Here it refers to collection of tweets. Search API is the official authentication API for Twitter, which returns the tweets that, matches the given string and writes it into the object. Total 4000 tweets were collected on "#BARCELONATERRORATTACK". [3][6]

VI. DATA PREPARATION AND MODELING

The data gathered is not pure. It contains hash tags, urls, abbreviations, punctuation, stopwords, etc. The tweet is needed to be cleaned to perform better analysis. The tm and stringr packages are to be used from the library of R for performing text mining. The tweets are formed into corpus, and then the corpus is invoked. The corpus is cleaned in proper order which is retweets, links, @, punctuations and other symbols which don't express any emotions. Corpus is used to carry out further text mining. [4][5]

Various functions are used to remove unwanted strings from the tweets.[7]

- removePunctuation() is used to eliminate punctuation marks from tweets.
- removeNumbers() is used to eliminate numbers as they don't have any underlying sentiment.
- tolower() converts the entire corpus content into lower case.
- Stopwords removes English words without sentiments like articles, conjunction etc.
- removewords is used to eliminate some specific words.
- stemDocument method reduces a word to its original root word. For example, the word "running" will be changed to its root form run.
- stripWhitespace is used to eliminate extra white spaces.

After performing data cleaning, the data is transformed into required format. Now, a cleaned corpus is transformed into document term matrix. A document terms matrix represent frequency of every word present in the corpus. [5]

VII. FORMATION OF WORDCLOUD

WORDCLOUD is the visual representation of words in the tweets. Here word cloud is used to visualize

essential information from the tweets.

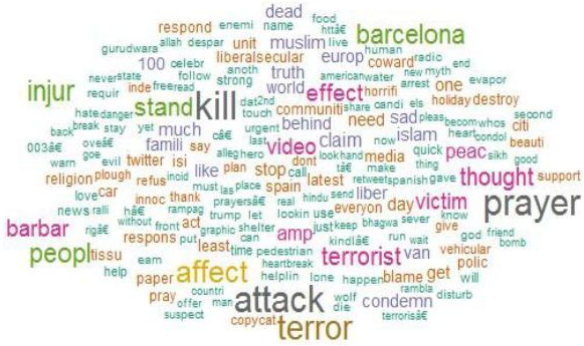

Fig 2: Most frequently used words

The above word cloud shows that most frequently used words in the tweets are kill, prayer, attack, terror, affect, people and so on. The different colors and size of the words indicate their frequency for example kill, prayer and attack have higher frequency than other words. These words represent the immediate response and reaction of the people.[3]

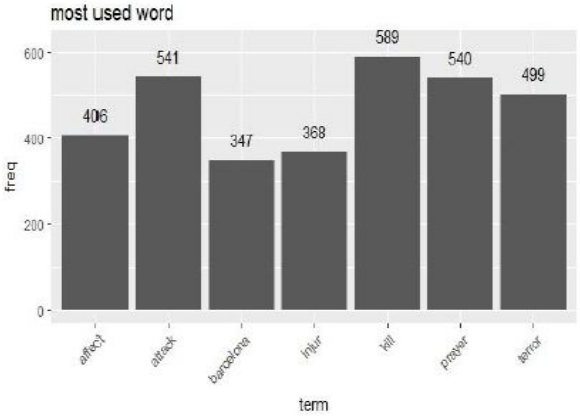

Fig 3: Most used words in the tweets

Fig 3 gives the visualization of the most used words in the tweets. It can be seen that the word kill has the highest frequency followed by words attack and prayer. It's quite evident as it was a terrorist attack and these words were expected to occur more frequently. Words like affect, injury and terror occur quite frequently too.[8]

Fig 4: Hash tag words

Hash tag is used to emphasis on certain issue or a topic of concern. Hash tags are used in tweets to show a subtopic or something which is important in the message the person is conveying. It is kind of highlighting a certain word. As it can be seen from Fig 4 some of the hash tags are

#prayforbarcelona, #Stopterrorism, #globalterrorism, #why muslimskillforislam. These kinds of hash tags highlights opinion of people further and show some deep insights.

Hash tagging certain words or phrase make it easy to search some particular tweets. Fig 4 has #CNN which is a news broadcasting channel so when someone searches for #CNN the tweet regarding Barcelona terror attack will be easily available too. So it can be considered as a successful method for drawing attention.

Fig 5: Word cloud for words without sentiment

Fig 5 it can be observed that the words without sentiments are stand, get, people, barbar, media etc. and so on. These are the helpful words which stand with no sentiments but definitely help in expressing the sentiments of people by forming correct framework of sentences. But here it can be also said that this word could have been part of a sarcastic or humorous comment. At the same time we can see some words like sad, prayer, sadden still present in the non sentiment word cloud which shows that at some level due to some anomaly all sentiments were not extracted correctly which is later corrected by including these words in sentiment word cloud through one of the text mining step. To be able to extract all the words with sentiment text mining must be performed with proper caution.

Fig 6: Word cloud for words with sentiment

Fig 6 shows the words with sentiments are terrorist, victim, thought, kill, terror, truth, condemn, blame etc. From words like terrorist, victim it is easy to judge the situation here. The major emotion for word terror is fear, for kill is anger and fear and for both words the sentiment will be negative. On the contrary for words like bless, safe the emotion will be trust, joy and the sentiment will be positive. Some of the like boom can be taken in both negative and positive sentiment.[8][9]

VIII. THE SENTIMENT ANALYSIS

The major aim of this paper was study and analyzes the sentiments of people for the BARCELONA TERROR ATTACK. The analysis will consist of eight emotions and two sentiments positive and negative.
Bar graph representation is used to visualize the various sentiments behind tweets. It is quite evident that the negative bar is highest because of use of words like terror, kill which shows high negative inclination. The next bar is positive indicating the use of optimistic words like pray which were used by people for the victims of the attack. The get_nrc sentiment function from the package syuzhet is used which will compare all the tokenized words with the wordsentinet EmoLex which contain a large number of words with different emotions. If a word matches with the word present in the sentinet then the prelisted emotions for the word will be increased by one. On adding all the values total emotion and sentiment can be calculated.[1][4]

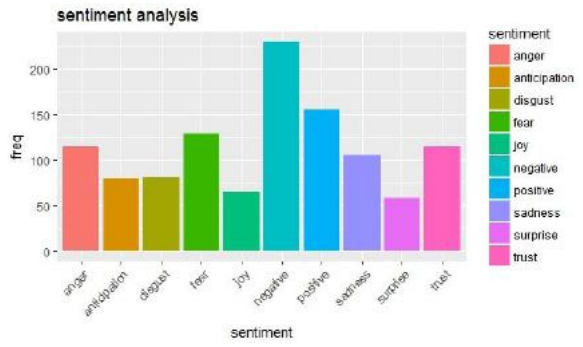

Fig 7: Bar graph showing sentiment analysis

Fig 7 here shows the bar graph representing all the different emotions and the sentiment for our analysis. As expected negative is highest followed by positive and fear. But the bar for surprise is quite high too which may be due to use of words like boom which contain some surprise element too.

IX. LIMITATIONS AND CHALLENGES

- The emotions which are presented through irony, humor sarcasm cannot be treated very well by present sentiment analysis tools.
- Emotions expressed through emoticons cannot be decoded into proper sentiment.
- Use of abbreviations, slang words or some local language derived words stand nowhere in the sentiment dictionary.
- Number of tweets used 4000 which is not much compared to the overall tweets posted by people across the globe.
- Use of mixed language words that is transliterated words makes it difficult for analysis.
- Limitation of collection of words in wordnet which compares the sentiments.
- Only recent tweets maximum of one week and only in text format can be analyzed, while other forms of media and communication result can affect the analysis if taken into considerations.

X. SOLUTION

- Expansion of wordnet for various languages which makes analyzing the sentiments easy.
- Developing tools or algorithm which can determine the context of humor or sarcasm can improve analysis further.
- Tools to convert the transliterated word into one language.
- Advanced mechanism to include other forms of social media into considerations.

CONCLUSION

The above analysis done on the tweets of #BARCELONA TERRORATTACK shows that people came out quite aggressively. The negative sentiment is quite high which was expected since people will definitely condemn the terror activities. Use of negative words definitely influenced our sentiments. But at the same time positive index was quite high which was because of use of some positive words like prayer which definitely raise the bar. In general, people condemned the terrorist activity and prayed for well-being of victims. The analysis shows the application of analytics to determine the sentiment of people and with development in technology and doing further research in text mining and analysis this analysis can be improved more.

REFERENCES

[1] Vivek Sharma, Apoorv Agarwal, “Sentiments Mining and Classification of Music Lyrics using SentiWordNet”, Symposium on Colossal Data Analysis and Networking 2016.
[2] Ilir Keka, Betim Çiço, “Statistical Treatment for Trend Detection and Analyzing of Electrical Load Using Programming Language R”, 4th Mediterranean Conference on Embedded Computing 2015.
[3] Rabia Batool, Asad Masood Khattak, “Precise Tweet Classification and Sentiment Analysis.”
[4] Giuseppe Bruno, “Text Mining and Sentiment Extraction in Central Bank Documents”, 2016 IEEE International Conference on Big Data.

[5] Yerzhan Baiburin, Aliya Nugumanova, "The case study approach to learning Text Mining."
[6] Tran Duc Chung, Rosdiazli Ibrahim, Sabo Miya Hassan, "Fast Approach for Automatic Data Retrieval using R Programming Language", Mediterranean Conference on Embedded Computing 2015.
[7] Raju Ranjan, Sumana Gupta, "Supervised Texture Identification Using Dictionary Based Data Modelling", 978-1-4799-1812-6/14/$31.00 ©2014 IEEE.
[8] Zhang Xiangyu, Li Hong, " A Context-Based Regularization Method for Short-Text Sentiment Analysis"
[9] Youngsub Han, Kwangmi Ko Kim, "Sentiment Analysis on Social Media Using Morphological Sentence Pattern Model", Proceedings of the fifth ACM international conference on Web search and data mining, February 2012.

★★★