Sentiment Analysis of Twitter Hashtag Data

Amanpreet Kaur
School of Computer Science & Engineering,
Lovely Professional University
Phagwara, Punjab, India.

Akshay V,
School of Computer Science & Engineering,
Lovely Professional University
Phagwara, Punjab, India.

Abstract—Nowadays social media being the center of attention for all the reasons, the micro-blogging social networking platform Twitter allows users to share their views and opinions with a maximum length of 140 characters. The way these tweets help the analysts to analyze them to make the business effective is incomparable. This could be seen by the number of frequent words used to analyze the sentiment of tweets which precisely describe the emotions of the people on what they tweet about the particular information. The analysis is all about to visualize, plot a graph and understand the emotion of the tweet. The important objective of the paper work was all about to carry out the twitter data analysis and understand the emotions of the tweet using sentiment analysis. With Twitter developer allows to extract the tweets real-time, there was no need of using the tweets dataset to analyze. Twitter Developer account permits the usage of access tokens to load the tweets real time when called upon. All the necessary packages were loaded beforehand to ensure that there were no mis happenings during the analysis. Package named twitter provides the access to the twitter API. For this visualization, there is a requirement of twitter developer account and R platform to perform the analysis. [1]

Index Terms—Twitter API, twitterR, Sentiment.

I. INTRODUCTION

Social Media is the most buzzing word of today’s world. Everyday people use most of the time in order to escape from this busy world. Each person has his/her opinion or view on some matter which is happening around the globe. It might be a football match, or a new cinema being released or politics which always becomes a new trend. These trending issues or matters in particular make users to express their viewpoints on these matters so that they can have their say on these things. Now when it comes to expressing their views, it all goes down to Twitter platform. This micro-blogging social networking platform creates a space to express the users view in a short, crisp format under 140 characters making it the best platform to express. Since Twitter allows the freedom of tagging and including hashtags to express on the particular matter, it makes it unique from other social networking platform. [4]

Now hashtags are those which are some words which are followed by a # symbol which customizes the tweets which make it easier for the users to select or find the topic. With the help of these hashtags, it is easy for this project to analyze the trending topic and visualize them. R is the language which is specifically designed for the statistical purpose for analyzing and visualizing the data. Text analysis is one which is vastly performed on R platform. This is the main reason to use this platform for the analysis. R has simple yet powerful commands for one’s who have prior programming knowledge. The packages provided by R gives more functionality and are powerful. Now the analysis requires some packages that needs to be pre-loaded in the code such as twitterR, wordcloud, ROAuth which allows to analyze the twitter data easily. Sentiment Analysis is all about predicting the mood of the users on what they are thinking into different categories like positive, negative or neutral. Even though we analyze, it is slightly difficult to predict and analyze precisely under some circumstances. It is to see what people have opinions on the topics of sports, politics or movies and so on. In order to analyze the tweets, there is a necessity of removing the slang words and articles which repeatedly occurs in a sentence. [8] R language is used to fetch the tweets from Twitter using authentication keys and are analyzed since it is most used open source platform running on almost all of the operating systems.

II. LITERATURE SURVEY

R language consists of enormous packages which allows for easier manipulation of data as well as analzyation. Since there is a lot of packages used for the easy analyzation of the twitter data, the analysis runs better. It consists of 11 packages in the code used which are used from the repository CRAN (Comprehensive R Archive Network). Since the sentiment package has been removed in the newer version of the R, a greater number of packages have been used so to perform the operations. The packages used are twitterR, ROAuth, plyr, dplyr, stringr, ggplot2, httr, wordcloud, RCurl, syuzhet, corpus. [5]
1) twitterR- this provides an access to API of Twitter since greater number of functionalities of the API is granted.
2) ROAuth- it provides some medium of interface to OAuth 1.0 specification which authenticates through OAuth to the users to servers.
3) plyr- these are the collection of tools that solves the common problems.
4) dplyr- The grammar of data manipulation, package for making data manipulation easier.
5) stringr- this package allows some set of functions that are designed which makes the working with the strings or tweets extracted easier and simpler.
6) ggplot2- A system for 'declaratively' creating graphics. The ggplot2 is so easy to plot the graph for any type of data as it just requires the data, the labels and a graph would be plotted.
7) httr- httr basically helps in establishing the network related functions so that when using live streaming of data, there should be no errors occurring.
8) wordcloud- Wordcloud provides a text visualization which appears with the greater number of usage of a particular text appears to be bigger and bolder and so on.
9) RCurl- this package provides the functionality of providing an interface with the website related url’s and much more. It usually fetches the URL’s and the requests and process the results which are given by the server.
10) syuzhet- Since the old sentiment is no longer available in the new version of R language, the new version is this. This allows the user to extract the sentimental analysis which are required to perform the operation.
11) corpus- corpus is the mini version of machine learning algorithm used to perform text mining and and the functions are used to create corpora that is the collection of documents from the files or the documents.

III. THE DESIGN OF THE SYSTEM

Creating a system, which takes in continuous stream of raw tweets, processes these tweets to filter out “noise” and get relevant informative tweets. Further these “filtered” tweets to detect and predict evolving “trending” topics in their very early stage. Furthermore, the data analyzed is plotted in a graph to see the emotions of the tweets.

![Figure 1: Data Flow Diagram of the System](image)

IV. METHODOLOGIES OF SENTIMENT ANALYSIS

Step 1: R Packages need to be loaded. So firstly, they should be a check to see whether the packages are already installed. If not, the packages are needed to be installed and loaded by library function which are required for the twitter data analysis and sentiment analysis for that data. All the required packages for the analysis and extraction of twitter data is mentioned above.

Step 2: To connect to the Twitter for the real time tweets, a developer account from Twitter is required. Create a Twitter developer account and create a Twitter application. Application allows the user to connect the R Console to Twitter API.

Step 3: After completing the application process, mention why the account and application is needed. The purpose of usage is the project and other things are filled out. Once created, four Twitter API secret keys are available. They are consumer key, consumer secret key, access token, access secret token. Now setup the connection and allow authorization using the functions of twitteR package.

Step 4: This step needs to ensure that handshake or the authentication is performed or not, that means that the authorization is done and ready to go. Now the tweets from Twitter are ready to be fetched for the analysis.

Step 5: Once the tweets authentication is available, using some functions, we would extract some meaningful tweets for the analysis. These tweets need to be cleaned in order to analyze them in a better way and predict the sentiment of the tweets in the correct path. All the tweets are cleaned by the functions and inspected if any glitches are there.

Step 6: Wordcloud function is used to form the wordcloud to see the visual differences and similarities between the words of the tweets.

Step 7: Finally, the tweets are checked for scores to do the sentiment analysis to classify them into positive, negative or neutral sentiments.

Authentication to Twitter

The interface establishes a connection to the Twitter API using the developer account’s access and consumer keys of the twitter account. When the authorization is done and handshake is done with the Twitter API, the tweets are readily available for the particular topic that the user requires to analyze. The number of tweets required that needs to be extracted could be specified in the function with particular topic that is in current trend. The extracted tweets need to be cleaned in order to exactly predict the sentiment of the tweets. Due to this modern technology and usage of slang words more often, the words in the sentence or the tweets need to be checked which might contain some misspelled words. Without checking, if the analysis is done, we might end up losing the exact prediction of the sentiment of the tweets. [9]

Sentiment Analysis

The feelings, emotions, attitudes or moods of the tweets relate
to the sentiment analysis. It is a natural language processing technique to determine the information of the data. To analyze a person’s opinion, we may find it difficult but with the help of sentiment analysis, it is easier to do it. On applying the sentiment analysis to the tweets, we try to understand the attitude of the tweets from the users for a particular topic and its polarity; positive, negative or neutral. [3]

Stop-words removal
Stop words are those common words like pronouns or articles which are frequently used in a sentence which would make no sense when an analysis is made and in addition to this they hold no extraneous information which seems to be useless. For example ‘in’, ‘an’, ‘a’, ‘the’, ‘he’, ‘she’, ‘they’, ‘on’, etc. On a contrary note, these words would build up to make wordcloud and sentiment analysis less significant in predicting the mood or the sentiment of the tweets by the user.

- Converting to Corpus
- Remove Newline characters
- Remove URL’s
- Remove Unicode Characters

Figure 2: Working flowchart and detailed design of analysis

Positive or Negative tweets
Now the next step is to see whether the tweets are positive or negative or neutral. Positive meaning the tweets has some good words usage which may reflect the positivity of the tweet by the user. If positive, the algorithm which runs for the sentiment analysis gives a value greater or equal to 1 and if the value goes beyond the zero mark that is the negative marking, it reflects that the tweet is a negative tweet. For this sentimental analysis, NLP- Natural Language Processing machine learning algorithm is run behind the system to determine the sentiment of the tweets.

V. ANALYSIS AND VISUALIZATION
Analysis and Visualization is the important factor in determining the output of the emotions or the sentiment of the tweets. Using R packages, and all the necessary functions, the analysis would be done with the visualization or the analysis that needs to be appeared in a picturized format. For example, when you are staring the tweets of so many users, you cannot determine which word has been used most frequently and to find out the words that are used more frequently, we use word cloud. They give a visualization format to make it easier. The graphical representation of the sentiment analysis and the formation of the wordcloud is the main part of this paper.

1) Wordcloud
Word clouds also known as tag clouds or text clouds, is the representation of words in a visualization format. The bigger, the bolder the words, the more frequent that is used in the tweets. Since the tweets contains wordings than numbers, wordcloud is much easier to visualize than any graphical representation.

In the figure 4 below, the word cloud analysis is of the hashtag- amazon. The word amazon looks the bolder, bigger which reflects that this word has been the frequent of all the words used in the extracted tweets and followed by the words cybermonday and protecting which are also used in the greatest number of tweets. This is what wordcloud is all about. Word cloud is the easiest yet powerful way to communicate on the usage of words in a sentence more frequently.

It gives an idea of the mindset of the users who tweet in Twitter. The usage of slang words would hamper the word cloud because they might be used more frequently, and they
would be highlighted at all cost which will make wordcloud inefficient. Without removal of those articles, slang words and stops words make wordcloud more efficient and precise. [2]

Figure 4: Wordcloud

2) Sentiment Analysis
a) Collecting tweets
For this step, the word or the hashtag which needs to be the keyword is given in the function to extract it from the range of tweets. A value is given for the number of tweets that has to be extracted from the account for the particular topic.

b) Pre-process tweets
Pre-processing is nothing but checking the usage of slang words or stop words which create unnecessary decrease in the effectiveness of the sentiment analysis.

c) Positive or Negative tweets
Now the next step is to see whether the tweets are positive or negative or neutral. Positive meaning the tweets has some good words usage which may reflect the positivity of the tweet by the user. If positive, the algorithm which runs for the sentiment analysis gives a value greater or equal to 1 and if the value goes beyond the zero mark that is the negative marking, it reflects that the tweet is a negative tweet. For this sentimental analysis, NLP- Natural Language Processing machine learning algorithm is run behind the system to determine the sentiment of the tweets.

d) Result
The sentiment analysis after cleaning, pre-processing finally is plotted on a histogram using ggplot2 package which gives the graphical visualization of the sentiment of the user’s tweets. In the figure 6, the sentiment for the hashtag-amazon has been plotted which shows positive tweets and very few negative tweets. This particular hashtag is a positive tweet. The users have tweeted positively about the hashtag and the negative words is also reflecting on the tweets analysis. [7]

Figure 6: Sentiment analysis of hashtag

VI. APPLICATIONS OF SENTIMENT ANALYSIS
Most of the company’s use the sentiment analysis for their business purposes to understand their customers to make sure that they hear what their customers say about the company or anything relating to the company.

a) Social Media Monitoring
Nowadays, people use social media like never before. Anything which is happening in the world is actually happening on the social media. Companies make use of social media to understand the clients or customers to improve the marketing and business so to satisfy the needs of the customer. With loads of amount of data, it will be difficult for the normal person to analyze each and everything which leads to years
together of analysis. Sentiment analysis provide the ease within few minutes to understand the mindset of the people.

b) Competitor Monitoring
Every business has its own rival. To compete with the rival the company should also make an attempt to understand how the rival company is performing. One’s loss is the other’s gain, with sentiment analysis the company make full use of the application to excel in their business.

c) Feedbacks
Majority of the company’s reputation depends on feedback given by the customers. Customers feedbacks are gold mine for the company to check their reach to the people.
An instance where customers tweeted about IRCTC about the food issue that was troubling the passengers for a long time and the anger and negative tweets helped the department to overcome the problem. [6]

VII. LIMITATION AND FUTURE SCOPE
The limitations of this paper are that the sentiment analysis is confined only to mentioned hashtags and not any random trending hashtags. This paper has the sentiment analysis which is most common among the analysis part. Sentiment is applied to data analysis. To add further to this, the real time extraction of hashtags and performing more operations would make this overcome the limitations. Adding to this, the future of sentiment could be done automatically with a web interface searching the current trends and expressing the sentiments of the data.

VIII. CONCLUSION
In this paper, the sentiment analysis has been the significance in the business and study of the tweets. Due to increase in analysis of the data, the business, corporates use the sentiment analysis to understand the emotions of the data and the users. Sentiment analysis play a major role in determining the future of the business by understanding the needs of the user. With more advancing technology and compatibility of R language, more packages with better analysis of the data would help the data scientists to predict the incoming of new.

REFERENCES
[1] Statistical Analysis with R Book by John M. Quick
[9] https://dev.twitter.com/overview/api