**Real- Time Twitter Sentiment Analysis on Hurricane Harvey Using R**

**Bhavani Prasad Rao Ejanthkar**

**Introduction**

The growth of several microblogging websites such as Twitter has been tremendous for the past few years where people share the real-time texts/messages/opinions about a variety of topics, products, and other issues in their day-to-day activities. In recent years, it was found to be challenging to understand the variety of information available on these websites and many organizations are increasingly exploring various methods to mine the social media data to analyze the information about what their users think and feel about their services or products in real-time. Earlier, companies used to depend on the customer feedbacks obtained through various channels such as customer feedback forms, customer calls, complaints etc. But these methods are time consuming and not reliable. In recent times, several techniques such as natural language processing (NLP), statistics, or machine learning methods are being used to identify, extract and summarize the useful information by analyzing the sentiment of the text on microblogging sites. One of the current challenges in analyzing the data is to build a technology that can detect and summarize an overall sentiment from the available data.

Sentiment Analysis is the process of analyzing whether a piece of text/information is positive, negative or neutral. It is also known as opinion mining or emotion AI or opinion extraction, extracting the opinion or attitude of a speaker. A general use case of sentiment analysis is to determine how people feel about a product/service/topic based on the real-time data. Although, there are several research approaches available to study how sentiments of the users are expressed in online reviews, news articles, tv shows, product reviews, and few issues on Twitter data namely iPhone product review, US Presidential elections etc. I did not come across analyzing the Twitter data on natural calamities such as Hurricane Harvey using Lexical Analysis to understand the people’s perception during disasters which helps in building an efficient disaster management.

In this project, I explored the Twitter data (about 3000 tweets) during Hurricane Harvey that occurred during August 17, 2017 – September 3, 2017 which was recorded as the costliest tropical cyclone till date in the USA. I performed Lexical Analysis, a sentiment analysis approach using R programming to categorize the tweets into positive, negative and neutral tweets.

**Background**

Sentiment analysis is one of the rapidly growing areas in the field of Natural Language Processing where research focus ranges from a document level classification (Pang and Lee 2008) to determining the polarity of words and phrases in the text (*Agarwal et al., 2009* and *Esuli and Sebastiani 2007*).

Like sentence-level sentiment analysis, Twitter data has been analyzed using character limitations (*Kim and Hovy 2004*). Few recent studies (*Go et al. 2009* and *Pak and Paroubek (2010)* used distant learning to study the sentiment of twitter data where they used positive and negative emoticons to build Naive Bayes and Support Vector Machines (SVM) models.

*Barbosa and Feng (2010)* used polarity prediction among words in their study. They used collected data from three websites as noisy labels to build a training model and used 1000 labelled tweets for tuning and another 1000 for testing purpose. In addition to the use of polarity of words, they introduced the use of syntax features in the tweet such as hashtag, retweet, URL, punctuation and exclamatory marks.

**Methodology**

Our approach to analyze the twitter data using sentiment analysis was based on the following series of methods described in Figure 1.

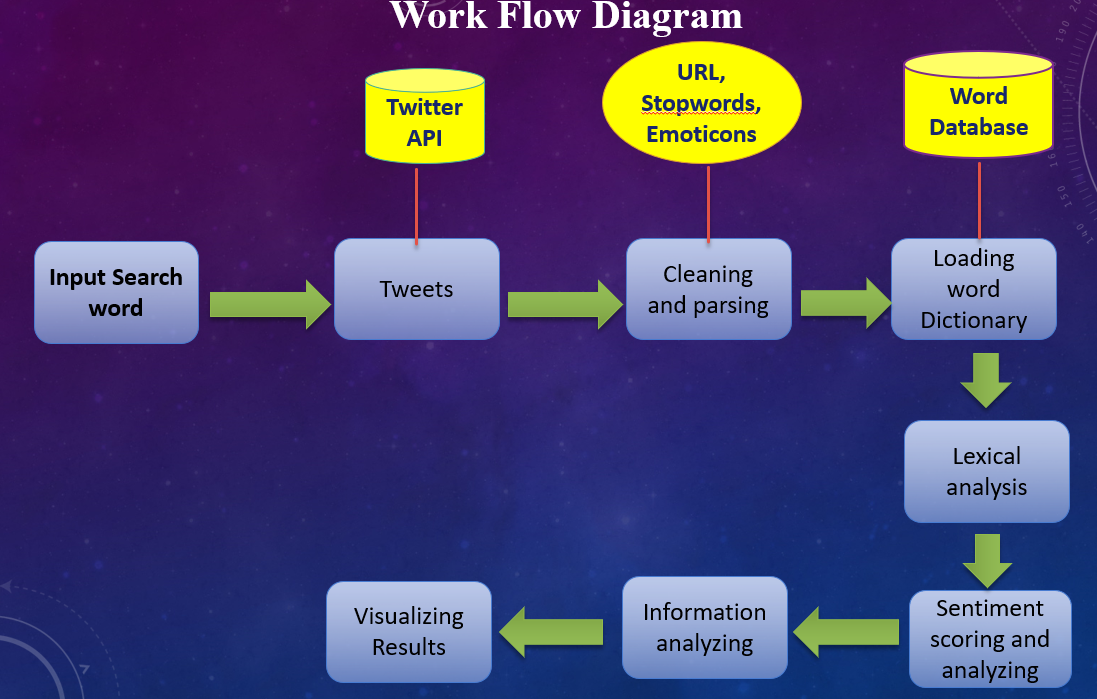


Fig. 1 Work Flow Diagram of our approach

1. Authentication
2. Loading the Reference Word Database
3. Preprocessing of Tweets
4. Lexical Analysis
5. Calculating Sentiment Score of Tweets
6. Information Analysis
7. Visualization
8. Dashboard

**1.Authentication**

The first step in implementing our project is to authenticate with the Twitter API for accessing the tweets directly from the twitter for analysis. To extract the tweets from twitter I need to create a twitter app for authenticating and getting permission to use the tweet data, this process is called authentication. For this step, I need to provide our twitter app credentials to the authentication program to establish a connection. The four credentials are namely consumer key, consumer secrete, access token, access secrete, and along with the certificate for the adhering to the twitter policies. After providing the credentials in the program, that will navigate to the twitter app for the pin as a last step in the authentication this will provide the three-way handshake for our connection with RStudio and Twitter App.



Fig. 2. Twitter Authentication Using Twitter API’s in RStudio

**2.Loading the Reference Word Database**

The is the second step in our project, this will allow us to perform the lexical analysis and calculating the sentiment scores for the tweet. There are positive and negative words file that has 2006 positive words and 4783 negative words as a reference.

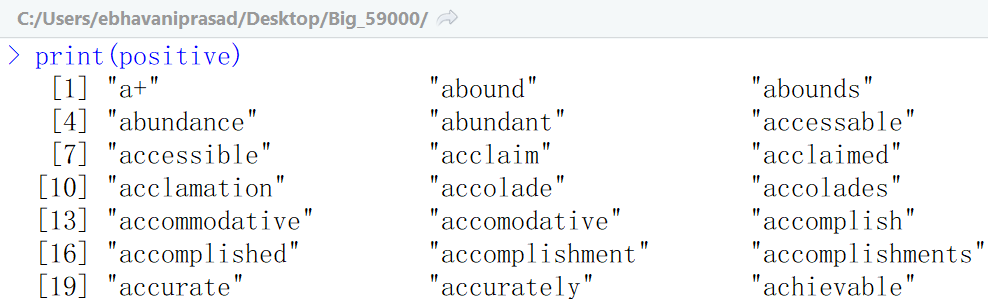


Fig.3. Sample of Positive Reference word list

**3.Extracting and Preprocessing Tweet Data**

In order to extract the tweets from the twitter app, I need to use the function. searchTwitter(“hurricaneharvey”, n= 1000, lang =’en’), this will extract the tweets and stores in RStudio.

Since the extracted tweets will have the all tweet related data in it and it is unstructured format, so in order to make the data clean for our analysis we have to first do a preprocess the tweet data. This involves the removing the emoticons from the tweets. The emoticons are the smiley symbols and other pictures that conveys the meaning, since this are not text and it is not useful in the analysis, so I will remove this from the tweets.

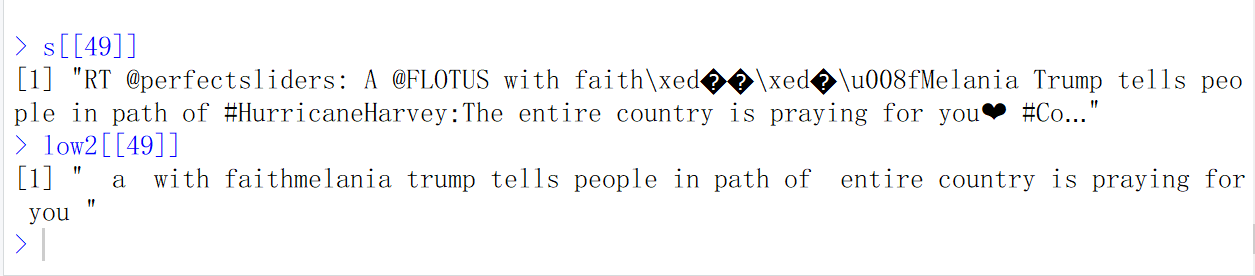


Fig. 4. Removal of emoticons from the Tweets

**Removing the URLs**

In this step, I will remove the URL’s that are present in the tweet. The image below will show the tweet before removing the URL and after removing the URL.

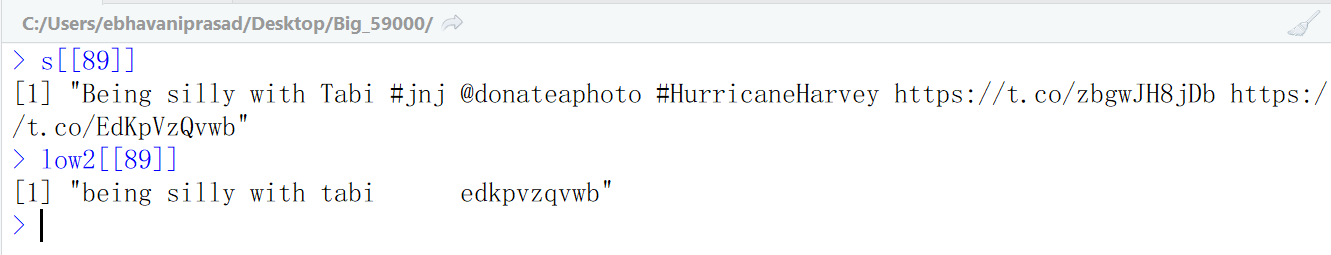


Fig. 5. Removal of URL from the Tweets

**Removing retweets and hashtags from the tweet**

In this step, I have removed the retweets that are in the original tweet. RT in the tweet indicates the ReTweet. It will remove the hashtags that are present in the tweet.

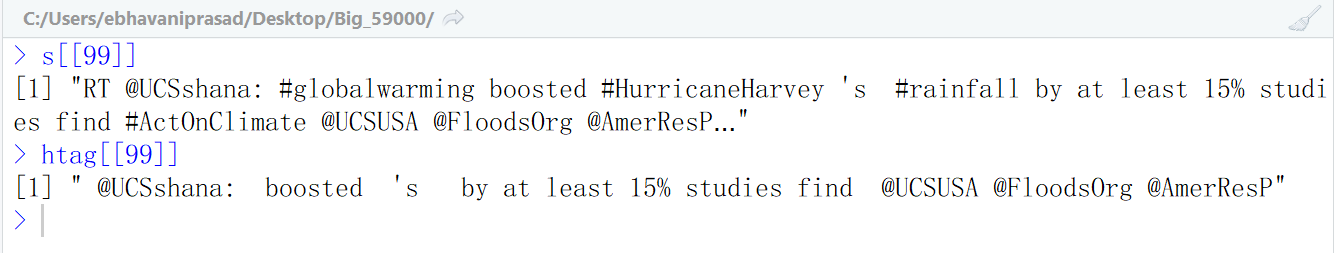


Fig.6. Removal of retweets and hashtags from the Tweets

The hashtags like #globalwarming, #HurricaneHarvey, #rainfall is removed from the tweets.

**Removing Mentions in the tweet**

In the tweet below there are some text that @UCSshana: @FloodsOrg @AmerResP are called mentions that are there in the tweet and this mention are quite useless for analyzing the sentiment, so this are removed from the tweet.

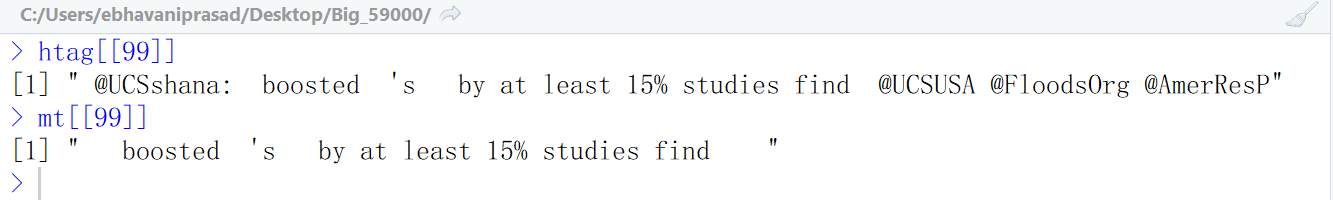
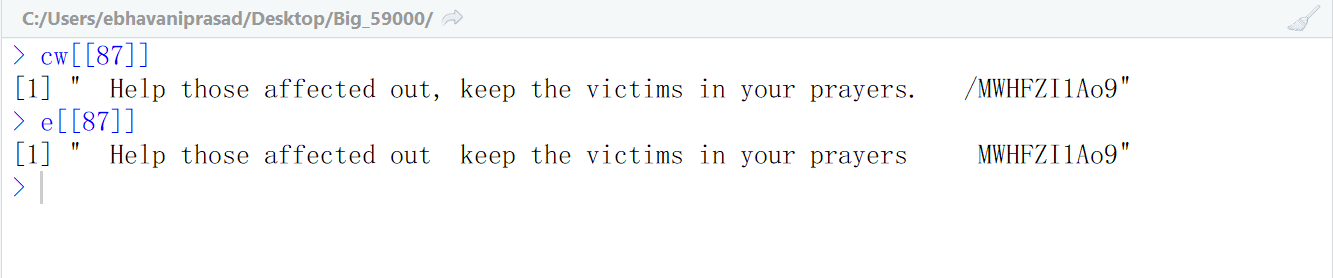


Fig.7. Removal of mentions from the Tweets

**Removing the Punctuations**

Since the tweet are text so it will contain some of the punctuations from the English grammar and it is don’t hold any emotions in it, so I had removed the punctuations from the tweet. The punctuations like comma (, ), Periods (.), backslash (/) are removed from the tweet below.

Fig.8. Removal of punctuation marks from the Tweets

**Removing the Newline characters, Numbers**

In this step, I have removed the Newlines characters like \n, \t, are removed from the tweet and it also removed the numbers are digits that are in the tweet are removed.

**Converting to a lowercase characters**

The tweets are generally informal text that are undercapitalized and overcapitalized words in it. So, in this step I had converted the tweet to a lowercase text that had helped us to comparing the tweets words with the reference positive and negative words.

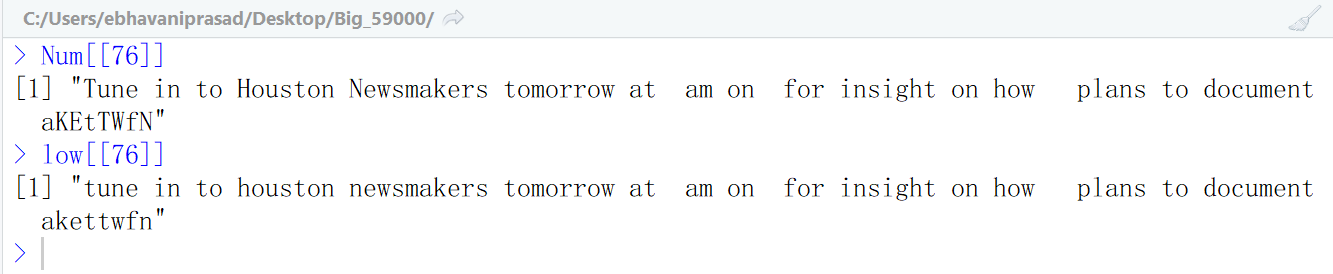


Fig. 9. Converting of tweet to lower case

**4.Lexical analysis**

During Lexical Analysis, I had analyzed the tweet and performed the operations that prepare for the calculating the sentiment score for the tweet.

**Calculating the Sentiment score for the tweets**

This is the main step in the project, that it will perform the analysis and generates the sentiment score for the tweet and gives the overall score and polarity of the tweet.

**Algorithm**

In order to generate the sentiment score for the tweet, I have used a binary classifier for the scoring the tweet. As the sentiment of a people toward a topic can be a positive, negative, or neutral. This essentially means that if the tweet is a positive then the algorithm will

1. The tweets are compared with the reference word database, i.e., the tweet will be compared with the positive, negative words.
2. The sentences are breakdown into words and each of these words are compared with the positive and negative words.
3. If the word from the tweet matches with the reference words database. Then it will give a score of 1 for each match. All the number of matches of the tweet words in each tweet are calculated.
4. The positive matches will have the positive score that indicate the number of the positive sentiment words, similarly for negative and neutral words.
5. If the sentiment in a tweet is neither positive or negative, then the tweet is considered as the neutral words.
6. Finally, it will return an overall sentiment score for each tweet. This is can be calculated by substracting the negative sentiment score from the positive sentiment score. That can have positive, negative, or neutral sentiment.

**Code Snippet**

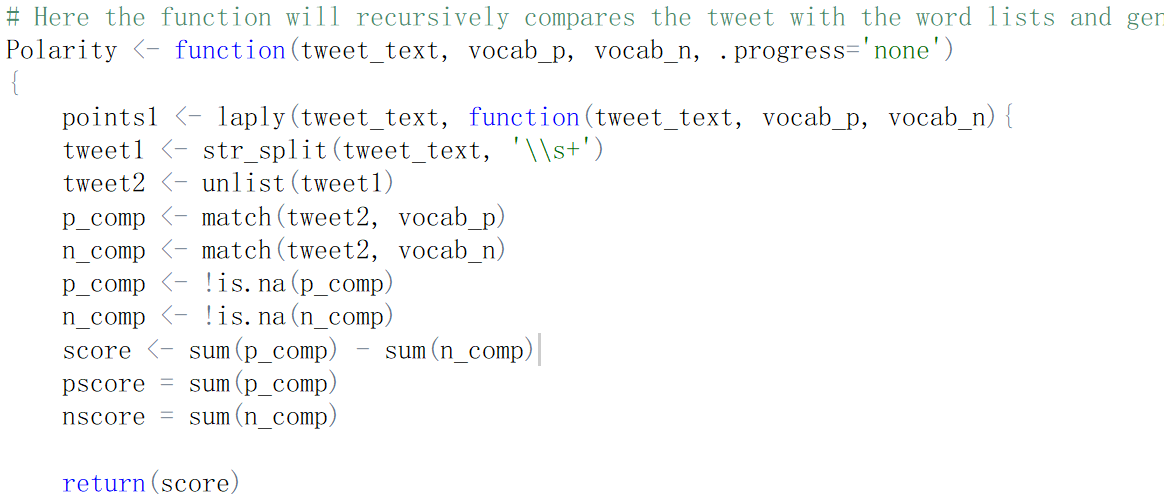


Fig.10. Code Snippet

I implemented this by using three recursive function that calculates the positive sentiment score, negative sentiment score, and total sentiment score (overall tweet score) separately.

**Positive Sentiment Score**

The sentiment score for the positive matches in a tweet are generates and the score is given upon the number of matches with the positive reference word database.

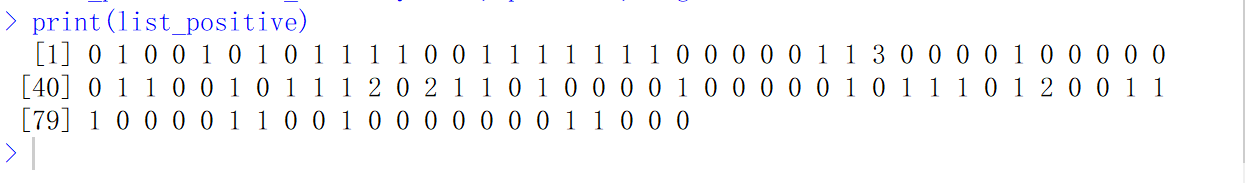


Fig.11. Positive sentiment score for 100 tweets

**Negative Sentiment Score**

The sentiment score for the negative matches in a tweet are calculated and the score is given upon the number of matches with the negative reference word database.

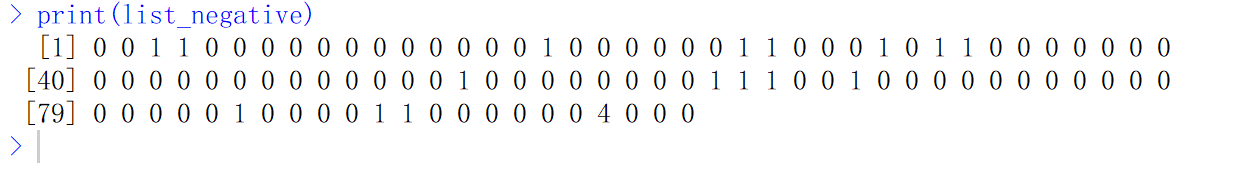


Fig.12. Negative sentiment score for 100 tweets

**Total Sentiment Score**

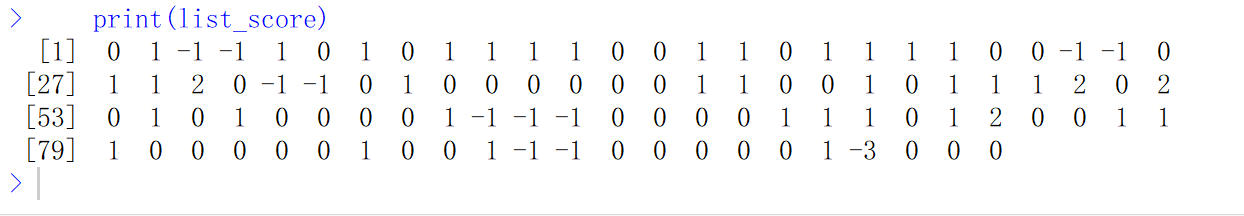
**The total sentiment score for sample of 100 tweets generated. If the score is negative, then the tweet is overall having negative emotion in it.**

Fig.13. Total sentiment score for 100 tweets

**5.Information Analysis**

I have initially extracted the tweet data from the twitter that are basically a raw data an unstructured data and after preprocessing the data I converted it to a structured data that is not much meaningful, and I performed the lexical analysis part on the tweets and calculated the sentiment score then it generates some information of the tweets. So, this next step will completely make the data into an information and gives the insightful information for the generated sentiment score of tweet and gives the facts and inferences of the people’s emotions toward he Hurricane Harvey.

1. Most positive tweet on the Hurricane Harvey: After performing the analysis on the information I have got one of the positive sentiment tweet as.

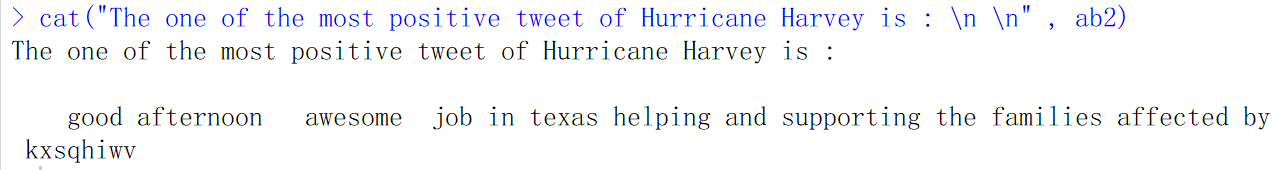


Fig.14. Most positive tweet on the Hurricane Harvey

The tweet above has words like ***awesome job, helping, supporting*** and the overall sentiment of the tweeted person is happy and positive. Since the words in the tweet has the positive meaning in it and it indicates the positivity of the person.

1. Most negative tweet on the Hurricane Harvey: After performing the analysis on the information I have got one of the negative sentiment tweet as

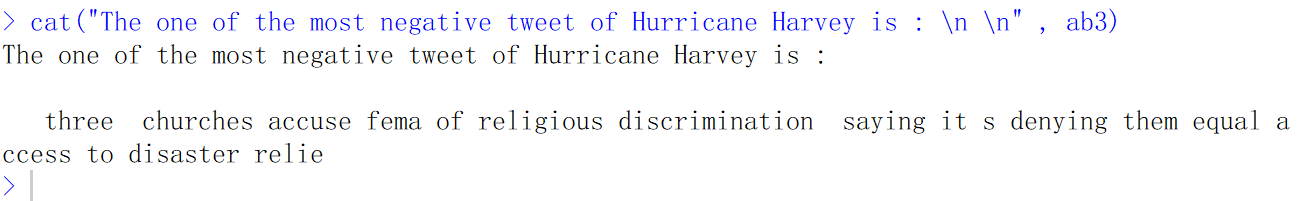


Fig.15. Most negative tweet on the Hurricane Harvey

The tweet above has words like ***discrimination, accuse, denying, disaster*** and the overall sentiment of the tweeted person is unhappy and negative. Since the words in the tweet has the negative meaning in it and this indicates the negativity of the person.

**Overall Score count for the tweet**

I have calculated the sentiment score count for the all tweets and it gives the polarity of the tweet.

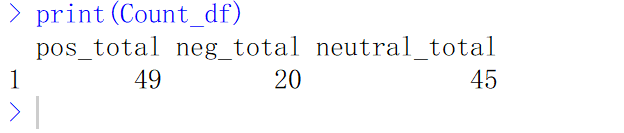


Fig.16. Overall score count for 100 tweets

**People’s Sentiment towards the Hurricane Harvey in Percentage**

After calculating the sentiment score and analyzing the people’s tweets and I got that the people are overall shows positive sentiment. I have got a majority of the people tweet have positive emotions in it.

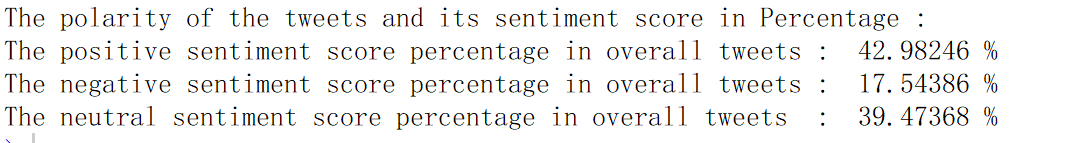
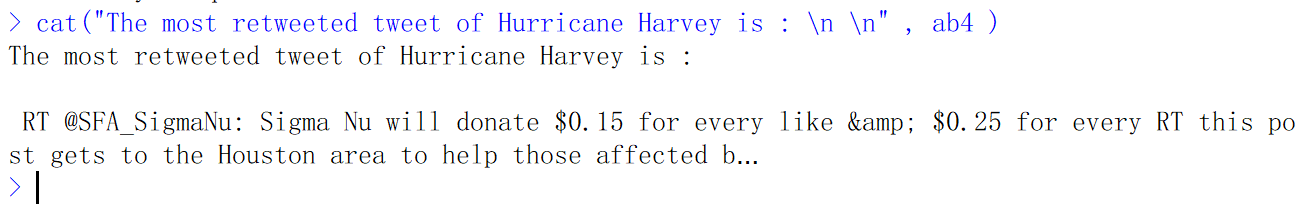


Fig.17. People’s Sentiment towards the Hurricane Harvey in Percentage for 1500 tweets

**Highest retweeted tweet on Hurricane Harvey**

After analyzing the retweet information, I have got a most number of times retweeted tweet as the

Fig.18. Highest retweeted tweet

This tweet was retweeted for 4318 times by the people. This is the most popular tweet on the Hurricane Harvey.

**Person retweeted most number of times on Hurricane Harvey**

After analyzing the retweet information, I have got a most number of times retweeted person as the

Beahavila as **ScreenName**

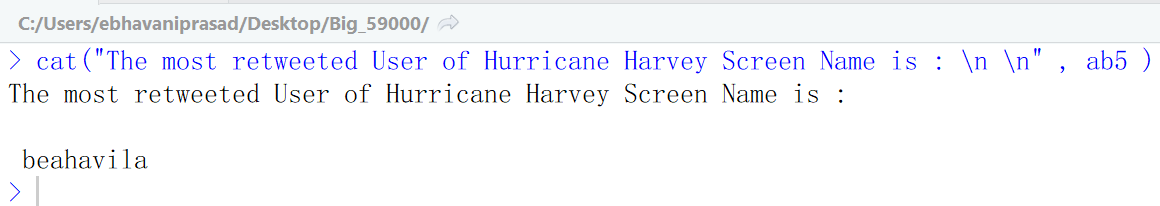


Fig.19. Person retweeted most number of times

I am attaching an excel file that have the sentiment scores and Percentages, Sentiment Score Count for all tweets. I tested the project with three datasets of sizes 100, 500, 1000, 3000, 10000 number of tweets. I have overall, I got a people showing positive sentiment on the Hurricane Harvey.

**6.Visualization Using Tableau**

I used Tableau V10.3 to visualize the analyzed information and to produce statistical results using various bar charts, pie charts and graphs. **Frequency Distribution charts** are used to show the frequency of the count of a positive, negative, neutral sentiment score. P**ie chart** to visualize the percentage of positive, negative, neutral sentiment on Hurricane Harvey.

**7.Dashboard Using Tableau**

Further, a dashboard is created to view the visualization results in tableau. Since to make it more interactive, I have attached a visualization charts, and dashboard in **.twbx** file in tableau visualization folder in the original zipped folder.

**Results**

**After Preprocessing the data**

The initial raw tweets are text that contain the all unwanted information JSON format and it is cleaned and tweet part text is extracted and RT – **Retweet**, @Honeycolony- **mentions**, https:… - **URLs**, #HurricaneHarvey- **hashtags** are removed and returns a plain text in a lowercase.

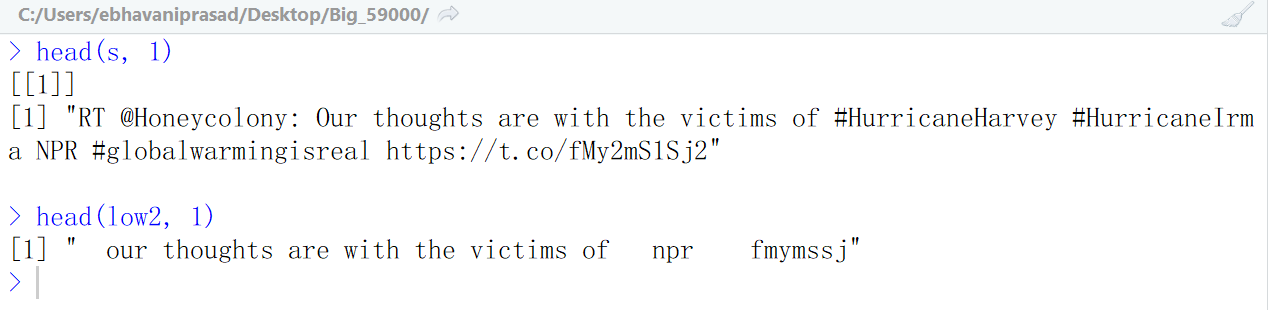


Fig. 20. Tweets before and after preprocessing

**Information Analysis**

Here, I have taken 3000 tweets and calculated the sentiment score and percentage that each score contributes in the overall score. The Result will display the tables with columns tweet (text), total score, positive score, negative score, positive sentiment percentage, negative sentiment percentage.

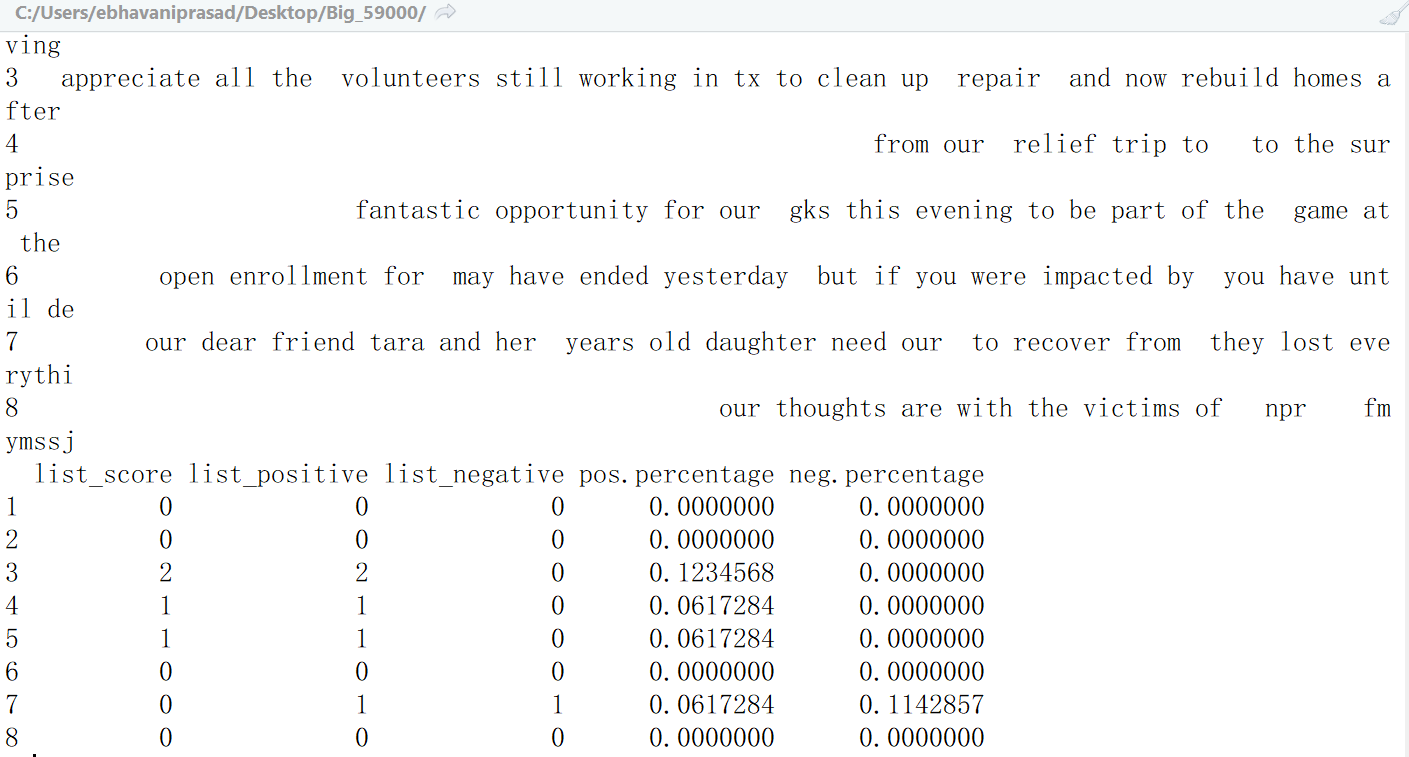


Fig.21. Final outcome after calculating sentiment score (3000 tweets)

**Visualization of analyzed information**

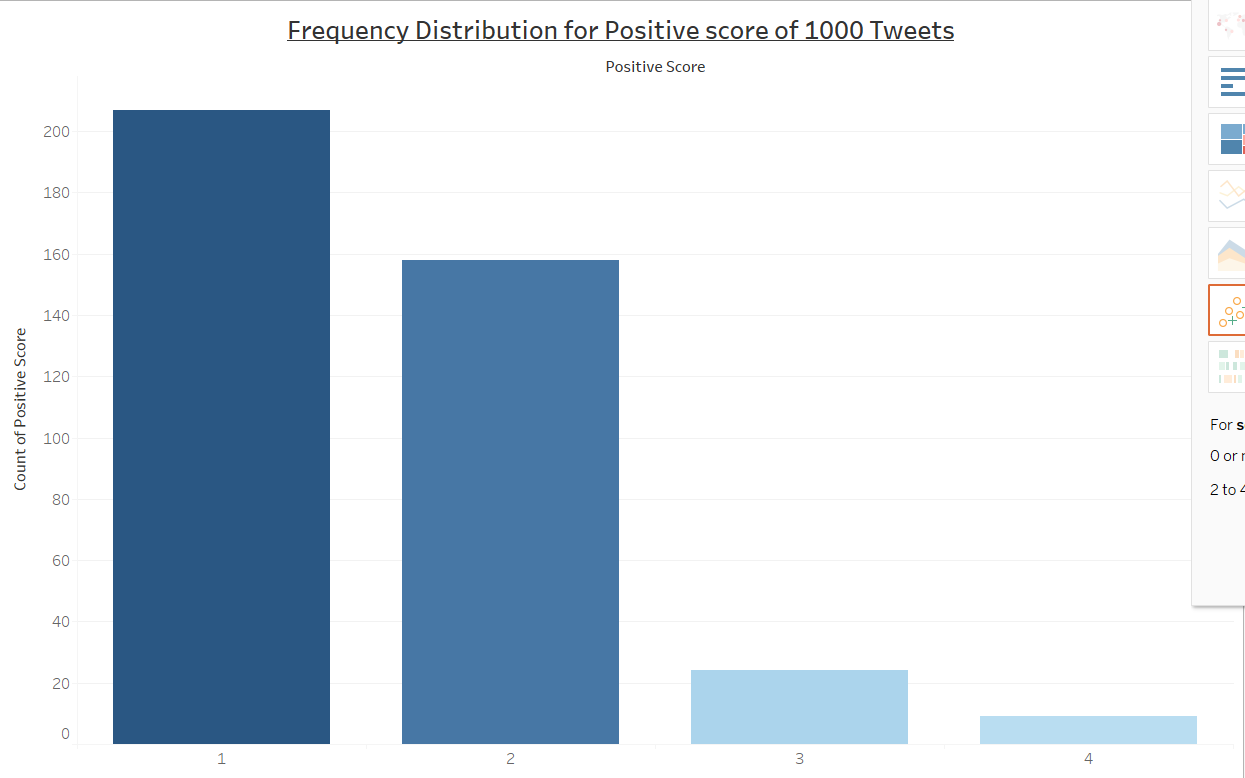


Fig.22. Frequency distribution for positive score of 1000 tweets

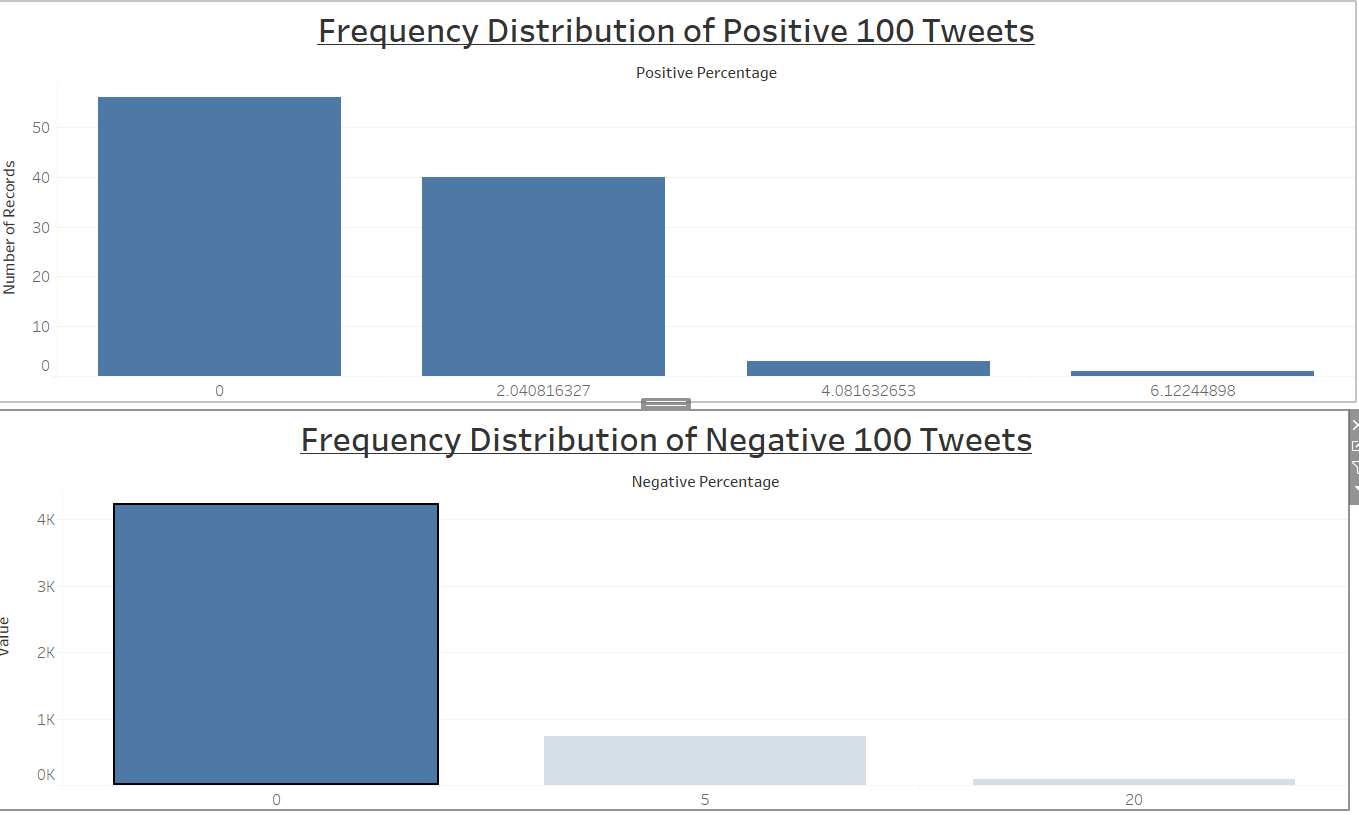


Fig.23. Frequency distribution for positive and negative scores of 100 tweets

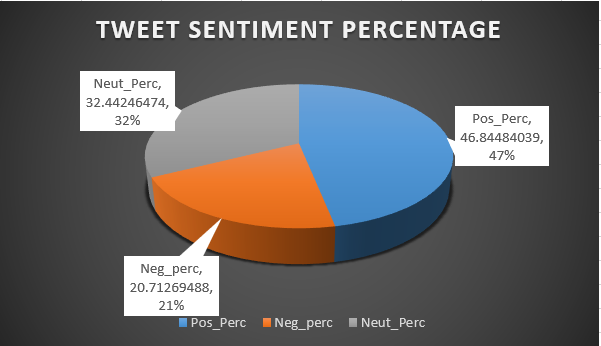


Fig.24. Percentage of sentiment score for 1000 tweets

**Challenges Faced**

Since, I used R Programming, and text data mining, initially I faced challenges implementing the sentiment analysis in R. I have taken a R programming tutorial on the DataCamp and referred YouTube tutorial videos before implementing the project. I have learned Tableau for visualizing and creating an interactive dashboard for various range tweets [100, 500, 1,000, 3,000, 10,000 tweets].

**Helper Code/ Libraries**

I used many libraries in order to use the functions in the RScript that are provided in the R Documentation. For the authentication part of the project, I used the source code form the sources. For calculating the sentiment score for the tweet, I used the idea and some logic part of the code from the source.

**Conclusion**

Our initial approach of Sentiment Analysis on Twitter data (input data is 1500 tweets) using R programming showed a negative sentiment towards Hurricane Harvey and the statistical results on tweets were found to be Positive -24.86%, Negative- 64.52 %, Neutral- 10.61% respectively. The sentiment score sums for positive, negative and neutral tweets were found to be 410, 1056 and 175 respectively.

But when I analyzed the tweets after a week for sentiment determination, surprisingly I got a positive sentiment i.e., Positive- 43%, Negative- 17.5%, Neutral- 39.5%. This is because our analysis is based on the real-time tweets data. The overall sentiment of people has moved from negative to positive which shows a significant variation in our study which concludes a positive sentiment for the analysis I performed on real-time twitter data.

**Future Scope**

With the time constraints, I have implemented simple binary classifier kind of algorithm for calculating the sentiment. In future, there is a scope to implement the following enhancements

1. To implement the project using the Supervised Machine Learning – linear classifier’s Support Vector Machine algorithm or using Probabilistic Classifier’s Naïve Bayes algorithm for accurately calculating the sentiment of the people’s tweets.
2. To analyze the information using the Pig Latin Script instead of MapReduce by computing over Amazon EMR cluster and EC2 instances.
3. I aspired to use MongoDB for storing the data and process the data. Since this is a Document database, it works better with the huge JSON- files with schemas.

**References**

1. Sentiment Analysis https://en.wikipedia.org/wiki/Sentiment analysis
2. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.
3. Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836.
4. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), 12.
5. Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012, December). Opinion mining and sentiment analysis on a twitter data stream. In Advances in ICT for emerging regions (ICTer), 2012 International Conference on (pp. 182-188). IEEE.
6. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics.
7. Esuli, A., & Sebastiani, F. (2007). SentiWordNet: a high-coverage lexical resource for opinion mining. Evaluation, 1-26.
8. Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.
9. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), 12.
10. Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010).
11. Barbosa, L., & Feng, J. (2010, August). Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 36-44). Association for Computational Linguistics.
12. H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 416-419
13. M. Trupthi, S. Pabboju and G. Narasimha, "Sentiment Analysis on Twitter Using Streaming API," 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, 2017, pp. 915-919.
14. P. Ray and A. Chakrabarti, "Twitter sentiment analysis for product review using lexicon method," 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), Pune, 2017, pp. 211-216.
15. M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-5.
16. A. P. Jain and V. D. Katkar, "Sentiments analysis of Twitter data using data mining," 2015 International Conference on Information Processing (ICIP), Pune, 2015, pp. 807-810.
17. RStudio Documentation. n.d.(2017). Rstudio software. Retrieved from <https://www.r-project.org/other-docs.html>
18. Larson B (2016). R: Twitter Sentiment Analysis. Analytics4All. Retrieved from <https://analyzecore.com/2017/02/08/twitter-sentiment-analysis-doc2vec/>