Indian Journal of Science and Technology, *Vol 9(45), DOI: 10.17485/ijst/2016/v9i45/97914, December 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Twitter Streaming and Analysis through R

Vaddadi Vasudha Rani[1] and K. Sandhya Rani[2]

[1]Department of Information Technology, GMRIT, Rajam - 532127, Andhra Pradesh, India; Vasudharani.v@gmrit.org
[2]Department of Computer Science, SPMVV, Tirupati - 517502, Andhra Pradesh, India; sandhyaranikasireddy@yahoo.co.in

Abstract

Objectives: To retrieve tweets from Twitter through Twitter API. The domain chosen for analysis is Make-In-India Dataset. **Methods/Statistical Analysis**: This paper consists of two phases of work: 1. Data Streaming from Twitter, 2. Knowledge Mining through R-Studio. Methods used for the two key operations are: 1. Twitter API and 2. Sentiment Analysis through R. Twitter application is created to request for connection with the Twitter database. Once connection establishes, authentication keys are generated. Providing the search key "Make-In-India" and number of keys required, a file with .df (data frame) is generated with the tweets and is converted into .CSV (Comma Separated Values) file which is suitable Analysis. Sentiment Analysis[1] is also called Opinion mining talks about retrieving facts from the tweets such as how many people supporting Make-In-India (or) how many are negative with the scheme (or) how many are neutral with it. For this process, a negative words file and a positive words file is taken for comparison with the tweet data to calculate positive score and negative score of the tweet. The difference of these scores gives us with the final score of the tweet. **Findings:** The number of tweets identified as positive (or) negative (or) neutral so that the status of Make-In-India can be visualised in a graph. Firstly, the extraction of Tweets is from Twitter through R-Studio Environment About "Make-In-India". Secondly, we parse the extracted raw tweets using R according to the types and store in .CSV format in R database. Scores are calculated for all the tweets and stored in a file. In the third, we perform visual analysis from the stored data using R statistical software to conclude the impact of the program. **Application/Improvements:** Application of the methodology is to get findings[2] from the public opinions which are available from Twitter tweets on a particular government issue, political parties and medical status around the country. Also it is useful to assess the popularity of the political leader and the program. Decision making is possible through sentiment analysis of user tweets.

Keywords: Big Data, Data Analytics, Make-In-India Data Set, Streaming, R-Studio, Tweets, Twitter API

1. Introduction

Big data encompasses social networking web sites including Twitter, Facebook and LinkedIn. All these data sources are having many Applications in the real world[3]. It consists of both structured and unstructured data of text, pictures and videos from which mining of knowledge regarding the latest workings of governments can be understood. It has been initially characterized by three V's but now through five V's. Recent days this is one of the most upcoming ones in the headlines; it is also fast-becoming a genuine force in originating planned insight and business intelligence which leads to strategic decisions from social media. Two additional dimensions of big data are variability and complexity i.e. varying data loads and to extract meaningful value from big data, we need ideal processing power and analytics capabilities[9].

1.1 Big Data Analytics

Processing large sets of data, either descriptive analytics to understand the inter relations or predictive analytics to discover new patterns of the current trends in the market. These data patterns are converted into actionable knowledge that can be used for decision making. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most

*Author for correspondence

important to the business and future business decisions. To analyse such a large volume of data, big data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, forcasting and data optimization. Big Data tools like Clojure, Scala, Python, Hadoop and Java for NLP and Text Mining and R, MAT can be used data analytics. For example considering new government initiative "Make-In-India". This paper analyses MII Dataset to make a decision on the future stabilization of the project.

1.2 Twitter

The messages that are created through Twitter are called as Tweets. This data is available as public data. So it can be taken as raw data mainly for opinion extraction, for analyzing customer satisfaction and for rating different government schemes and finally to do sentiment analysis. Nowadays online purchases are happening based on the opinions posted by the people regarding different products. So, marketers and purchase departments need to spend more time in analysis of user opinion for its positiveness. Here Bayesian classification method is used for classification of tweets into positive, negative and neutral. Using the sentiment analysis the customer can know the quality about the product or services before making a purchase[5]. The company can use sentiment analysis to know the feedback of customers about their products, so that they can analyze customer satisfaction and according to that they can improve their product quality. Sentiment analysis has become one of popular research area in computational field, because of the explosion of sentiment information from social web sites, online forums and blogs as in paper.

1.3 Make-in-India

This initiative is to make India a Global Manufacturing Hub giving a call to business leaders, potential partners and investors throughout the world. It should be shown as a credible initiative to its upcoming depositors. There is visible momentum, energy and optimism. So to observe the moment of Make-In-India, we have targeted the information from SN web sites like Twitter to do sentiment analysis to see that among the world population, how much positive and negative opinion is there on Make-In-India taking a sample of 200 current tweets about Make-In-India.

1.4 Sentiment Analysis

It is to calculate sentiments of people i.e. opinions about a particular context like product reviews etc. Analysis[3–5] part takes care of knowing the opinions positive, negative or neutral. For example, by knowing how many reviews are positive about a product, customer can take decision to buy or not.

2. Architecture

2.1 Twitter Environment

Twitter API Authentication process is carried out using OAuth package of R. Figure 1 precises the steps involved in usage of OAuth to Access Twitter API. A Twitter application needs to be created to run Twitter API.

- Consumers need to register with Twitter. It continues in providing a key and secret key to consumer which can be used in the application to be authenticated.
- These keys are to be used to create a Twitter link through which Authentication process gets initiated. Verification of users identity is done by Twitter and issues PIN 3 called as verifier. The user needs to offer this pin to the application.
- Next application process uses this PIN to request for an Access Token and Access Secret, exclusive to the user from Twitter API.
- Token and secret key information are cached for further use. It can be accomplished through GetUserAccessKeySecret.

2.2 R- Studio

R-Studio is the environment developed for statistical analysis and a Graphical view of the large data sets.

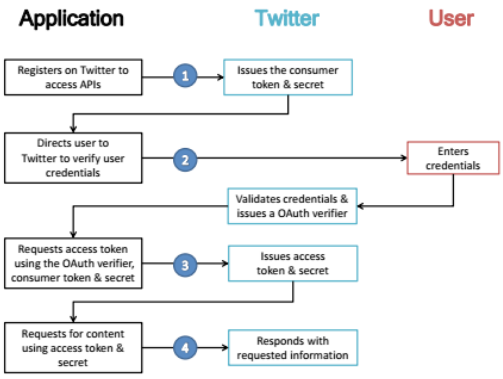


Figure 1. Twitter architecture.

R-Studio is rich in packages, nearly 8000 packages are available.

2.3 R-Packages

R-packages are a collection of R functions which is a compiled code on sample data. These functions are stored under the name of R-Library in its environment. During installation period, by default R installs a set of packages. Remaining packages need to be installed and loaded separately as and when they are required by the specific application.

The below given packages are used in the implementation of this paper.

- **twitteR**: Whose main purpose is to provide an interface to Twitter API.
- **ROAuth**: Allows users to authenticate to the server through OAuth (1.0 specification).
- **plyr:** The main purpose of this package is to manage a big problem by dividing it into manageable pieces, work on it and to put them together.
- **stringr**: Stringr is responsible for all string functions in R environment. It is a simpler and easier to use package. These functions can effectively handle zero length characters data and even for NA's also.
- **ggplot2**: To implement graphics in R. It is for both base and lattice graphs and supports multiple data sources.
- **RColorBrewer**: Used for drawing nice maps shared according to a variable through palettes.
- **Devtools**: The aim of Devtools is to make your life as a package developer easier by providing R functions that simplify many common tasks. R packages are actually.

3. Methodology

For Twitter Analysis, the proposed system has the following steps involved:

- ✓ Creating Twitter Application.
- ✓ Execute Twitter API code through R-Studio.
- ✓ Collecting Twitter data archives.
- ✓ Classifying the Data with R Tool commands.
- ✓ Running R commands for processing the tweets.
- ✓ Establishing R Plotter to view results.

3.1 Creating Twitter Application

Twitter search API is used to access a part of Twitter recent tweets of the past one week. This is a collecting and cleaning tweet data sets used for research work.

To do this Twitter application is to be created. It is shown in Figure 2.

3.2 Execute Twitter API Code through R-Studio

Twitter Search Application Program Interface code has to be executed from R Console. To have the interface to Twitter tweets, connection has to be established to Twitter web site. Then we need to search for our tweets. And save them to CSV file. CSV stands for comma separated Vector. As part of Twitter API process there are many R packages has to be installed first through install command of R. and imported to R through library command.

makeinindia.list<- searchTwitter('makeinindia', n=100,lang="en")

The above command returns the tweets from source for the last one week data about the product i.e Make-In-India for my example, mentioned in the command.

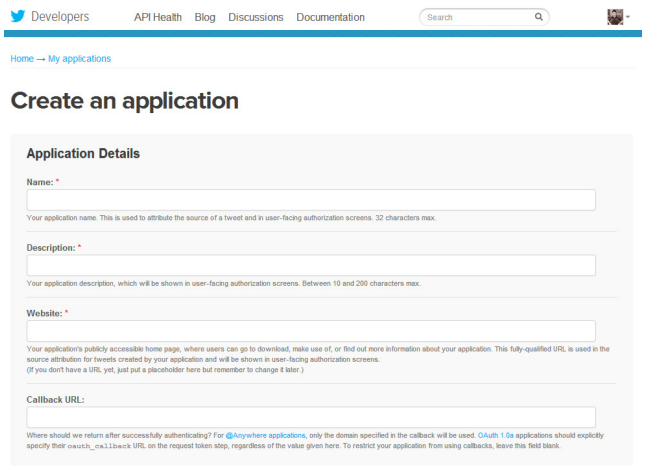


Figure 2. Twitter application.

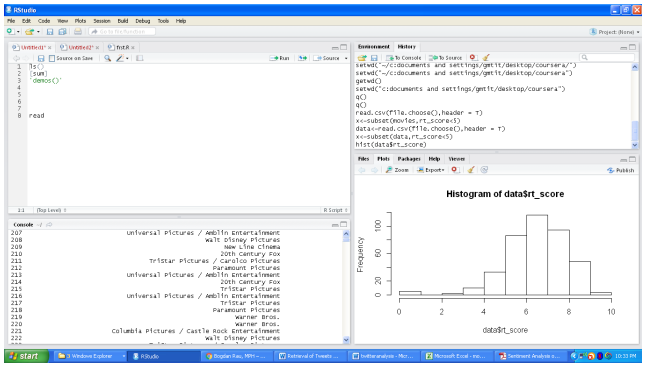


Figure 3. R-Studio environment.

Having four window environment: Left bottom Area is the R console which is the work area, where we implement R scripts. Top left is the R Script area, where R scripts can be written similarly, when we run can be implemented through console. Right top area is Global Environment where variables are defined, data sets are read. Right bottom is plotting area to plot charts for the data.

R code is written in the Script area of R-Studio and it is run to retrive data. The dataset is imported and shown right top area. Plot for the attributes no. of tweets and and their scores is shown in right–left area.

3.3 Import Dataset through Twitter API

In this module, firstly we need to perform handshake property with Twitter. After that we can retrieve latest tweets associated to any keyword of the area. The Search Twitter function is used to extend the final phase of downloading tweets from the timeline[6]. Now this list of tweets are converted into data frame (.df). The .df data frame is converted into .csv format file.

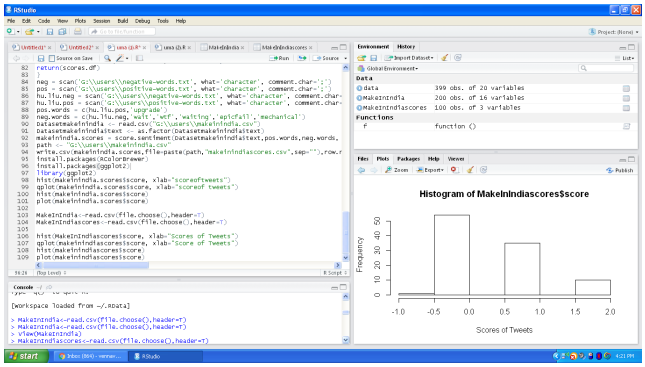


Figure 4. Execution of code in R-Studio.

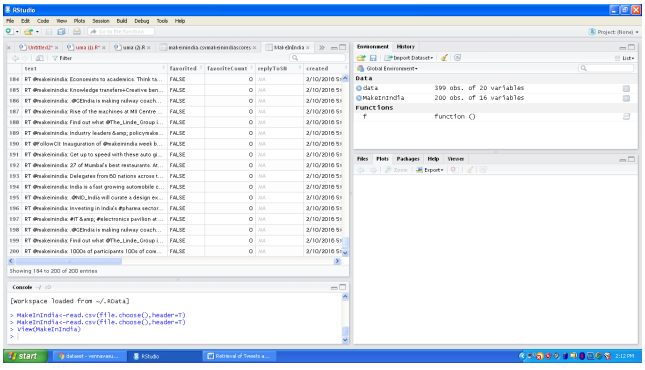


Figure 5. Make–In-India Tweets in R-studio. Imported Make-In-India data set is displayed in R-studio.

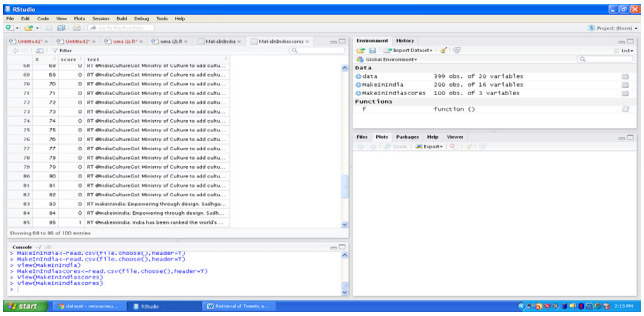


Figure 6. Make–In-India Tweets in R-Studio with scores. Make–In-India dataset after the scores are generated with added scores column.

3.4 Standardizing the Data

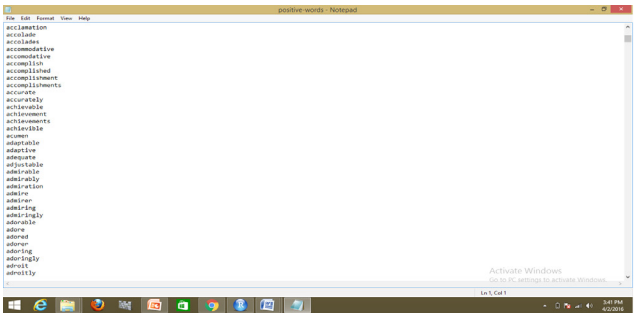
Once we have the tweets we just need to apply some functions to convert these tweets into some useful information. This process is called as standardizing the data. Removal of extra symbols which doesn't give any meaning to the tweets reduces the burden for classification.

3.5 Classification of the Data

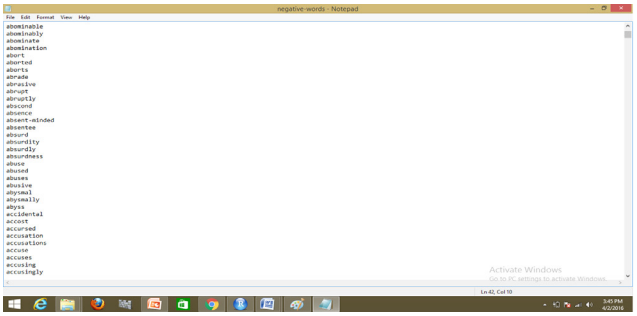
The process of sentiment analysis is to calculate the synchronization of the words of the tweets with respect to Positive word list[6] and negative word list. For this negative word list and positive word list to be downloaded and need to be saved to working directory.

Sentiment analysis requires two additional packages PlyR and StringR to manipulate strings.

File of Positive words:



File of Negative words:



3.6 Getting Scores

The sentiment function[7] calculates score for each individual tweet. It first calculates the positive score by comparing words with the positive words list and then calculates negative score by comparing words with negative words list. The more no. of words that matches the positive list from the tweet gives positive score and the no. of words that matches the negative list from the tweet gives negative score. The final score[8,9] is calculated as Score = count of positive words – count of negative scores.

4. Results

Make-In-India dataset status in terms of public opinions is visualized here. We can create visual histograms and other plots to visualize[10] the sentiments of the user. This can be done by using hist function[11]. We have used a package RColorBrewer to play with colors. All the tweets are considered on to X-axis and the corresponding scores are on the Y-axis. The bar chart[12] tells us that either score that is highlighted to give a decision on the Make-In-India schemes into people's opinions.

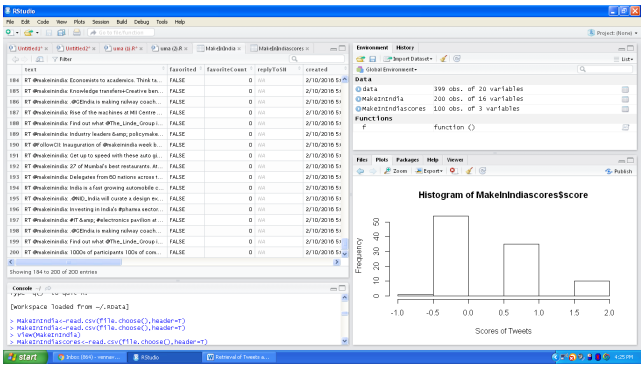


Figure 7. Shows the final analysis of Make-In-India Tweets (i) Tweets with scores and code and plot for the scores data.

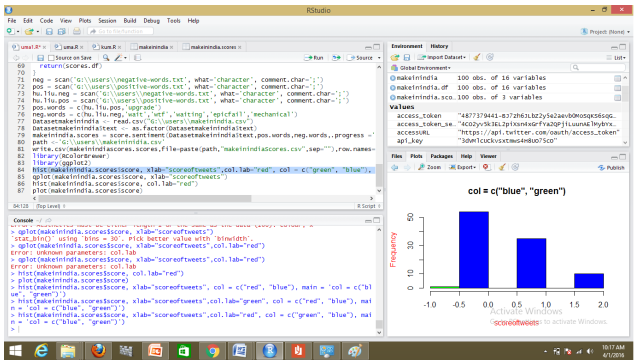


Figure 8. Same as Figure 7 with added colour for plot.

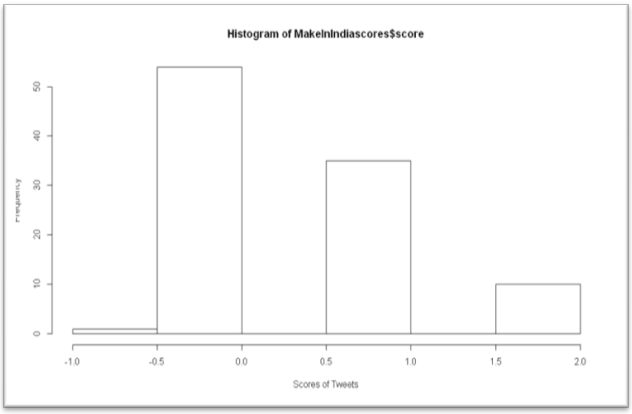


Figure 9. Histogram for the scores of "Make-In-India" Tweets and Tweets count.

5. Conclusion and Future Works

With the rapidly expanding social networks, it is challenging to analyze its large data using existing data mining tools. We have shown that our Architecture to access Twitter and R-Studio Environment analyzes large data for decision making. We have shown through our experiments to do Sentiment Analysis on retrieved 'Make-In-India" data from Twitter that the number of people have given positive and negative opinions on the scheme "MII". With this, it is advisable to conclude R Statistical Tool is sufficiently used for the analysis of Big data. This can be further extended to use PYTHON for more analysis of big data.

6. References

1. Agarwal A, Xie B, Vovsha I, Rambow O. Sentiment analysis of Twitter Data. Proceedings of the …, 2011. Available from: dl.acm.org.
2. Rahmath H. Opinion mining and sentiment analysis-challenges and applications. IJAIEM. 2014 May; 3(5):1–3.
3. Spencer J, Uchyigit G. Sentimentor: Sentiment analysis of Twitter Data. CiteSeer^X 10M; 2012.
4. Sharma Y, Mangat V, Mandeep K. Sentiment analysis and opinion mining. International Journal of Soft Computing and Artificial Intelligence. 2015 May; 3(1).
5. Rao NP, Srinivas SN, Prashanth CM. Real time opinion mining of Twitter Data. International Journal of Computer Science and Information Technologies. 2015; 6(3):2923–7.
6. Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10); 2010 May 19-21.

7. Ding X, Liu B, Yu PS. A holistic lexicon-based approach to opinion mining. Proceedings of First ACM International Conference on Web Search and Data Mining WSDM; 2008.
8. Bifet A, Frank E. Sentiment knowledge discovery in Twitter Streaming Data. New Zealand: Springer Link; 2010. p. 1–15.
9. Bifet A, Holmes G, Pfahringer B, Gavalda R. Detecting sentiment change in Twitter streaming data. Journal of Machine Learning Research. Proceedings Track. 2011; 17:5–11.
10. Fiaidhi J, Mohammed O, Mohammed S, Fong S, Kim TH. Opinion mining over Twitter space: Classifying tweets programmatically using the R Approach**.** IEEE. 978-1-4673-2430-4/12.
11. Parthiban P, Selvakumar S. Big data architecture for capturing, storing, analyzing and visualizing of web server logs. Indian Journal of Science and Technology. 2016 Jan; 9(4). DOI: 10.17485/ijst/2016/v9i4/84173.
12. Liang M, Trejo C, Muthu L, Ngo LB, Luckow A, Amy W. Evaluating R-based big data analytic frameworks. IEEE International Conference on Cluster Computing; 2015, p. 508–9.
13. F Morstatter F, Pfeffer J. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. The 7th International AAAI Conference on Weblogs and Social Media. 2013.
14. Sukhpal K, Rashid EM. Web news mining using back propagation neural network and clustering using K-Means algorithm in Big data. Indian Journal of Science and Technology**.** 2016 Nov**;** 9(41).