Twitter Data Analysis using R

[1]Shubham S. Deshmukh, [2]Harshal Joshi, [3]Pranali Pandhare,[4] Aniket More [5]Prof.Aniket M. Junghare

[1,2,3,4]*BE in Computer engg. ZCOER,Pune,Maharashtra,India*
[5]*Asst.Professor at ZCOER,Pune,Maharashtra,India*

***Abstract*— The growth of Technology has changed the way of expressing people's opinions, views and Sentiments about specific product,services,people and more, by using social media services such as Facebook,Instagram and Twitter. Due to this is massive amount of data gets generated. To find insights from this Data generated and make certain decision we implement web application that collects twitter data and shows it in different statistical forms. The main objective of the work presented within this paper was to design and implement twitter data analysis and visualization in R platform. Our primary approach was to focus on real-time analysis rather than historic datasets. Twitter API allow for collecting the sentiments information in the form of either positive score, negative score or neutral. Then we decided to build our back-end on top of Hadoop platform which includes Hadoop HDFS as distributed file system and Map-reduce as distributed computation. Package twitteR allows to use different twitter function. Hadoop provides integration with R known as RHadoop, it provides different packages used to connect R environment to Hadoop and to perform the analysis of tweets data that are having a size of GB s .To Visualize data we used Rshiny application that generally helps to represent data easily.**

***Index Terms*— RHadoop,HDFS,Sentiments, twitteR, Twitter API.**

I. INTRODUCTION

Currently the volume of the data available in multiple forms and increasing day by day. Rate of data gets increased repeatedly. Due to this processing become question ?Processing large amount of the data also generate new data. Alternatively, processing large amount of date requires computation to achieve results. Efficient tools are important to perform multiple task on big amount of collected Data. Most popular distributed computing paradigms nowadays uses MapReduce. The MapReduce processing paradigm is based on two main phases mapping and reducing.

On the other hand, for storage Hadoop offers HDFS (Hadoop Distributed Filesystem). Also, several machine learning algorithm are available which allows computation when exposed on new data. For integration of R with Hadoop, RHadoop is available as a set of R packages providing interfaces to HDFS and a set of functions to write MapReduce operations. R is a comprehensive statistical platform provides approximate 6000 packages and offers different data analytics techniques. It is a powerful platform for data analysis and exploration. The main goal of this paper is to provide information on tweets ,sentiment analysis and different visualize analytical
data. Therefore, our aim was to use R language and R Studio for development.

Sentimental analysis means to check the opinions, taste,views and interest of people regarding different prospectives such as celebrity, politicians, foods, places, or some other topic. In sentimental analysis we usually classify their mood in different category like positive, negative and neutral.

Lets take a example of "Shahrukh khan is second richest actor" how people react about this? Whether they think positive, negative or neutral. It is difficult to understand and extract information from this kind of data. So, we need such types of tools and Technologies that can efficiently store and process unstructured. There are different
techniques and tools are available that can handle this type of data and produce meaningful insights but in this paper we study R language and RHadoop Tool which perform more statistical computation.

Sentimental analysis of twitter contains slang words, misspelling in tweets and graphics words. So, R statistical computing language is used to perform sentimental analysis. Rhadoop include different packages such as rhdfs, rhbase, rmr and plyrmr which allows to integrate and communicate large with data.

Data analyst, Data Scientists, Data Engineers ,Big Data analyst use R language for Statistical computing and analysis purpose. R is a most popular open source platform with different version on Windows, Linux and mac OS. In some cases when size of data is large and it exceeds memory limits, then it performs slow and gives poor results. Therefore,RHadoop is introduced which stores data into Hadoop(Hdfs) and R is used for fetching data from (HDFS file system) and performs analysis on on that data.

II. TECHNOLOGY USED

RHadoop is a collection of five packages which is mainly designed to support the processing and analyze purpose in the R Environment. It consist of five R packages: rhdfs, rmr2, rhbase, plyrmr and ravro. It is Developed by Revolution

Analytics, this packages are compatible with various distributions of Hadoop frameworks.
RHadoop consists of five R packages: rhdfs, rmr2, rhbase, plyrmr and ravro.
1) ravro – package used to connect to the Avro filesfrom the HDFS.
2) rhdfs – It mainly provides connectivity to a distributed Hadoop file system (HDFS).It able to view,read and edit data stored in HDFS.
3)rhbase – package using to connect to the Hbase and NOSQL distributed database.
4)rmr2 – package providing the set of functions to write a R code that can be transformed into the MapReduce task.
5) plyrmr– package that enables to execute data manipulation functions contained in packaged dplyr and reshape2, but on the large sets of data stored in Hadoop clusters. Similarly, to rmr2, it relies on translation of the R code into the MapReduce paradigm.
One thing is sure that in future more and more data will come on our doorstep. So, there is a need to analyze effectively and use data to improve the outcomes, operations and processes.

III. ARCHITECTURE OF THE PROPOSED SYSTEM

In this work we used a small-sized cluster infrastructure . The configuration of the Master is as follows: 4GB RAM, 4 CPU cores. Overall architecture of proposed system is depicted on Figure 1.Hadoop cluster is used for data storage and processing of the analytical functions written in R. Preprocessing and analysis methods are written using the RHadoop packages functions, which enables the code to utilize the cluster framework . On top of the R implemented scripts, we have developed a R Shiny application which serves as in user end.

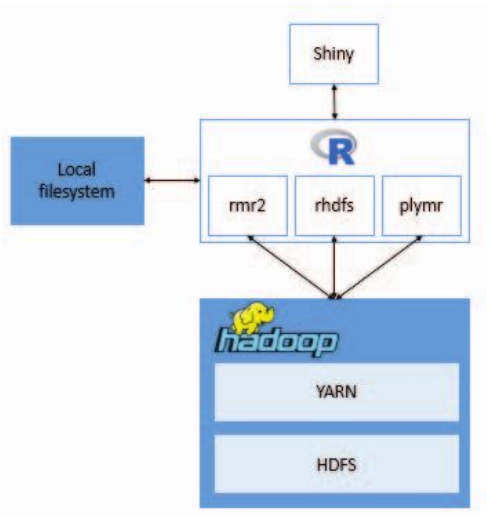

Figure 1. Architecture of Proposed System
[1] http://www.cloudera.com/

IV. METHODOLOGIES OF SENTIMENT ANALYSIS

Step 1. Install and Load all the packages and tools which are required for the sentimental analysis of twitter data. In this paper RStudio GUI is used to install the packages. Packages are given below:
twitteR: This package provides interface to connect with the twitter web API and different functions provided by twitter.
ROAuth: This allows users for the authentication with the twitter via setup_twitter_oauth method.
Plyr: This package have some sets of different methods to solve a diiferent problem. It breaks problem into manageable parts and then combine them.
Stringr: It provides some set of wrapper function that is simple to use.
ggplot2 : This is a graphical system in R for building different graphics by using qplot and plot function.
There are some other packages which are required for installing packages mentioned above. They are dependences which are required for the proper installation of other packages. Some dependencies are like httr, httpuv,dplyr etc.

Step 2. Create a Twitter application by using the twitter developer account. Application will allow to connect from R console to twitter API.

Step 3. Now to download the certificate file from the internet in order to valid the user authentication on the twitter API. After that handshake is performed using consumer key and customer secret key.

Step 4. When the handshake is done, it means we are authorized by twitter. Now we can fetch the data from twitter server using some methods which are provided by twitteR package.

Step 5. When we have the tweets data in our hand, we would apply some functions and commands to extract some meaningful information from tweets data.

Step 6. Finally the output is in the form of scores is fetched in order to classify them in the form of the positive and negative sentiments of different users.

Twitter Authentication
The proposed system connects to the Twitter Streaming API using the developer‘s customer access token and consumer key of twitter account. As soon as the consumer key and customer token is authenticated and handshake is done with twitter API, it provides different methods through which we access the data from twitter server and perform different computation on it. The tweets data contain a lot of misspelling in words, the presence of slang words etc. So, it is very necessary to do per-processing of data and removable of all repeated words and hash tags.

Sentimental Analysis
In this step we check the sentiments of different tweets on the basis of positive score, negative score neutral. Thus, it is helpful for determining the
opinion of the sentiment for the users.
For example: The Sentence in the tweet is:“you are hired”, the word “hired” being an word is hired and compared with the positive and negative word list. Similar with other words.

Using R Language and Hadoop: RHadoop
The integration of R and Hadoop probably did for data analysis and storage technologies to build

a powerful system that have combine the features of both of them. Fig2 shows about the integration of r and Hadoop. There are many ways to integrate them by using different connectors like (Rhadoop, Rhipe, and Hadoop Streaming) but in this paper we used RHADOOP connector.

Figure 2. Integration of R and Hadoop
[2]http://www.r-bloggers.com/

V. ANALYSIS AND VISUALIZATIONS

Now we will provide the details of implemented analyses and their visualizations on the processed real-time tweets .We implemented four main group of data visualizations. For visualization purposes we utilized several R provided packages, such as tm,reshape,gridextra, ggplot, and wordcloud. This analytical task is based on the computation of the trending tweets around different countries around world. For this reason, we needed to prepare extraction functions of tweets for a specific country.

1) Wordcloud
This analytical task developed within the proposed system was the extraction of words from tweets based on the analysis of hashtag (words starting with the #). In this case "content" attribute of hashtag was used for map function and their respective counts were acquired using reduce function to get counts for different hashtag. The example of wordcloud visualization was implemented using wordcloud package and is depicted in Figure 3. In RShiny interface the user has the possibility to select the maximum number of words which will be included in the wordcloud, as well as minimum required frequency of the words which can be used for visualization.

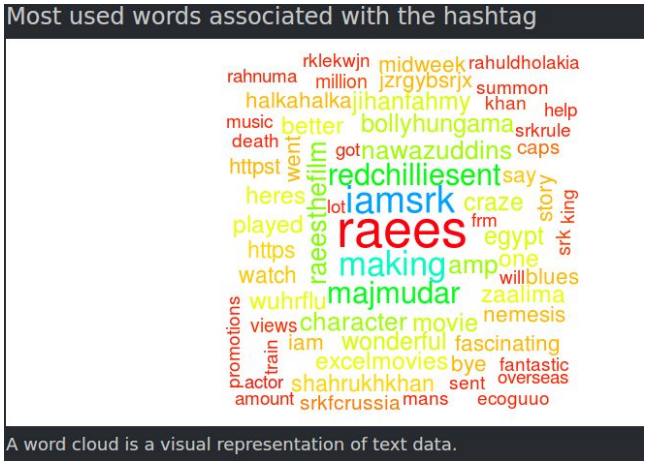

Figure 3. Word Cloud

2) Top 20 tweeters of Hashtag
Here visualization of top 20 tweets of specific hashtag is displayed. It Creates barplot with the **barplot** function, where *height* is a vector and If **height is a vector** then values determine the heights of the bars in the plot.

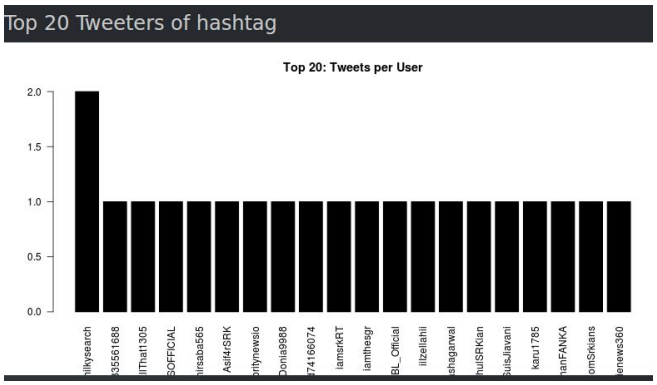

Figure 4. Top 20 Tweets

3) Pie chart
To display sentiments of particular hashtag (people,place,product or service) we used piecharts, it gradually shows opinions of human. For proper outcomes data needs to be represented in multiple forms so that is can help to take actions on certain projects.

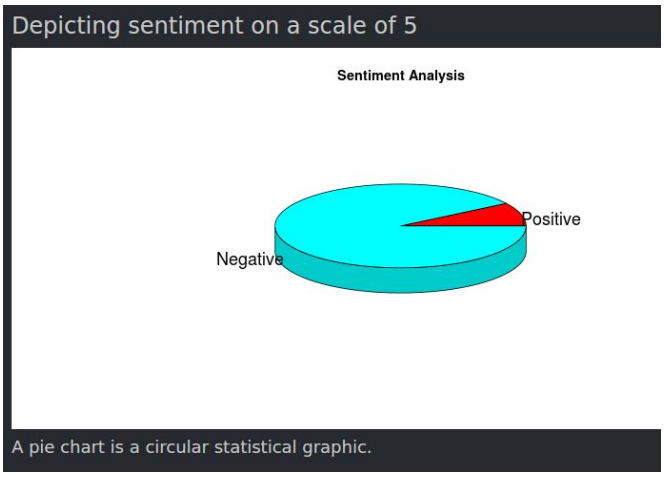

Figure 5. Sentiment in Circular Pie chart

4) Depicting sentiment in a tabular form on scale of 5. In the last analytical task presented here, we provide a tool for analysis of tweets counts produced for specific hashtag. Most of the tweets are created can't be judge according to score. This provides sentiments in tabular form.

Depicting sentiment in a tablular form on a scale of 5

Text	Positive	Negative	Score	PosPercent	NegPercent
@IamSRK_Rahul wo neck kissi wali SRK ki pics u post	0	0	0	0.00	0.00
RT @bharatmatrimony: ___ is a great source of #Inspiration for me! #bollywood #SRK #AkshayKumar #AjayDevgn #HrithikRoshan #InspiredLiving	0	0	0	0.00	0.00
RT @RVCJ_FB: SRK on Actors /VGAY71M8XX	0	0	0	0.00	0.00
RT @SRKCHENNAIFC: I truly belive my job is to make sure people smile - SRK #SRKQuote /hV6Xm6BZ6L	0	0	0	0.00	0.00
RT @aveskhan20: #srk #srkuniverse #rehnuma @SRKUniverse @iamsrk @kamaalrkhan @RedChilliesEnt @SRKFC_Russia watch srk fans this vido https:	0	0	0	0.00	0.00
Good night @iamsrk /YpI85dTEKM	0	0	0	0.00	0.00
#srk #srkuniverse #rehnuma @SRKUniverse @iamsrk @kamaalrkhan @RedChilliesEnt @SRKFC_Russia watch srk fans this vido /8GOQL2pvYC	0	0	0	0.00	0.00
RT @IamMayaSharma: This turned the tables.It gave Akki the confidence to clash wid SRK n he has taken him on.And Overconfident SRK is worri	0	0	0	0.00	0.00

Figure 6. Depicting sentiment in table

5) Top Hashtag of user

After tweets analysis , user now needs to find out frequencies of his own tweets that are retweet by another user. It generally tries to represent retweets of tweets user sends

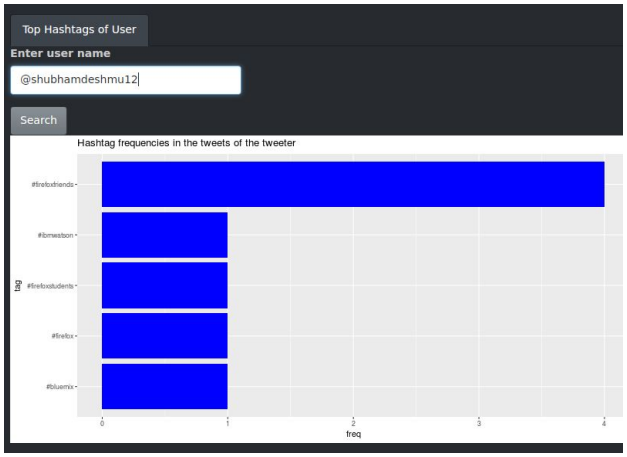

Figure 7. Top Hashtag of user

6) Finally here comes sentiment analysis which shows histogram with score according to positive,negative and neutral opinions of users.

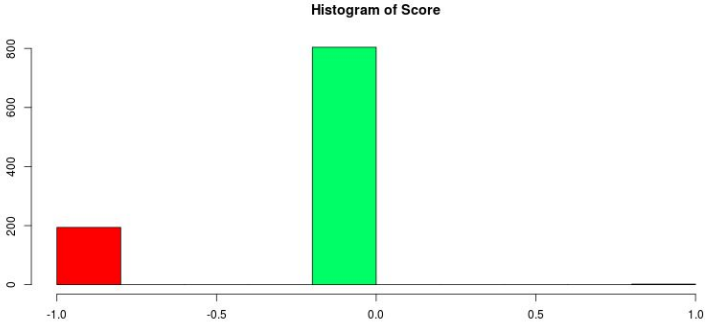

Figure 8. Sentiment analysis of Hashtag

CONCLUSION

The main objective of this paper was to describe and design system for twitter data analysis and visualization. It was developed using R and the big data processing technologies called Hadoop. Different RHadoop packages were used to process large amount of data and to support distributed processing in R. We developed a set of analytical representation which helps user to identify product,people,services and movies data and can gain insights from it. We took a set of visualizations, implemented in Shiny web applications which helps to integrate user interface with RHadoop. RHadoop functions were used and utilized in numerous preprocessing, data cleaning and querying methods.

REFERENCES

[1] R Graphics Cookbook Practical Recipes for Visualizing Data By Winston chang.
[2] Learn Ggplot2 Using Shiny App Book by Keon-Woong Moon.
[3] Web Application Development With R Using Shiny Chris Beeley.
[4] The R Book Book by Michael J Crawley.
[5] A Handbook of Statistical Analyses Using R Book by Brian Everitt and Torsten Hothorn.
[6] Statistical Analysis with R Book by John M. Quick.
[7] https://www.r-bloggers.com/integrating-r-with-apache-hadoop/.
[8] http://blog.revolutionanalytics.com/2015/06/using-hadoop-with-r-it-depends.html.
[9] http://stackoverflow.com/questions/33046510/r-hadoop-integration-how-to-connect-r-to-remote-hdfs.
[10] http://stackoverflow.com/questions/23828643/rstudio-to-connect-to-remote-hadoop-server.
[11] Hadoop: The Definitive Guide Book by Tom White.
[12] Hadoop Operations Book by Eric Sammer.
[13] Real-World Hadoop Book by Ted Dunning.
[14] Data Analytics with Hadoop: An Introduction for Data Scientists Book by Benjamin Bengfort and Jenny Kim.
[15] https://dev.twitter.com/docs.
[16] https://dev.twitter.com/overview/api.