Data Analytics using R Programming

Albert Tungoe[1], Chandan Sharma[2], Pranav Kumar[3]

[1]*Department of Information Technology, Kaziranga University, Jorhat, Assam-785006, India,*
[2]*Department of Information Technology, Kaziranga University, Jorhat, Assam-785006, India,*
[3]Department of Computer Science and Engineering, Kaziranga University, Jorhat, Assam-785006, India,

---

Abstract -

Nowadays the analytics companies has developed the capability to support their decisions through analytical reasoning using various statistics and mathematical techniques. "Data analytics using R programming" not only adds to the existing analytical knowledge and methodology, but also equips us with the exposure into latest analytics techniques including forecasting, social media analytics, text mining, etc. It also gives an opportunity to work on a real time data from social networking sites like twitter, Facebook, Instagram, WhatsApp and many more. Here R programming language and environment uses in Statistical computing, data analytics and scientific research.

1. INTRODUCTION

A statistical analysis package called S was developed by Bell labs in the States. Later in 1994, Ross Ihaka and Robert Gentleman wrote the first version of S at Auckland University and named it R [1]. R is an open-source implementation of S, and differs from S largely in its command-line. For statistical analyses, R has a broad set of facilities that has been specially constructed. As a result, R is said to be a very powerful statistical programming language. The open-source nature of R indicates that, as new techniques for statistics are developed, new packages for R usually become freely available very soon after. It consists of its own inbuilt statistical algorithms – the sheer amount of machine learning algorithms and mathematical models available to users in R and third-party packages is staggering and continues to grow. R can also carry out important analyses that are difficult or next to impossible in many other such packages, including Generalized Additive Models, Linear Mixed Models and Non-Linear Models. R consists of broad range of graph-drawing tools, which makes it easy to produce standard graphs of your data. In traditional analysis, developing a statistical model takes more time than by performing the calculation by the computer. In case of Big Data this proportion is turned upside down. Big Data comes into picture when the CPU time for the calculation takes longer than the process of designing a model. Data sets that contain up to millions of records can easily be processed with standard R. Data sets with almost one million to one billion records can also be processed in R, but requires some additional effort. Worldwide, millions of statisticians as well as data scientists use R in order to solve their most challenging problems in the field, right from quantitative marketing to computational biology [1]. R have become the most popular language for data science and an most essential tool for analytics-driven companies such as Google, Facebook, LinkedIn and Finance .

1.1. BIG DATA ANALYTICS:

The business value is not generated by data stored and this is true as for traditional databases, data warehouses, also for the new technology like Hadoop and many more for storing big data. Once the data is appropriately stored, it can be analyzed, and thus immense value can be created. In-memory analytics, in-database analytics and a variety of analysis, technologies and products have arrived that are mainly applicable to big data.

1.2. R's GROWTH:

In 2020, IEEE had listed R at 8th position in the top 10 languages of 2020 [4]. In addition to this, as the amount of intensive data work increases, demand for tools like R for data-mining, processing and visualization will also increase.

2. LITERATURE REVIEW

2.1. Literature Review 1.

Title: Twitter Data Analysis using R

Author: Shubham S. Deshmukh, Harshal Joshi, PranaliPandhare, Aniket More, Prof.Aniket M. Junghare.

Publication: International Journal of Science, Engineering and Technology Research (IJSETR) Volume 6, Issue 4, April 2017, ISSN: 2278 -7798

2.1.1. Overview:

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.

2.1.2. 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.

2.2. Literature Review 2.

Title: Sentiment Analysis on Twitter Data using R.

Author: Sonia Saini,

Amity Institute of Information Technology, sector-125

Noida, U.P, India.

RituPunhani,

Amity School of Engineering & Technology, Sector-125

Noida, U.P, India.

RuchikaBathla,

Amity Institute of Information & Technology, Sector-125

Noida, U.P., India.

Vinod Kumar Shukla,

Department of Engineering & Architecture, Amity University

Dubai, UAE

Publication: 2019 International Conference on Automation, Computational and technology Management, Amity University.

2.2.1. Overview:

Sentimental Analysis is an ongoing research field in Text Mining Arena to determine the situation of market on particular entity such as Product, Services...Etc. and it can be called as computational treatment of reviews, subjectivity and sentiment of text. Crypto currency can be explained as a type of digital estate and devised to mechanize as a form of trade and exchanges that uses cryptography as an encryption technique to secure the transactions and acts as decentralized controlled transaction which is opposed to centralized transactions. Crypto currency are a type of virtual currency, digital currency and alternative currency, On basis of categorical, there are different architecture and security protocols which are used in the crypto currencies to secure transactions, the different types of crypto currency are available in the market such as Bitcoin, Lite coin, and Name coin...etc. This paper focuses on survey on different types of sentimental analysis methods and main contribution of this paper include sentimental analysis of social media data on different types of crypto currencies on basis of categorical and different terms of crypto currency such as Crypto currency, virtual currency, digital currency and discussed on trends of crypto currency in present market.

3. ABOUT R

3.1. R in business:

R was originated as an open-source version of the S programming language in the 90s. It has gained the support of a number of companies since then, mostly R Studio and Revolution Analytics that are used to create various packages, and services related to the language. R has support from large companies that power to some of

the largest relational databases in the world. Oracle, as one, has incorporated R into its offerings.

3.2. R in higher education:

R is also originated in academia. Ross Ihaka and Robert Gentleman in New Zealand at the University of Auckland created it, and it's also been widely adopted in graduate programs that include intensive study of statistics. Massive open online course such as the Coursera Data Science Program also makes use of R.

3.3. R has a diverse community:

The R community is diverse, along with many individuals coming from unique professional backgrounds. This list includes statisticians, business analytics, academics, scientists and professional programmers. The comprehensive R Archive Network (CRAN), maintains packages that are been created by community members that reflect this background. Packages exist in order to create maps, perform stock market analysis, engage in high-throughput genomic analysis and perform natural language processing.

3.4 R is fun:

R has an ability to generate charts and plots in very few lines of code. Tasks that would require multiple lines of code in some other language could be accomplished in R in only a few lines of code. While it's been considered strange when you compare it with many popular languages, it includes powerful features specifically when geared towards data analysis.

3.5. R's CHALLENGES:

For all of its benefits, R has its share of shortcomings as follows-

Memory management

Speed

Efficiency.

These are probably the biggest challenges R faces. Also, people coming to R from other languages might also consider R odd. When working with very large data sets the design of the language can sometime lead to problems. Data has to be stored in physical memory. But this can become a minor issue, as nowadays computers have plenty of memory. Abilities such as security were not built into the R language. Also, R cannot be embedded in a Web browser. You can't use it for Web-like or Internet-like apps. It was primarily next to impossible to use R as back-end server to perform calculations due to lack of security over the Web. For a long time, there was not a lot of interactivity in the language. Languages such as JavaScript still have to enter in to fill this gap. Although an analysis may be done in R, the furnishing of results might be accomplished in different language like JavaScript.

4. SCOPE OF PROJECT

Every individual is said to be addicted to WhatsApp group or not and this is done using the R statistics software programmed. The conclusion is expected to clarify the level of addiction of an individual to WhatsApp Group and regards to Gender and Age with Frequency of use.

5. METHODS AND APPROACHES

a. Setting up Machine

Once R is installed you can choose to work with an integrated development environment (IDE) R Studio. It is the most popular IDE for R and supports debugging, workspace management, plotting and much more [6].The Window of R Studio is divided into four panes namely source pane, console pane, workspace pane and plots pane.

b. Source Pane

On the left side on top is the source pane where you can write and edit your R programs and documents.

c. Console Pane

It is located on the left side at the bottom, where results are displayed.

d. Workspace Pane

It is located on the right side on top and allows quick access to additional tools. It is used to perform the following functions

- Environment that exhibits data objects defined in the current R session.
- History is previous commands list that needs to be executed.
- Plots Pane

Located on the right side on bottom, allowing quick access to additional tools and performing the following functions

- Files to browse folders.
- Plots illustrates plots created by the user.
- Packages option from where packages can be installed and loaded.
- Helps to get help on R commands.

e. Data cleansing

Data cleansing is also known as Data Scrubbing in which inaccurate records from a particular dataset

are corrected and eliminated. The purpose of data cleansing is to detect incorrect, irrelevant or insufficient parts of the data to either alter or delete it to ensure that a given set of data is accurate and

	Multimedia	SmileyText	Text
Female	661	1583	567
Male	800	1402	550

V l any
is to R:
Ir served for the header

Avoid blank space in names, values or fields, else each word will be treated as a separate variable, resulting in errors that are related to the number of elements per line in your data set.

To concatenate words, do this by making using of a dot (.) .For example Sender. Age

Short names are preferred over long names.

To avoid special symbols such as !, @, #, $, ^, ***, ,(,) ,- , ?,,,< ,> , / , | , \ , [,] ,{ , and }.

Values missing in the data set are tend to be indicated with NA.

Performing all such validations on text file and hence converting it into csv file via the help of excel proved to be a tedious job and hence automating the process of conversion of text file into csv file by writing few lines of code proved to be more efficient, less time consuming as well as reduced manual work. The process of conversion with all the above mentioned validations thus shifted from hours to minutes.

f. Import Data File

In R Studio, click on the Workspace tab, and then click on “Import Dataset”. Select “From local file”. A file browser will open up, locate the .csv file and click Open. A dialog box will appear that would mention a few options on the import. Make sure if you have column names in your file then the Header is set to “Yes”. Mention the separator as used for a particular csv file. In this case semicolon is been used as a separator. Finally Click “Import”. Note that the name of your data appears in the Workspace pane. A preview of the data opens in the file-viewing pane.

The aim of this research is to classify the number of users as those addicted and not addicted to WhatsApp group chat and thus predicting the level of addiction as well as to find a way to answer:

To find what type of communication medium people prefer the most in WhatsApp Group chat.

As our dataset consists of records from a WhatsApp group of approximately 2 years of the communication medium people preferred more in both the years is Smiley and Text.

The Communication medium in the given dataset is divided into three parts mainly Multimedia (Audio, Video),Text , Smiley and Text.

- To find most active day of week.

As per the performed analysis and visualization illustrated in Fig. 5.1. The most active day of the week is “Monday“ with total number of messages send as well as received are 587 and 628 respectively. So maximum participation of the senders on the WhatsApp group chat takes place on Monday and the least is on Friday.

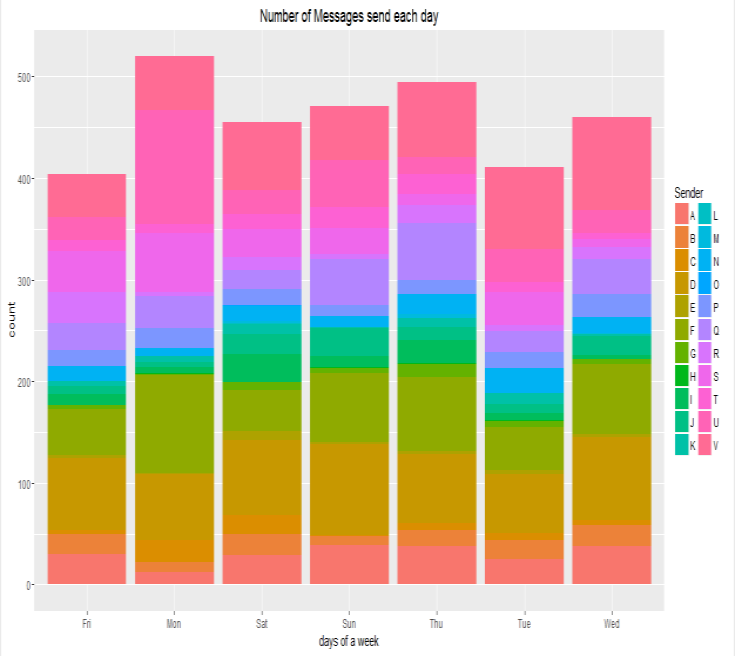

Fig 5.1: most active day of the week.

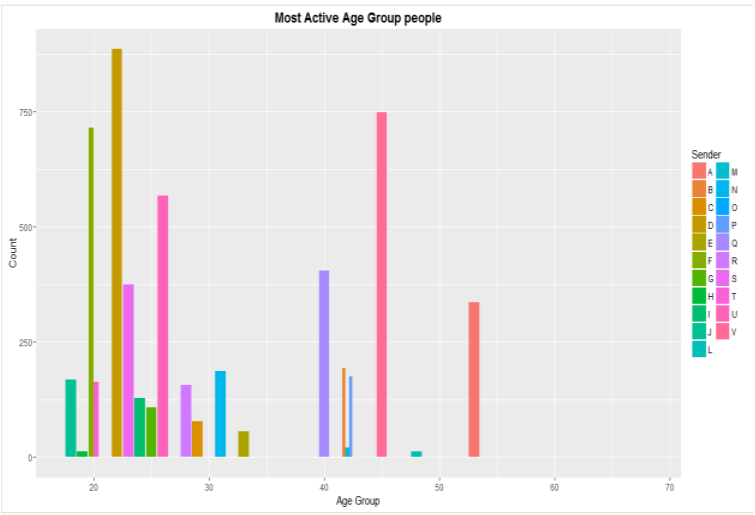

Fig 5.2: Most active age Group.

- To find which age group participants are more active on WhatsApp group and number of messages send by each age group participants per month, day, and hour.

The age group of the participants in the Dataset ranges from 18 to 58. So the youngest person of the group is of age 18 and the most eldest person has an age of 58.

It is clear enough that the most active participants on the WhatsApp Group are of Age group 20 to 30 and least active participants are of Age group 50 to 60.

As per the analysis, total number of messages send by every age group person per hour is shown in Fig. 5.2 ,where 00,01 … 11 are number of hours and the column on the left depicts names of participants between age group 18 to 58. The filled rows and columns are the number of messages send by every participants per hour. Similar analysis can also be performed as per months and per days instead of hours.

- To find whether Males are more addicted to the WhatsApp group or Females:

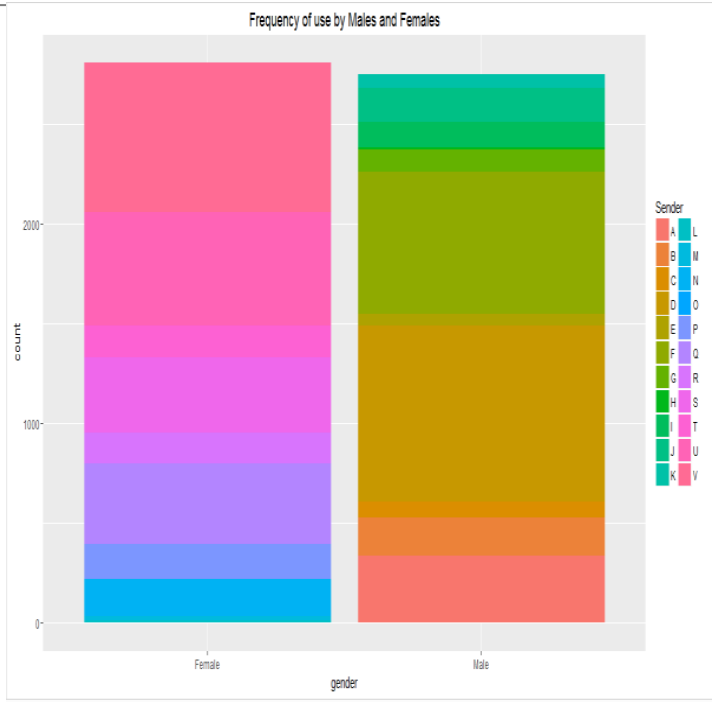

Fig 5.3: Frequency of using WhatsApp by the two gender.

According to the analysis, Females are likely to be more addicted to the WhatsApp group as compared to Males. Total number of messages send by the Females are 2957 and by Males are 2762 as shown in Fig. 3.12 .This clearly concludes that Females are more involved in the Group.

- Total number of messages send as per timestamp.

Total of 3674 messages had been delivered at PM while a whole of 1889 messages were delivered at AM. As a result maximum amount of interactions took place after noon as shown in Fig 5.4 with the help of a Pie chart.

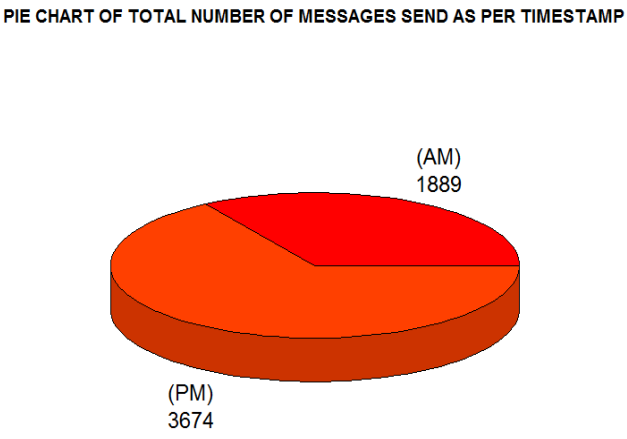

Fig 5.4: Total Number of Messages send as per timestamp.

6. RESULTS AND DISCUSSION

The following are the finding from the above analysis:

- The dataset consisted of equal number of male gender and female gender respondents.
- Most of the respondents were in the age group of 20 to 30 years implying a young sample.
- Least respondents were in the age group of 50 to 60 years implying a senior sample.
- The most active day of the week was Monday.
- Females have been found more addicted to WhatsApp group than Males.
- Users were keener to share Smiley and text as compared to other multimedia items.
- According to the Timestamp most of the messages were send after noon (PM).

7. CONCLUSION

From the performed analysis and visualization it is found that total number of active users in WhatsApp group chat are 24 consisting of equal number of males and females. Majority of the female users tend to be more addicted to WhatsApp group chat as compared to male users, due to various features provided by WhatsApp such as multimedia, Smiley and Text. The most addicted respondents were in the age group of 20 to 30 years representing a young sample. So as to conclude we would like to say that WhatsApp is a great communication platform where there will be pros and cons subjected but it is upto the user how they utilize this technology.

ACKNOWLEDGEMENT

The project is undertaken by **Albert Tungoe and Chandan Sharma**as our MCA 6th semester project on "Data Analytics using R Programming", under the guidance and supervision of **Mr. Pranav Kumar**.

Our primary thanks goes to him who poured over every inch of our project with paint stalking attention and helped us throughout the working of the project. It's our privilege to acknowledge our deepest sense of gratitude to him for his inspiration which has helped us immensely.

We are extremely grateful to him for his unstilted support and encouragement in the preparation of this project.

We show our gratitude to our Dean **Dr. SajalSaha** and HOD "**Dr. Manoj Kumar Muchahari**" and our Project Co-ordinator**Dr. PurnenduBikashAcharjee** for providing the best of facilities and environment to bring out our innovation, talent and spirit of inquiry through this project.

URL REFERENCES

[1] http://www.r-statistics.com/tag/hadley-wickham/, accessed on 28/1/2020 at 7:35 AM.

[2] http://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html, accessed on 28/1/2020 at 7:45 AM.

[3] http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages, accessed on 30/1/2020 at 6:42 PM.

[4] http://www.analytics-tools.com/2012/04/r-basics-introduction-to-r-analytics.html , accessed on 30/1/2020 at 8:45 PM.

[5] http://blog.revolutionanalytics.com/, accessed on 2/2/2020 at 7:50 AM.

[6] http://www.r-bloggers.com/handling-large-datasets-in-r/, accessed on 4/2/2020 at 11:30 AM.

[7] http://www.analytics-tools.com/2012/04/r-basics-introduction-to-r-analytics.html, accessed on 5/2/2020 at 9:13 PM.

[8] http://data.vanderbilt.edu/~hornerj/brew/useR2007.rhtml, accessed on 7/2/2020 at 5:30 PM.

[9] http://blog.ukdataservice.ac.uk/the-power-of-r-methods-for-processing-big-data/, accessed on 7/2/2020 at 6:12 PM.

[10] http://bigdatauniversity.com/moodle/course/view.php?id=522, accessed on 11/2/2020 at 9:12 AM.

[11] http://aisel.aisnet.org//viewcontent.cgi?article=385&context=cais, accessed on 11/2/2020 at 5:40 PM.

[12] http://www.revolutionanalytics.com/what-r, accessed on 15/2/2020 at 8:45 AM.

[13] http://blog.revolutionanalytics.com/2013/12/tips-on-computing-with-big-data-in-r.html, accessed on 15/2/2020 at 9:50 AM.

[14] http://blog.ukdataservice.ac.uk/the-power-of-r-methods-for-processing-big-data/, accessed on 19/2/2020 at 7:28 AM.

[15] http://www.odbms.org/blog/2013/02/on-big-data-analytics-interview-with-david-smith/, accessed on 27/2/2020 at 8:24 AM.

[16] http://www.slideshare.net/bytemining/r-hpc, accessed on 27/2/2020 at 5:17 PM.

[17] https://www.pluralsight.com/blog/software-development/r-programming-language, accessed on 1/3/2020 at 6:45 AM.

[18] http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages, accessed on 1/3/2020 at 7:24 AM.

[19] http://www.stat.yale.edu/~mjk56/temp/bigmemory-vignette.pdf, accessed on 19/3/2020 at 8:25 PM.