WhatsApp Group Data Chat Analysis Using Big Data Analytics

Abhishek Shah[1], **Siddh Patel**[2], **Divyesh Patel**[3]

Student[1,2], Assistant Professor[3] , Department of Computer Engineering[1,2,3],

Chandubhai S. Patel Institute of Technology, Changa, Gujarat, India.

1. ABSTRACT

In recent days, the means of communication has changed over time due to the advancement in technology. So, the amount of data floating across the internet has also increased and has become more scalable, sceptical and diverse compared to the structured data. Also, the process of transferring data is no longer restricted to in the form of text and between two people. In fact, now we can have the exchange of data among a large group and can share data in terms of text, audio, Images and videos. So, due to this storage handling and analysis of such huge unstructured data results in complicated and complex processing and visualisation problems. To overcome such problems, we use Big data analysis. It consists of various tools and techniques that solve visualisation and analysis problems. Big data tools provide a vast range of visualisation techniques that result in the graphical view of the data. We had considered analysing WhatsApp Application's group chats as it is one of the most used messaging application worldwide. This research paper focused on the analysis including emotions, the most active user and day, the total number of messages sent by each user, etc. of the group data in which we used text mining to fetch the raw chat and later used analytical tools to generate graphical results.

2. INTRODUCTION

In this era due to various features of WhatsApp as a messaging application its use has increased for communication. People in every field tend to be an active user of WhatsApp regardless of them being on a business account or a Personal account. WhatsApp had 450 million active users in 2018 and its going on increasing every day. Thus, due to this increased usage of messaging applications the data has not only been limited to structural but has also moved towards semi-structural and unstructured types, which create a variety of data mixed together. Such type of mixed data can be filtered and used with the help of big data analytics. It provides diverse, scalable and intelligent approach for data visualization and analysis.

The objective of this paper is to provide basic idea of analyse done on a particular WhatsApp group data. Following are the sections in which the research has been carried out.**[9]**

- Number of post per user.
- Most common words.
- Word cloud of the chat.
- Number of post sent by each user at a particular time in a day.
- Analysis of chat history of each user.
- sentiment analysis of the chat using three lexicons i.e. NRC, Bing and Loughran. The value of the sentiments depends on these lexicons[1].

In particular, this paper also emphasises on the usage of R studio software for data analysis and how it can be used to extract and work with a dataset. R is an open-source data analysis programming language which is mostly used for big data analysis.**[8]**

3. BIG DATA ANALYSIS

Procedure of cleaning, changing, assessing and demonstrating data with the aim of revealing helpful data, and thus supporting decision-making is Data Analysis. Big data is a field devoted to the analysis, processing and storage of huge capacity of data which are frequently generated from different sources. In particular, Big Data information tends to unmistakable necessities. Such as, consolidating of various irrelevant data sets, handling of vast measure of unstructured information and separating of concealed data in a period delicate way. Big Data life cycle for the most part includes recognizing, obtaining, getting ready and breaking down a lot of crude. Unstructured information to separate significant data that can fill in as a contribution for distinguishing designs, improving existing information and performing vast scale looks.**[1]**

[1] NRC is an emotion lexicon developed by Saif Mohammad and Peter Turney, Bind is the sentiment lexicon developed by Bing Liu and his team. furthermore, Loughran lexicon is mostly appropriate for financial text and it was developed by Tim Loughran and Bill McDonald.

4. TYPES OF DATA

Data sources can be varied and it can be any format like audio, video, text, table, etc., such that it is categorized as follows:

4.1. Structured Data

Structured Data alludes to information that goes into a social database, exists in predefined fixed fields, and is findable by means of hunt tasks or calculations. Structured Data is very easy to enter, spare, find and break down; in any case, it must be all around characterized with respect to handle name and character type. Consequently, Structured Data is regularly confined in use on account of its firmness. A few instances of Structured Data are budgetary subtleties, call detail records, web server logs and human information. Analysts and programmers taking a shot at this sort of information use SQL innovation for RDBM.**[2]**

4.2. Unstructured Data

Unstructured Data does not fit into a data store. Notwithstanding, it might have its inner structure. While unstructured Data appears to be sorted out in nature, it is additionally loved and progressively accessible as perplexing data designs, for example, messages, content documents, website pages, computerized pictures, interactive media content, route subtleties and internet based life posts. Actually, most of business connections appear to be sloppy in nature. There are a few different ways to begin gathering a database of unstructured Data and handling it.**[6]**

4.3. Semi-structured Data

semi-structured data as a kind of information that contains semantic labels, yet does not adjust to the structure related with commonplace social databases. While semi-structured substances have a place in a similar class, they may have distinctive traits.

4.4. Quasi-structured data

It is absolutely instinctive, rising, pseudo, surmise, apply a standard and refine process.

5. TOOLS IN BIG DATA ANALYTICS

R Language gives rich graphical offices to information investigation and backings fundamental charts to cutting edge diagrams. R supports illustrations to make essential diagrams like pie, bar and line. R likewise has augmentation of numerous packages for information perception. Representation of information comprehends the patterns in extensive data sets. R offers numerous packages for information examination and gives interfacing different instruments for information investigation. There are in excess of 7769 packages.Accessible including GitHub, CRAN and Bioconductor. CRAN records the perspective on fundamental essential packages in a sorted-out way.

R package is an accumulation of capacity, information and assembled code in a characterized configuration and put away as library. Standard packages are introduced in R for fundamental information the executives, examination and graphical presentations. Extra packages can be installed and stacked as required.**[3]**

6. METHODOLOGY

The method of WhatsApp chat data analysis is divided into different stages. The process includes gathering relevant data, importing into R studio and analysing it to obtain the results.

6.1. Data collection

Collecting data is the first stage of the process flow which includes defining project, setting up the machine and later understanding the data. The process flow has been described below in the figure 1.

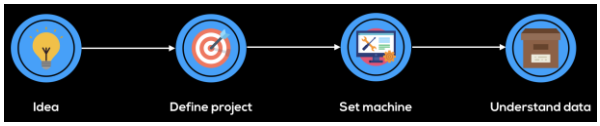

Figure 1 : Process flow for data collection

Data collected should be accurate for the research to give perfect results.

6.1.1. Idea

The user must know the source of data and its usefulness. The data from WhatsApp can be extracted by the following steps.

Open Group -> Click menu button -> Click more

->Select Export Chat -> Send via mail.

The output of the chat will be a .txt file which you will receive.

6.1.2. Define project

This research paper analyses various things like the duration a user spends chatting in the group, his/her sentimental analysis and various other analysis are carried out. This is done with the help of R programming. The results obtained at the end are expected to help in analysing the group chat.

6.1.3. Set Machine

First R is installed on the system then the Integrated development environment i.e. R studio is installed to carry out Big data analysis.

The R studio is divided into four panes which are named as source pane, workspace pane, console pane and plots pane.

6.1.3.1. Source Pane

This pane is located on the top-left corner of the R studio screen and is used to write and edit the R programs and document.

6.1.3.2. Workspace Pane

This pane is located on the top-right corner of R studio screen and it allows quick access for additional tools. It is used for performing following functions.

History: It is a list of the commands used previously which are to be executed.

Environment: It shows the data objects defined in the current R studio session.

6.1.3.3. Console Pane

This pane is located on the bottom-left corner of the R studio screen. It is used to display the results.

6.1.3.4. Plots Pane

This pane is located on the bottom-right corner of the R studio screen. It allows quick access to additional tools and performs the following functions.

- It consists of files used for browsing the folders.
- Help to get the help on R commands.
- Plots illustrate plots created by the user.
- Packages option used for checking the total packages installed and loaded into the R studio.

6.2. Data Conversion

After fetching the raw data from the WhatsApp chat in the .TXT format we need to convert it into usable data format i.e. to (comma separated value).CSV data so that data analysis can be performed on it.**[4]**

6.2.1. Data Scrubbing

Data scrubbing means cleaning the data in which the irrelevant data is removed from the dataset. The purpose behind this is to detect incorrect and useless data and to ensure that such messy data is either altered or removed. This process is must to obtain a clean, accurate and efficient dataset.

This also includes imposing various validations on the text source file so to avoid any issues of irrelevancy in future. Validations which have to be imposed are mentioned below.

- Blank spaces should be avoided in names, values or any other field.
- Including special characters in the data should be avoided.
- Short names should be used instead of long names.
- While using spreadsheets the first row should be reserved for the header.
- If for any record their value is missing it has to be indicated by NA.

After performing this validations on the text file, it has to be converted into the csv file format so that analysis on the data becomes easier.

6.3. Data Loading

This step involves importing the csv file in the R studio.

6.3.1. Import Data file

In R studio, click on workspace Pane then click on “import dataset”. Select “From local file”. A file browser opens up and then import the .csv file. The preview of the data will open in the file-viewing pane.

7. Data Analysis and Visualisation

The main aim of this research is to classify Number of post per user, most common words, word cloud of the chat, number of post sent by each user at a particular time in a day, analysis of chat history of each user and sentiment analysis of the chat.**[5]**

The results of the analysis are explained below.

7.1. Post per user

The Figure 2 shows the total number of chat posted by each

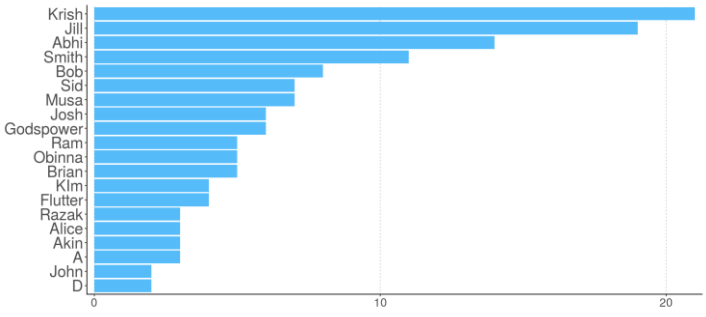

Figure 2: Post Count per user

7.2. Word Frequency

The Figure 3 shows the most common used words in the chat by all the users. We can also set the word length and according to the specified length the most used words will be displayed.**[7]**

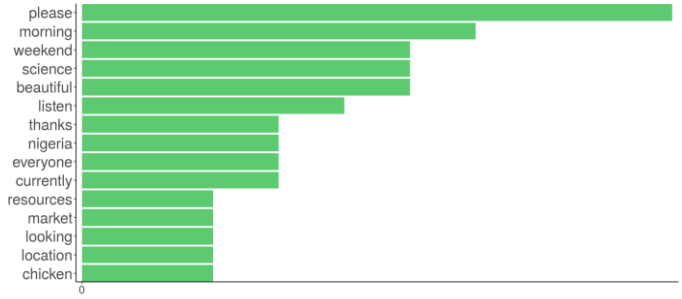

Figure 3: Word Frequency

7.3. Word Cloud

The Figure 4 shows the word cloud formed by the most used words in the whole group chat.

Figure 4: Word Cloud

7.4. Time of the Day

The figure 5 shows the number of post sent by each user at a particular time in a day in the group. The figure 6 shows the analysis of only one user(Bob).

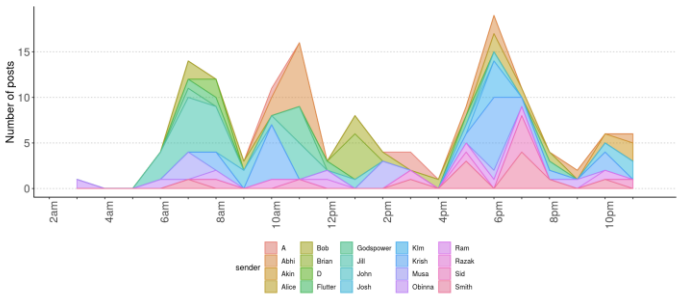

Figure 5: Time of the day (All users)

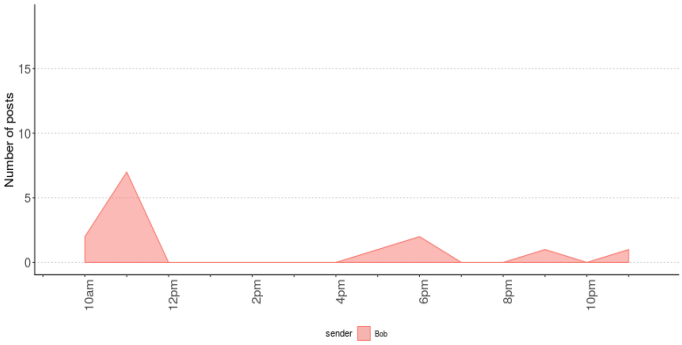

Figure 6: Time of the day (Single User)

7.5. Date

The Figure 7 shows the chat history for every user in the group. It shows the number of post by each user in a particular year. The figure 8 shows the same result for a specific user (Jill).

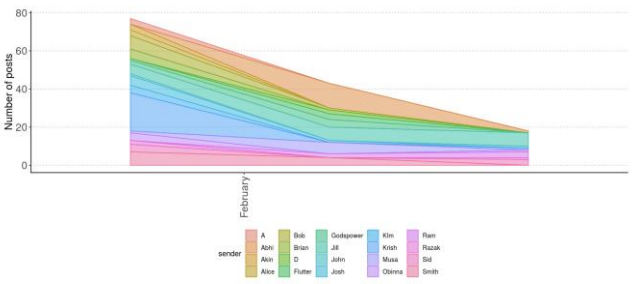

Figure 7: Date (All Users)

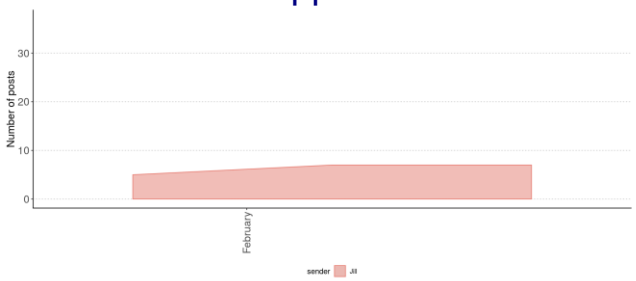

Figure 8: Date (Single User)

7.6. Sentiments

The result of the emotional analysis is done using three lexicons namely Nrc, Bing and Loughran. This sentiment analysis shows emotions of every user in the group by predicting their chat.

Figure 9 shows sentiment analysis using Nrc lexicon, Figure 10 shows sentiment analysis using Bing Lexicon and Figure 11 shows sentiment analysis using Loughran Lexicon.

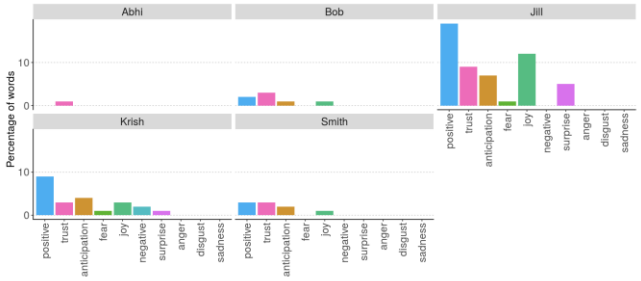

Figure 9: Sentimental Analysis (NRC Lexicon)

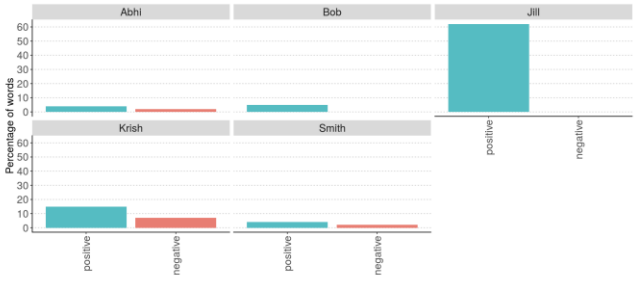

Figure 10: Sentimental Analysis (Bing Lexicon)

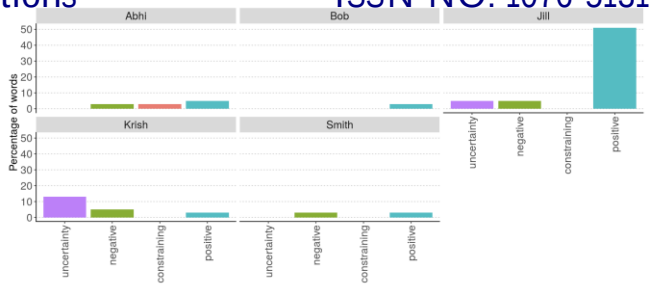

Figure 11: Sentimental Analysis (Loughran Lexicon)

8. Result and Conclusion

The complete group chat analysis helps the users to understand the amount of post they contribute to the group and help them understand the emotions they share with each other in the group. All the figures from 2-11 are the outputs of all the observations carried out for the group chat using Big Data Analysis in R studio.

From the performed analysis and visualisation, we get the idea about how the user and extract the meaningful data from their WhatsApp group chat and analyse them to see and achieve various objectives.

9. References

[1] Patel, Divyesh. (2018). Sentiment Analysis of Harry Potter Book Series using R.

[2] Priyank Jain, Manasi Gyanchandani and Nilay

Khare, “Big data privacy: a technological perspective and review”, Journal of Big Data, DOI 10.1186/s40537-016-0059-y

[3] Thomas Erl, Wajid Khattak, Paul Buhler, “Big Data Fundamentals: Concepts, Drivers & Techniques”,Pearson India Education Service Pvt. Ltd. 2016.

[4] Kharde, Vishal & Sonawane, Sheetal. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications. 139. 5-15. 10.5120/ijca2016908625.

[5] V. Bhuvaneswari,“Data Analytics with R”, published by budca.in,ISBN:978-81-929131-2-4, Edition, 2016

[6] Gautam, G., Yadav, D.: Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In: Seventh International Conference on Contemporary Computing(IC3), India. IEEE (2014)

[7] The R Project for Statistical Computing. https://www.rstudio.com/

[8] Samiddha Mukherjee, Ravi Shaw, " Big Data – Concepts, Applications, Challenges and Future Scope", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016

[9] Sagar Deshmukh, "Analysis of WhatsApp Users and Its Usage worldwide", International Journal of Scientific and Research Publications, Volume 5, Issue 8, pp. 1- 3, August 2015 1 ISSN 2250-3153.