# WhatsApp Data Analysis

Text and Sentiment Analysis in R

Akshay Shah
*School of Computer Science*
*University of Windsor*
Windsor, Canada
shah1bz@uwindsor.ca

Manil Patel
*School of Computer Science*
*University of Windsor*
Windsor, Canada
patel3h@uwindsor.ca

Srihari Jayachandran
*School of Computer Science*
*University of Windsor*
Windsor, Canada
jayacha1@uwindsor.ca

Pooya Moradian Zadeh
*School of Computer Science*
*University of Windsor*
Windsor, Canada
pooya@uwindsor.ca

***Abstract*—The sudden outburst of the Internet has made it easier for people to communicate with people across the world. Social media provides humongous data to carry out various algorithms and predict the output. WhatApp is considered to be the most popular application due to its huge customer base. In this project, we have performed the text and opinion mining on WhatsApp chat data to understand the people in a better way. We chose WhatsApp data due to its number of users and way of retrieving data. Our project will deliver the graphical visualizations for Text analysis and sentimental analysis which include message count, emojis, frequency of the texts and emojis, lexical diversity, word cloud and sentiment score. We believe that the results of this project could help the Government, companies and Organizations to gain better insights about the people. This analysis might also help to avoid conflicts that might arise due to emotional mismatch or misunderstanding.**

***Index Terms*—text mining, opinion mining, WhatsApp, sentiment, emojis, text**

## I. Introduction

In the current age of the internet, the fastest mode to communicate with our family, friends and other people is the WhatsApp mobile application. It is the one single app where each smartphone user shares their opinions, thoughts and sentiments with other people [1]. The innovative feature that was employed by WhatsApp was the group chat feature, by which information can be shared to many people. This was the main reason for any news to spread quickly through WhatsApp irrespective of its fakeness. So, it has become the need of the hour to analyze the WhatsApp texts to find out the positivity of the people. WhatsApp has provided an easy facility to convert a group chat into a text file through email. However, it is difficult to analyze the chats manually due to a very large dataset. Opinion mining is a process in which each statement from a huge dataset could be analyzed [2].

R is a programming language that contains many packages and functions for data cleaning, mining and sentimental study. The text and sentimental analysis can be executed by using R Studio [3]. In our project, we have reviewed the work made in the sentimental and text analysis field. We have retrieved the WhatsApp chat of a friends group. After pre-processing the data, we have applied text analysis techniques and sentimental study approaches using R Studio. We have also analyzed the emoticons and calculated the sentiment score of the group chat.

The rest of the paper is as follows: section 2 contains the literature review, section 3 contains the project details, section 4 depicts the architecture followed, setup used to carry out the experiments is described in section 5 ,our work is concluded in section 6 and section 7 contains some ideas for future work.

## II. Literature Review

Opinion mining, also called sentimental analysis, is a research area that is trending in the big data analysis field [4]. Lots of research has been carried out and still being conducted to analyze the text and sentiments in real-world scenarios like messaging applications, online blogs, product reviews etc [8]. There are also three main approaches for sentiment mining. The first being Machine Learning, second is lexicon-based and the third approach is hybrid. Tools like SentiWordNet, Emoticons etc. can be used for analyzing sentiments [5].

WhatsApp has changed the way people communicate with each other [11]. Researchers have found that the most unpleasant day was Monday and the messages were transferred in more numbers during the afternoon. 750 million users make use of WhatsApp everyday and 20 million users are getting added up every month making it a very popular mobile application. According to a survey, WhatsApp users are mainly between the age of 26 to 35 with the senior citizens being the least. WhatsApp has the feature to send different forms of information like texts, pictures, videos, emojis etc. Among these, the text messages are more in number. There are many communication services available at present. Reports have shown that the users would move to other communication services vendor, if one service stops working, making it a very competitive business [12].

## III. Project Details

### *A. Definition*

Textual analysis is a methodology that involves understanding language, symbols, and/or pictures present in texts to gain information regarding how people make sense of and communicate life experiences. Visual, written, or spoken messages provide cues to ways through which communication may be understood.

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service. Every person shares his or her information on social network sites, messaging application, blogs, product review websites and web-forums.

### *B. Specification*

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate record from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleaning may be performed interactively with data wrangling tools, or as batch processing through scripting.

- Messages per day: This allows us to see popular periods for messages over the time period of the year.
- Number of messages : This shows who sent the most messages on the chat throughout the year. Some people clearly contributed more than others.
- Most common words by author : This shows the words used most frequently by each author
- Important words used by each author : This snippet of code finds words that are common within the messages of one author but uncommon in the rest of the messages.
- Most used emojis: This shows the emojis used most frequently by each author
- Lexical Diversity : Let's calculate lexical diversity. Basically, you just check how many unique words are used by an author
- Unique words of each author : The lexical diversity plot indicates that author uses the most unique words and what words these are.
- Sentiment analysis : In Sentiment analysis, we had classified the messages into three main classes such as positive , negative and neutral based on bag of word model and Syuzhet model .

## IV. Architecture

In order to make one singular effective application, we created an application in several continuous steps. After exporting data from Whatsapp, we used rwhatsapp library to create a data frame. Then it requires several cleaning processes in order to get better results. After cleaning, we apply some methods of text mining and sentiment analysis. Figure 1 displays the flow diagram of our application.

### *A. Data cleaning*

We collect the data from whatsapp using the extract feature which is provided by whatsapp that data file ill be in formed of .txt or .zip. Here, we had chosen .txt form of data and load that into rwhatsapp library which is provided by Rstudio for whatsapp data analysis. rwhatsapp is a small yet robust package that provides some infrastructure to work with WhatsApp text data in R. rwhatsapp include function such as emoji function which List all emojis and corresponding descriptions and rwaread which Read WhatsApp history into R and takes a history file from the "WhatsApp" messenger app (txt or zip) and returns a formatted data.frame with descriptions of the used emojis.

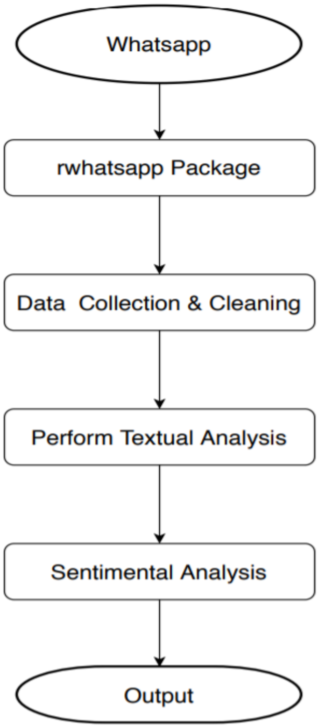

Fig. 1. Flow Diagram

### *B. Textual analysis*

In rwa_read function, we had passed the arguments like x : Path to a txt or zip file of a WhatsApp history or the history itself as character object and tz : A time zone for date conversion. Set NULL or "" for the default time zone or a single string with a timezone identifier. And third argument will be the Format : Most formats are automatically detected.

After loading the data, We need to perform data cleaning on the data to remove the data and inappropriate Messages . WhatsApp seems to become increasingly important not just as a messaging service but also as a social network—thanks to its group chat capabilities. To clean the data we have to remove the NULL values and inappropriate messages such as "This message was deleted, You deleted this message and Changed the subject by you or author".

- dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges:

- mutate() adds new variables that are functions of existing variables

- select() picks variables based on their names.

- filter() picks cases based on their values.
- summarise() reduces multiple values down to a single summary.
- arrange() changes the ordering of the rows.

These all combine naturally with group_by() which allows you to perform any operation “by group” in our project we had use group by authors in text mining. ggplot2 is a library for declaratively creating graphics, based on The Grammar of Graphics. we provide the data and tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details in most cases you start with ggplot(), supply a dataset and aesthetic mapping (with aes()). we then add on layers (like geompoint() or geom_histogram()), scales (like scalecolourbrewer()), faceting specifications (like facetwrap()) and coordinate systems (like coordflip()).

- ggplot() is used to construct the initial plot object, and is almost always followed by + to add component to the plot. ggplot(df, aes(x, y, other aesthetics)) this method is recommended if all layers use the same data and the same set of aesthetics, although this method can also be used to add a layer using data from another data frame and ggtitle function used to represent the title of the plot.
- Coord flip() Flipped cartesian coordinates so that horizontal becomes vertical, and vertical, horizontal. This is primarily useful for converting geoms and statistics which display y conditional on x, to x conditional on y.
- ggimage: Supports image files and graphic objects to be visualized in ’ggplot2’ graphic system.
- theme() : Themes are a powerful way to customize the non-data components of your plots: i.e. titles, labels, fonts, background, gridlines, and legends. Themes can be used to give plots a consistent customized look. Modify a single plot’s theme using theme().
- tidytext() : This package implements tidy data principles to make many text mining tasks easier, more effective, and consistent with tools already in wide use.

### C. *Sentimental analysis*

In Sentiment analysis, we had used bag of words model and Syuzhet Model to analyse whatsapp Message for each group and distributed messages into positive and negative and neutral message.

Syuzhet Model : Here, chat is taken as a sentence and assigned a positive or negative score based on the total score of all the words in it. As usual, we will start with some cleaning to remove html links, punctuations and non-alphanumeric characters like emojis. In the case of Bag of Words model does not capture the complete expression of the author’s emotions.

Syuzhet package : Extracts sentiment and sentiment-derived plot arcs from text using a variety of sentiment dictionaries conveniently packaged for consumption by R users. Implemented dictionaries include ”syuzhet” (default) developed in the Nebraska Literary Lab ”afinn” developed by Finn.

RSentiment package : Analyses sentiment of a sentence in English and assigns score to it. It can classify sentences to the following categories of sentiments:- Positive, Negative, very Positive, very negative, Neutral. For a vector of sentences, it counts the number of sentences in each category of sentiment.In calculating the score, negation and various degrees of adjectives are taken into consideration. It deals only with English sentences.

## V. Experimental setup

In this project, we used the Rstudio IDE Version 1.2.5033. We used the different inbuilt libraries to visualize and analysis of data.

### A. *Implementation Details*

We used rwhatsapp as a major library to analyze text. In addition to that, we also used the dplyr, tidyr, ggplot2, lubridate, tidyr, gganimate, ggimage, stopswords, rsentiments, tidytext, wordcloud, and syuzhet to perform a certain task in the project. In the table I displays about all libraries in detail.

TABLE I
Table to test captions and labels

| Name | Description | Version |
|---|---|---|
| dplyr | A Grammar of Data Manipulation | 0.8.4 |
| ggimage | Use image in ggplot2 | 0.2.7 |
| gganimate | A Grammar of Animated Graphics | 1.0.5 |
| graphics | The R graphics package | 3.6.3 |
| lubridate | Dealing with dates | 1.7.4 |
| RSentiments | Analyse Sentiment of English Sentences | 2.2.2 |
| rwhastapp | Handling WhatsApp chat | 0.2.1 |
| stopwords | Stop word list | 1.0 |
| syuzhet | Extract Sentiment and Sentiment Derived plot arcs from text | 1.0.4 |
| tidyr | Tidy messy data | 1.0.2 |
| tm | Text mining package | 0.7-7 |
| wordcloud | Sentiment analysis library | 2.6 |

### B. *Testing*

In order to test our code, we used three different datasets from our own WhatsApp group. Table II displays the detail of each dataset.

TABLE II
Dataset Detail

| WhatsApp Group Name | Number of Group Members | Number of Messages |
|---|---|---|
| Always Worthy | 4 | 1939 |
| LD ME | 25 | 1483 |
| Geek Club | 13 | 5399 |

### C. *Discussion of your findings, and challenges*

During our project implementation, we found lots of challenges. First, there is no way to get the direct database from WhatsApp because of end to end encryption feature. We extracted our own chat. The first main challenge was how to clean data. We removed all null values and then we removed the system generated field which is not related to chat like

”This message was deleted”, ”You deleted this message”, ”Changed the subject”, etcetera. Another problem is rwhatsapp library considers every word is an author which ended with ’:’. So, we removed every colon from the chat. After this, the timestamp is different in several countries. We also managed the problem of timestamp.

We found some interesting results in our project. First, this algorithm is not identical to a person who sends extremely few messages like less than 10 messages. Still, our algorithm can perform some analysis on that text analysis. We also found that the number of messages is increasing drastically during some specific time in each group. It could be exam time or festival time or some popular movie or tv series launch time. Our application is also capable to find who sends more emojis or who send more links. In the Unique word section, we found that our application is able to detect the wrong word. In some datasets, if a person uses the wrong spelling of a particular word or short form, then it indicates a unique word. Figure 2 display such a result, when one user often writes the wrong spelling of practical. We also found that more number of messages is not guaranteed that the user will achieve more lexical diversity.

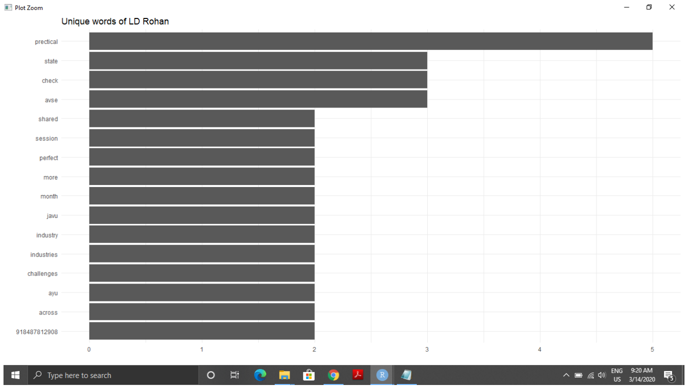

Fig. 2. Unique Words

TABLE III
PROPORTION OF SENTIMENTS IN DIFFERENT DATASET

| Dataset | Total Positive | Positive (%) | Total Negative | Negative (%) | Toatl Neutral | Neutral (%) |
|---|---|---|---|---|---|---|
| Geek Club | 230 | 7.7% | 210 | 7% | 2550 | 85.3% |
| Always Worthy | 104 | 10.9% | 49 | 5.1% | 801 | 84% |
| LD ME | 68 | 5.7% | 29 | 2.4% | 1093 | 91.8% |

We found that most of the messages are neutral in all datasets. Table III displays the result of all three datasets. With some more testing, we got some error in bag of words model because bag of words model work on each every word. That’s why we also used syuzhet model which can distinguish sentiments in ten different classes of emergence namely Positive, Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust and Negative. Figure 3 displays the result of all three datasets.

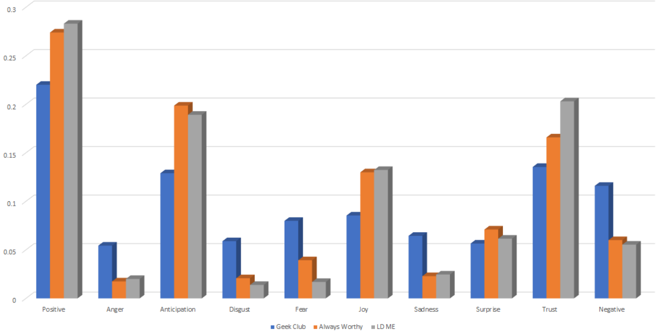

Fig. 3. Flow Diagram

## VI. Conclusion

WhatsApp is one of the most popular messaging applications in this era with 1.2 billion monthly active users in 2017 [14]. That’s why it’s necessary to analyze data. To analyze chat, we used R language because of lots of available good visualization libraries. The primary library which we use is rwhatsapp. First, we clean the data, then we perform some analysis like the number of messages, who send the most number of messages, the most number of emojis, most often used words by user, important or unique words used by a particular author, lexical diversity of author and sentiment analysis. we found some interesting result. This application provides user insights into their communication.

## VII. Future work

This application is currently suitable for WhatsApp only. But with little modification, it can be work in any messaging app which has an export chat feature. The other thing which we want to add is to the analysis media file including images. Currently, it’s application in the form of the console. We also want to create this application in a more user-friendly format where the user can export their chat and get a detailed analysis of data.

## References

[1] Seufert, M., Hoßfeld, T., Schwind, A., Burger, V., Tran-Gia, P. (2016, May). Group-based communication in WhatsApp. In 2016 IFIP networking conference (IFIP networking) and workshops (pp. 536-541). IEEE.
[2] Thakkar, H., Patel, D. (2015). Approaches for sentiment analysis on twitter: A state-of-art study. arXiv preprint arXiv:1512.01043.
[3] Torgo, L. (2016). Data mining with R: learning with case studies. CRC press.
[4] Alessia, D., Ferri, F., Grifoni, P., Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3).
[5] Fang, X., Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 5.
[6] Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 30(4), 330-338.

[7] Joshi, S. (2019). Sentiment Analysis on WhatsApp Group Chat Using R. In Data, Engineering and Applications (pp. 47-55). Springer, Singapore.

[8] Jotheeswaran, J., Koteeswaran, S. (2015). Sentiment analysis: a survey of current research and techniques. Int. J. Innov. Res. Comput. Commun. Eng, 3(5).

[9] Piryani, R., Madhavi, D., Singh, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. Information Processing Management, 53(1), 122-150.

[10] Medhat, W., Hassan, A., Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.

[11] Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., By, T. (2012, August). Sentiment analysis on social media. In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 919-926). IEEE.

[12] Pradhan, V. M., Vala, J., Balani, P. (2016). A survey on Sentiment Analysis Algorithms for opinion mining. International Journal of Computer Applications, 133(9), 7-11.

[13] Joshi, S. (2019). Sentiment Analysis on WhatsApp Group Chat Using R. In Data, Engineering and Applications (pp. 47-55). Springer, Singapore.

[14] Schwind and M. Seufert, ”WhatsAnalyzer: A Tool for Collecting and Analyzing WhatsApp Mobile Messaging Communication Data,”2018 30th International Teletraffic Congress (ITC 30), Vienna, 2018, pp. 85-88. doi: 10.1109/ITC30.2018.00020 URL:http://ieeexplore.ieee.org.ledproxy2.uwindsor.ca/stamp/stamp.jsp?tp =/arnumber=84930 58/isnumber=8493038