WhatsApp Chat Analyzer

Atharv Tiwari[1], Anshika Gangrade[2], Dr. Pritika Bahad[3], Diksha Bharawa[4]
Student of Department of Artificial Intelligence and Data Science, PIEMR Indore, Madhya Pradesh[1]
Student of Department of Artificial Intelligence and Data Science, PIEMR Indore, Madhya Pradesh[2]
Assistant Professor Artificial Intelligence and Data Science, PIEMR Indore, Madhya Pradesh[3,4]

Abstract: *The most used and efficient method of communication in recent times is an application called WhatsApp. WhatsApp chats consist of various kinds of conversations held among groups of people. This chat consists of various topics. This information can provide lots of data for the latest technologies such as machine learning. The most important thing for machine learning models is to provide the right learning experience which is indirectly affected by the data that we provide to the model. This tool aims to provide in depth analysis of this data which is provided by WhatsApp. Irrespective of whichever topic the conversation is based on, our developed code can be applied to obtain a better understanding of the data. The advantage of this tool is that it is implemented using simple python modules such as pandas, matplotlib, seaborn and sentiment analysis which are used to create data frames and plot different graphs, where then it is displayed in the flutter application which is efficient and less resources consuming algorithm, therefor it can be easily applied to largest dataset.*

Keywords: *WhatsApp chat data, Pandas, Seaborn, matplotlib, sentiment analyzer etc.*

1. Introduction

This tool is based on data analysis and processing. The first step in implementing a machine learning algorithm is to understand the right learning experience from which the model starts improving on. Data pre-processing plays a major role when it comes to machine learning. In order to make the model more efficient we need lots of data, so we turned our focus primarily on one of the largescale data producers owned by Facebook which is nothing but WhatsApp. WhatsApp claims that nearly 55 billion messages are sent each day. The average user spends 195 minutes per week on WhatsApp, and is a member of plenty of groups. With this treasure house of data right under our very noses, it is but imperative that we embark on a mission to gain insights on the messages which our phones are forced to bear witness to.

1.1 Problem Statement

WhatsApp-Analyzer is a statistical analysis tool for WhatsApp chats. Working on the chat files that can be exported from WhatsApp it generates various plots showing, for example, which another participant a user responds to the most. We propose to employ dataset manipulation techniques to have a better understanding of WhatsApp chat present in our phones.

1.2 Existing System

There is a lot of development in the current system. In the older version there was no feature to display status, there was no feature to share documents and there was no feature to share location. In the current version, all of these features are available. In older version we couldn’t share images through doc’s format. In this system user is able to access WhatsApp in windows through WhatsApp web application, which can be connected through QR code. There is another feature called export chat where user can send or share or get the chat detail for data analysis through email, Facebook or some messenger application.

1.3 Proposed System

Data pre-processing, the initial part of the project is to understand implementation and usage of various python-built modules. The above process helps us to understand why different modules are helpful rather than implementing those functions from scratch by the developer. These various modules provide better code representation and user understandability. The following libraries are used such as numpy, scipy pandas, csv, sklearn, matplotlib, sys, re, emoji, nltk seaborn etc.

Exploratory data analysis, first step in this to apply a sentiment analysis algorithm which provides positives negative and neutral part of th chat and is used to plot pie chart based on these parameters. To plot a line graph which shows author and message count of each date, to plot a line graph which shows author and message count of each author, Ordered graph of date vs message count, media sent by authors and their count, Display the message which is di not have authors, plot graph of hour vs message count.

1.4 Objective

In this decade the upcoming technologies are mainly dependent on data. This data can only be obtained if there is some research applied on the context of the requirements of the tool. Since a lot of machine learning enthusiasts develop models which helps solve multiple problems the requirements of appropriate data are very large scale this project aims to provide a better understanding towards various types of chats. This analysis proves to be better input to machine learning models which essentially explore the chat data. These models require proper learning instances which provides better accuracy for these models. Our project ensures to provide an in-depth exploratory data analysis on various types of WhatsApp chats

2. Literature Review

As a demo Survey analysis on the usage and Impact of WhatsApp Messenger [1]: Various Studies and analysis has been done on the usage and impact of WhatsApp. Some of these studies are for finding the impact of WhatsApp on the students and some are based on for the general public in a local region.

In a study of southern part of India was conducted on the age group of between 18 to 23 years to investigate the importance of WhatsApp among youth. Though this study, it was found that students spent 8 hours per day on using WhatsApp and remain online almost 16 hours a day. All the respondents agreed that they are using WhatsApp for communicating with their friends. They also exchange images, audio and video files with their friends using WhatsApp. It was also proved that the only application that the youth uses when they are spending time on their smart phone is WhatsApp. Methods used in this survey is to analyze the intensity of WhatsApp usage and its popular services and to identify the degree of positive or negative impacts of using WhatsApp.

Content Analysis of WhatsApp conversation [2]: An analytical study to evaluate the Effectiveness of WhatsApp Application in India. The Study will be an important research work for exploring the possibilities of emergence of WhatsApp as the leading mobile messaging application in India.

With the advancement of digital technology and the mergence of mobile phones in India the communication scenario has completely changed. The increasing trend of smart phones and social networking applications in India has made communication faster and easier than at any time in history. Now, people may not have enough money to eat, enough place to sleep and enough dress to wear but have mobile phone in their pockets to interact with their family members, friends and customers. With the changing scenario, use of quantitative and qualitative research techniques has also increased with the passage of time. Procedures were devised for the measurement of nature and effect of communication devises on human behavior. During the same period, smart phones and instant messaging application like WhatsApp, Viber and Skype took over the world of communication in India.

WhatsApp Group Data Analysis with R [3]: The dataset of WhatsApp group chat used for analysis is of 1 year(may, 2015-may,2016) which consists of 5,5563 records in total and comprises of certain characteristics that define how much a particular person is using WhatsApp chat group, such as the years of usage, duration of usage in a day, the response levels, type of messages posted by each individual in the group (Smiley, Text, Multitude), which age group people are more active and so on.

The main attributes set for this analysis are type of message been send, duration of use per year/month/week/day/hour, timestamp (AM/PM), age group of senders, gender (Male/Female). RStudio the most favored IDE for R is been used to perform exploratory data analysis and visualization for the collected data largely because of its open source nature.

Forensic analysis of WhatsApp Messenger [4]: WhatsApp provides its users with various forms of communications, namely user-to-user communications, broadcast messages, and group chats. When communicating, users may exchange plain text messages, as well as multimedia files (containing images, audio, and video), contact cards, and geolocation information. Each user is associated with its profile, set of information that includes his/her WhatsApp name, status line, and avatar (a graphic file, typically a picture). The profile of each user is stored on a central system, from which it is downloaded by other WhatsApp users that include that user in their contacts. The central systems provide also other services, like user registration, authentication, and message relay.

3. Software Requirement Analysis

If software requirement analysis in the field of systems engineering and software engineering, encompasses those tasks that are used for a new or altered product or tool, taking account of the possibly conflicting requirements of the various stakeholders, documenting, validating and managing software or system requirements.

3.1 Feasibility Study

The main objective of the feasibility study is to treat the technical operational and economic feasibility of developing the application. Feasibility is the determination of whether or not project is worth doing. The process followed in making this determination is called feasibility study. All systems are feasible, given unlimited resources and infinite time. The feasibility study to be conducted for this project involves:

- Technical Feasibility
- Operational Feasibility
- Economic Feasibility

3.1.1 Technical Feasibility

It is the measure of the specific technical solution and the availability of technical resources and expertise. It is one of the first studies that must be conducted after the tool has been identified. A technical study of feasibility is an assessment of the logistical aspects of business operation. This is considered with specifying equipment and software that will successfully satisfy the user’s requirements. The technical needs of the system may vary considerably but should include the facility to produce outputs in a given time, response time under certain conditions and the ability to process a certain number of transactions at a certain speed.

The proposed system is developed by using VS Code software. VS Code is a non-profit organization created to develop open-source software, open standards, and services for interactive computing across dozens of programming languages. The idea is to implement a data processing code using python to make better sense of WhatsApp group chat data.

3.1.2 Operational Feasibility

Operational feasibility is mainly concerned with issues like whether the system will be used if it is developed and implemented, whether there will be resistance from the users which will affect the possible application benefits. It is the ability to utilize, support and perform the necessary tasks of a system or program. It includes everyone who creates, operates or uses the system or program. It is the measure of how well a proposed system solves the problem and takes advantages of the opportunities identified during the scope definition and problem analysis phases. This system helps in many ways. It shows the number of users using WhatsApp and gives the data information of their sharing data. Which is organized in Pie-chart and Bar- chart.

3.1.3 Economic Feasibility

Economic feasibility is the most frequently used method for evaluating the effectiveness of the new system. Economic feasibility is the measure of the cost effectiveness of an information system solution. Without a doubt, this measure is most often and important one of the three. Information systems are often viewed as capital investments for the business, and, as such should be subjected to the same type of investment analysis as other capital investments.

Economic analysis is used for evaluating the effectiveness of the proposed system. In economic feasibility, the most important is cost-benefit analysis. This project is not economical as it mainly depends on the sharing of data between two phones

4. System Implementation

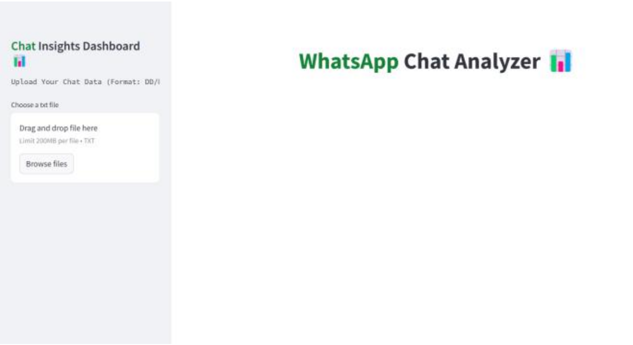
Python: It is an interpreted, high-level general-purpose programming language. Created by Guido Van Rossum and first released in 1991. Its language constructs and objects-oriented approach aim to help programmer with clear, logical code for small and large-scale tools. Python is used for web development (server-side), software development, mathematics, it can be used alongside software to create workflows, it can connect to database systems, it can also read and modify files, it can be used to handle big data and perform complex mathematics and can be used for rapid prototyping, or for production-ready software development.

JSON: Java Script Object Notation is an open standard file format, and data interchange format, that uses human-readable text to store and transmit data objects consisting of attribute-value pairs and array data types. It is very common data format, with diverse range of applications. Such as serving as a replacement for xml in ajax systems. Json is a language-independent data format. It was derived from JavaScript, but many modern programming languages include code to generate and parse JSON-format data. The official Internet media type for Json is application/json. Json filenames use the extension (.json). When exchanging data between a browser and a server, the data can only be text. Json is text, and we can convert any JavaScript object into json and json to the server. We can also convert any json received from the server into JavaScript objects. This way we work with the data as JavaScript objects, with no complicated parsing and transactions.

DART: It is a client-Optimized programming language for apps on multiple platforms. It is developed by google and is used top build mobile, desktop, server, and web applications. Dart is an object-oriented, class-based, garbage-collected language with C-style syntax. Dart can complete to either native code or JavaScript. It supports interfaces, mix-ins, abstract-classes, refined generics and type inference. To run in mainstream web browsers, Dart relies on source-to-source compiler to JavaScript. According to the tool site. Dart was "designed to be easy to write development tools for, well-suited to modern app development, and capable of high-performance implementations". When running dart code in a web browser the code is precompiled into JavaScript using dart2.js compiler. Compiled as JavaScript, Dart code is compatible with all major browsers with no need for browsers to adapt dart. Though optimizing the compiled JavaScript output to avoid expensive checks operations, code written in dart can, in some cases, run faster than equivalent code hand-written using JavaScript idioms.

5. Result Analysis

The results of this work, showed several activities on specific dates as specified by the system at given time. The results showed that the most active date was 15th June, 2018. The number of messages sent on that most active date was 190. Also, the overall, most active user was recorded and the user was shown to have posted over 972 messages to the group. The system also recorded the list of active authors, emojis etc. Furthermore, the total number of users on that group was shown to be 230. In addition, a full list of all users on the platform with their name or their phone number were also outputted plus the number of times each individual on the platform has made a post. And the most used word was also given as "the" and it was used 43313 times. Below fig. shows a snap shot of the screen that shows the output of the analysis.



Chat Insights Dashboard 📊

Upload Your Chat Data (Format: DD/I

Choose a txt file

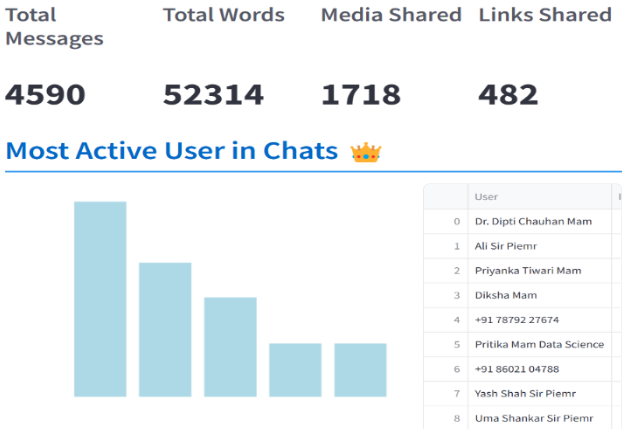
Drag and drop file here
Limit 200MB per file • TXT

Browse files

WhatsApp Chat Analyzer 📊

Fig 1: Sample output of WhatsApp plot

The following represents the output of the result of the analysis done with Python on the given group chat.



Total Messages	Total Words	Media Shared	Links Shared
4590	52314	1718	482

Most Active User in Chats 👑

	User
0	Dr. Dipti Chauhan Mam
1	Ali Sir Piemr
2	Priyanka Tiwari Mam
3	Diksha Mam
4	+91 78792 27674
5	Pritika Mam Data Science
6	+91 86021 04788
7	Yash Shah Sir Piemr
8	Uma Shankar Sir Piemr

Fig 2: Top 10 active users

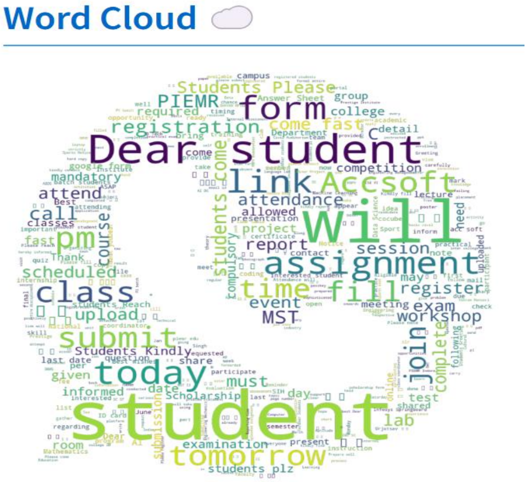


Fig. 3 : most repeated messages

Fig. 3 on the other hand shows the most repeated messages in the group. This gives a narrower scope of comparison.

As explained for fig 2, fig 3 also has some percentage and some with names. This figure clearly gives an analysis of the top 10 users on the group chart, showing them in percentages.

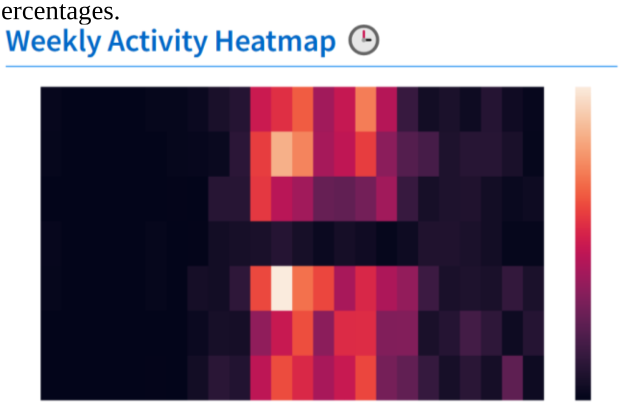


Fig 4: weekly heatmap of chats and usage

6. Conclusion

In conclusion, it can be said that the capabilities of the WhatsApp application and the power of the python programming language in implementing whatever network data analysis intended, cannot be overemphasized. This work was able to discuss the WhatsApp application and its libraries, to create an analysis of a WhatsApp group chat and visually represent the top 10 and top 20 users in the chat groups. A pseudocode of the plot was given and at the end, visual representation of the plot was implemented. Also, an analysis of the top 10 and top 20 users were done. The system was done with python, and the python libraries that were implemented includes, NumPy, Pandas, Matplotlib and Seaborn. At the end of the work expected results were obtained and the analysis was able to show the level of participation of the various individuals on the given WhatsApp group. On serious note this system has the ability to analyze any WhatsApp group data input into it.

Reference

[1] Available from: http://www. statista.com/statistics/260819/number- of-monthly-active-WhatsApp-users. Number of monthly active WhatsApp users worldwide from April 2013 to February 2016(in millions).

[2] Ahmed, I., Fiaz, T., "Mobile phone to youngsters: Necessity or addiction", African Journal of Business Management Vol.5 (32), pp. 12512-12519, Aijaz, K. (2011).

[3] Aharony, N., T., G., The Importance of the WhatsApp Family Group: An Exploratory Analysis. "Aslib Journal of Information Management, Vol. 68, Issue 2, pp.1-37" (2016).

[4] Access Data Corporation. FTK Imager, 2013. Available at http://www.accessdata.com/support/product-downloads.

[5] Singh, Harsh Pratap, et al. "AVATRY: Virtual Fitting Room Solution." 2024 2nd International Conference on Computer, Communication and Control (IC4). IEEE, 2024.

[6] Singh, Nagendra, et al. "Blockchain Cloud Computing: Comparative study on DDoS, MITM and SQL Injection Attack." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[7] Singh, Harsh Pratap, et al. "Logistic Regression based Sentiment Analysis System: Rectify." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[8] Naiyer, Vaseem, Jitendra Sheetlani, and Harsh Pratap Singh. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.

[9] Pasha, Shaik Imran, and Harsh Pratap Singh. "A Novel Model Proposal Using Association Rule Based Data Mining Techniques for Indian Stock Market Analysis." Annals of the Romanian Society for Cell Biology (2021): 9394-9399.

[10] Md, Abdul Rasool, Harsh Pratap Singh, and K. Nagi Reddy. "Data Mining Approaches to Identify Spontaneous Homeopathic Syndrome Treatment." Annals of the Romanian Society for Cell Biology (2021): 3275-3286.

[11] D.Radha, R. Jayaparvathy, D. Yamini, "Analysis on Social Media Addiction using Data Mining Technique", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, pp. 23-26, April 2016.

[12] Jessica Ho, Ping Ji, Weifang Chen, Raymond Hsieh, "Identifying google talk", IEEE International Conference on Intelligence and Security Informatics, ISI '09, pp. 285-290, 2009.

[13] Mike Dickson, "An examination into AOL instant messenger 5.5 contact identification.", Digital Investigation, ScienceDirect, vol. 3, issue 4, pp. 227-237, 2006.

[14] Mike Dickson, "An examination into yahoo messenger 7.0 contact identification", Digital Investigation, ScienceDirect, vol. 3, issue 3, pp. 159-165, 2006.