International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 635-640.
www.theijire.com **ISSN No: 2582-8746**

WhatsApp Chat Analysis Based on NLP Using Machine Learning

Dhanashri Hase[1], Junaid Khan[2], Sahil Khot[3], Rehaan Qureshi [4], Firoz Shaikh[5]
[1]*Assistant Professor, Department of Computer Engineering, Armiet, Maharashtra, India.*
[2,3,4,5] *Students, Department of Computer Engineering, Armiet, Maharashtra, India.*

How to cite this paper:
Dhanashri Hase[1], Junaid Khan[2], Sahil Khot[3], Rehaan Qureshi [4], Firoz Shaikh[5], "Whatsapp Chat Analysis Based on NLP UsingMachine Learning", IJIRE-V4I02-635-640.

https://www.doi.org/10.59256/ijire.2023040238

Abstract: *A program called WhatsApp has emerged as the most popular and effective means of communication in recent years. The program, called WhatsApp Chat Analyzer, is set up on Heroku Web and offers analysis of WhatsApp groups. Although there are many other approaches for analysis,matplotlib, stream lit, sea born, re, panda's libraries of Python, and certain NLP concepts are employed here. This is the fusion of NLP with machine learning. This WhatsApp conversation analyzer imports a user's WhatsApp chat file, analyses it, and outputs several visualization's. I've suggested a WhatsApp Chat Analyzer in this report. Different forms of communication between groups and individuals are included in WhatsApp chats. Different subjects are covered in this talk. This could give machine learning technology more data to work with. The correct learning experience is offered by machine learning models, which is a crucial factor that is indirectly impacted by the data supplied to that model. This program offers analysis of the information that WhatsApp provides. This application has the benefit of being implemented by straightforward python libraries, such as sea born, pandas, numpy, stream lit, and matplotlib, which are frequently used to build data frames and other graphs. This is shown on the web using a heroku link, which is accessible from any device with a browser.*

Key Word: *WhatsApp Chat, Python, Stream lit, Analysis, Nature Language Processing, Emoji, Pandas, Matplotlib.*

I.INTRODUCTION

Depending on the project's objectives and specific environment, WhatsApp chat analysis projects can differ. The overall goal of a WhatsApp chat analysis project is to learn more about how people communicate and then utilize this knowledge to accomplish a certain goal, such as enhancing communication, identifying fraud, or comprehending group dynamics. This project will raise awareness about how to use WhatsApp properly in academic and corporate settings to address new economic concernsin India. The business community in our country is still very far behind in understanding the beneficial role that WhatsApp is playing in emerging nations. The study will not only describe how WhatsApp is used in the commercial and academic worlds, but it will also promote WhatsApp use in a variety of public and private organization's in a sustainable way. When it comes to using WhatsApp to address new difficulties, the next five to ten years will be crucial. Educating people on how to use WhatsApp properly through training, seminars, and workshops can help achieve the goal. The study also lays a solid framework for future research to understand how we can use it more effectively. India may effectively utilize WhatsApp as a cost-effective tool in a variety of organizations by holding various training sessions with careful planning. This study focuses on the phenomena of how many students and working adults use WhatsApp and sheds insight on how important it is to their everyday lives. This study will contribute to the development of a plan for integrating WhatsApp into the academic and professional groups.

II.LITERATURE REVIEW

Various research and analyses have been found as a result of a survey analysis on the impact and usage of WhatsApp Messenger [1]. These studies cover WhatsApp's effects on students and young people. According to the survey, people between the ages of 18and 23 in the southern region of India use WhatsApp for roughly 8 hours per day and are occasionally online for up to 12 to 16 hours per day. The majority of them acknowledged that WhatsApp was their preferred website. They share videos, audio, and photos. This study also demonstrated that WhatsApp is the program that is used on smart phones the most, compared to all other apps. Since WhatsApp is the most popular app among young people and other generations, our project can give them insights into their discussions and reveal unknown truths to them. This study was done to determine WhatsApp usage's benefits anddrawbacks. As this survey has revealed, of course.

Author	Title & Journal Name	Paper Outcomes	Limitation
Barbosa et al.	"WhatsApp Chats in the Context of Health Professionals: Analysis of Characteristics andPossibilities"	Found that WhatsApp is widely used by health professionals for communication and coordination, but also identified potential risks to patient privacy and confidentiality.	The study focused solely on health professionals and did not include patient perspectives or experiences.

Chen and Chen	"A Study on the Analysis of Group Chat in WhatsApp"	Examined the linguistic features of group chats on WhatsApp and identified patterns in language use and communication behavior.	The study was limited to a small sample size of 20 participants and may not be representativeof larger populations.
Ghosh and Kar	"Analysis of WhatsApp Chat for Abusive Language Detection"	Developed a machine learning model to automatically identify abusive language in WhatsApp chats, achieving an accuracy rate of 86%.	The study was limited to identifying only one type of problematic language and maynot generalize to other types of problematic behavior.
Hassan et al.	"A System for Identifying Fake News in WhatsApp Groups"	Developed a system to automatically detect fake news in WhatsApp groups, achieving a precision of 90% and a recallof 85%.	The study was limited to identifying only one type of misinformation and may not generalize to other types of misinformation.
Majumder and Biswas	"Exploratory Analysis of WhatsApp Group Chat: A Study of Interpersonal Relationship, Group Dynamics and Interactions"	Analyzed the dynamics of WhatsApp group chats and identified patterns of social interaction and communication.	The study was limited to a small sample size of 40 participants and may not be representative of larger populations or groups with different characteristics.
Palacios et al.	"WhatsApp Chats: A Goldmine for Research?"	Explored the potential of WhatsApp chatsas a source of data for social research and identified the advantages and limitations of using this platform.	The study did not conduct any specific analysis of WhatsApp chats but provided a general overview of their potential as a research tool.
Purwarianti et al.	"Sentiment Analysis on Indonesian WhatsApp Group Conversations Using Hybrid Method of Naïve Bayes and Support Vector Machine"	Developed a machine learning model to analyze the sentiment of WhatsApp group chats in Indonesian, achieving an accuracy rate of 87%.	The study was limited to analyzing only one language and may not generalize to other languages or cultures.
Shahzad et al.	"WhatsApp Group Chats: Analyzing the Role of Social Capital in Information Sharing"	Examined the role of social capital in WhatsApp group chats and identified factors that contribute to the formation of strong ties and social connections.	The study was limited to a small sample size of 31 participants and may not be representative of larger populations or groups with different characteristics.

III. PROPOSED SYSTEM

The proposed system would gather information on WhatsApp chats using automated data extraction methods. Instead than asking users to manually copy and paste chat logs, this would entail utilizing specialized software or algorithms to extract the data directly from WhatsApp servers. This strategy would guarantee that the data is gathered consistently and dependably and would lessen the possibility of biases or errors. The suggested system would use sophisticated data cleaning and pre-processing techniques to prepare the WhatsApp chat data for analysis after it has been extracted. Duplicate data would need to be removed, missing values would need to be handled, and the data would need to be transformed into a structured format for analysis.. The proposed system would make sure that the data is of the highest quality and prepared for more sophisticated analysis by employingcutting-edge data cleaning and pre-processing techniques. It would mine the WhatsApp chat data using sophisticated analysis methods to glean insights and patterns. This would entail utilizing machine learning algorithms to find trends and patterns in the data as well as natural language processing (NLP) methods to recognize keywords and the emotional undertone of the chat. The proposed system would be able to offer more precise and in-depth insights about user behavior and preferences by utilizing advanced analysis techniques. Users would be able to examine the data and see the analysis' findings thanks to interactive visualization and reporting tools included in it. To do this, interactive dashboards and reports that enable users to drill down into and study the data in greater detail would need to be created. The proposed system would facilitate users' comprehension and interpretation of the outcomes of the study by supplying more interactive and user-friendly visualization and reporting tools. The suggested system for WhatsApp Chat Analysis comes with a number of crucial components that work to increase the analysis process' speed, precision, and scalability. The suggested system would be able to offer more precise and in-depth insights into userbehavior and preferences by utilizing automated data extraction, sophisticated analysis techniques, interactive visualization, and reporting. The suggested approach will also increase user confidence and guarantee that the analysis is done in an ethical and responsible manner by prioritizing data privacy and security. For enterprises and researchers, the suggested approach has the potential to dramatically increase the value and usability of WhatsApp Chat Analysis.

1. System Architecture

The study's main goal was to distinguish between churned and kept consumers and to pinpoint the factors that contribute to churn. According to the survey, male consumers who are single had a somewhat higher churn rate. Furthermore, a connection between customer churn and Mobile preferred order category was discovered. Additionally, it was discovered that

churning customers favoured phones/mobile phones slightly more than other devices, which may be related to the platform's phone version's improved user experience for e-commerce customers. The survey also discovered that consumers that have churned have higher averages forcomplaints, city tier, number of residences, and registered devices.

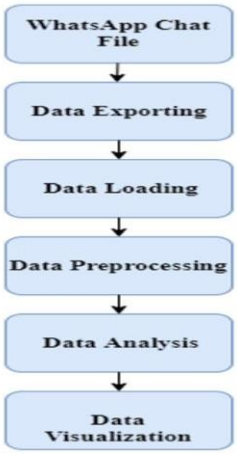
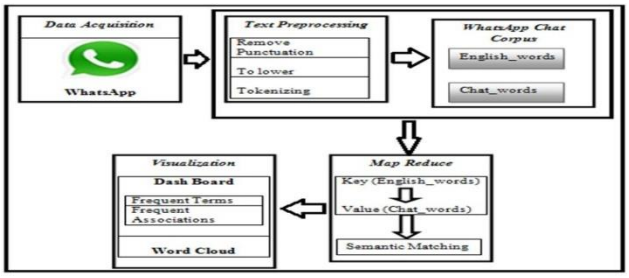


Fig-1: SystemArchitecture

2. Data Flow Diagram



In this diagram, the functionality of the architecture is as follows:

Data Extraction: This part is in charge of obtaining WhatsApp chat data directly from the source. To extract the data straight from WhatsApp servers, it could include either human extraction or automatic extraction using specialised software or algorithms. In a database, the extracted data is then kept for later processing.

Data Pre-processing and Cleaning: This step entails cleaning the extracted data and getting it ready for analysis. As part of this, the data may need to be cleaned up, handled for missing values, and transformed into a structured format for analysis.

Natural Language Processing (NLP): For this part, text data from WhatsApp chats will be analysed using NLP methods. Identifying keywords, sentiment analysis to gauge the conversation's emotional intensity, and named entity recognition to pinpoint specific entities referenced in the chat are possible steps in this process.

Machine Learning: In this step, methods for machine learning are used to find patterns and trends in the data. To categorise comparable discussions and find patterns, this may entail clustering and classification algorithms, or predictive models to foretell future trends.

The creation of interactive dashboards and reports that let users examine the data and see the outcomes of the analysis falls under the category of "visualisation and reporting" under this component. This could entail making interactive charts and graphs using data visualisation tools like Tableau, Power BI, or D3.js.

Data Security and Privacy: This part is in charge of making sure that the conversation data is secure and private. To prevent unauthorised access to the data, it might be necessary to apply robust encryption and access controls. It might also be necessary to follow the industry's best practises for data privacy and security.

In order to gather, clean, analyse, and visualise the WhatsApp chat data, a number of interrelated components are used in the architecture flow for the WhatsApp Chat analysis Project. The architecture flow for WhatsApp Chat analysis Project can offer insightful information into user behaviour and preferences, assist businesses and researchers in making better decisions, and provide valuable insights into user behaviour and preferences by utilising advanced analysis techniques like NLP and machine learning and prioritising data privacy and security.

3. Requirement Analysis

Hardware requirements:

For Development we need a machine of following configuration:

- CPU: Corei5 10thGen, 1.2GHz.
- RAM: DDR3 4GB.
- HDD: 256 GB.
- Systems: Monitor, Keyboard, Mouse

Software requirements

- Operating System: Windows 8/10/11.
- Programming Language: Python, JSON.
- Development IDE: Visual Studio Code Version:1.75
- Other Software's: Google Collab, Jupiter Notebook.

IV.RESEARCH AND METHODOLOGY

To transliterate WhatsApp chat text into its original English text within this framework, we created the messaging chat shown inFigure 2. There are four phases in the framework.

- Data Acquisition and Text Pre-Processing
- Creation of WhatsApp Chat Corpus
- Semantic Map Reduce Framework for WhatsApp
- Visualization

A WhatsApp Chat analysis project's methodology can be divided into many crucial steps:

Define the study's goals. Clearly defining the research objectives is the initial stage in any research effort. This entails defining theproject's scope and identifying the important problems that the analysis should address.

Data Gathering: The data from the WhatsApp chats that will be analyzed is collected in the second step. This may entail manuallyextracting the data from WhatsApp servers or automatically extracting the data using specialized software or algorithms. For future analysis, the gathered data should be kept in a safe and convenient location.

Data Cleaning and Pre-processing: The third stage is to clean and pre-process the gathered data in order to get it ready for analysis. To do this, the data may need to be cleaned up, missing values handled, and transformed into a structured format for analysis.

Data Analysis: The fourth phase entails applying cutting-edge analysis methods like NLP and machine learning to examine the WhatsApp chat data. This could entail locating keywords, sentiment analysis to determine the conversation's emotional undertone,and machine learning algorithms to find patterns and trends in the data.

Visualization and Reporting: The creation of interactive dashboards and reports that let users explore the data and see the outcomes of the analysis constitutes the fifth phase. This could entail making interactive charts and graphs using data visualizations tools like Tableau, Power BI, or D3.js.

Data Privacy and Security: The final step is to ensure the privacy and security of the chat data. This may involve implementing strong encryption and access controls to protect the data from unauthorized access, and adhering to best practices for data privacy and security.

In order to gather, clean, analyses, and visualize the WhatsApp chat data, a methodology for a WhatsApp Chat analysis project entails a number of connected phases. The methodology for a WhatsApp Chat analysis project can give important insights into user behavior and preferences by utilizing cutting-edge analysis tools like NLP and machine learning and by prioritizing data protection and security. This can help organizations and researchers make better decisions

V.IMPLEMENTATION AND RESULT

The outcomes of this work demonstrated a number of activities on particular dates as determined by the system at the time. The findings indicated that June 15, 2028, was the most active date. There were 190 messages sent on that day, which was the busiest. Additionally, the individual who was the most active overall was noted; it was shown that this user had contributed over 972 messages to the group. The system also saved information on emoji's, current authors, etc. Additionally, 230 users were shown tobe included in that group as a whole. Along with the quantity of posts each user on the platform has made, a complete list of all users was also outputted, along with their name or phone number. And "the" was the most frequently used word, appearing 43313times in total.

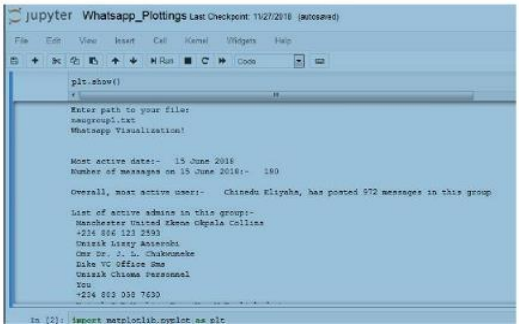


Fig. 5.1 Sample Output of WhatsApp chat plot

Activity Map: It shows the busy days and months. We have used the matplotlib library to plot the graph, the number of messagesin a particular month or day are mapped to the particular day or month.

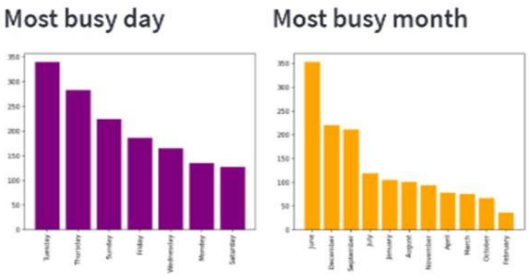


Fig. 5.2 Activity Map

Emoji Analysis: It shows the most commonly used emoji's. We have used the Emoji library to select or distinguish the emoji'sfrom the messages and plotted the pie chart using matplotlib

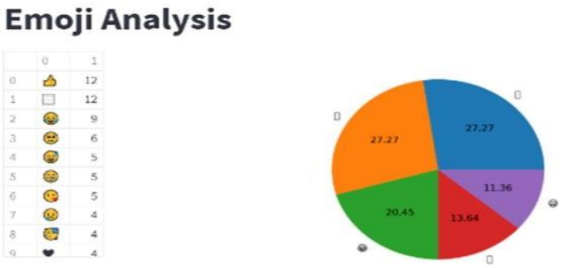


Fig. 5.3 Emoji Analysis

Top Statistics: It shows the most commonly used emoji's we have used the Emoji library to select or distinguish the emoji's fromthe messages and plotted the pie chart using matplotlib

Fig. 6.3.4 Top Statistics

Most Common words: It shows the statistics like total messages, words, and images links shared. We have converted the whole chat file into a data frame and then separated the words and messages and used URL extract to find links.

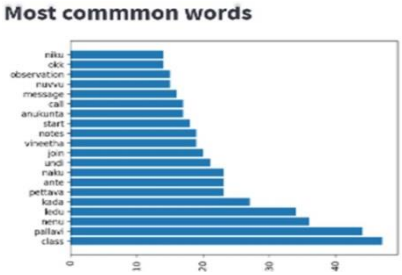


Fig. 6.3.5 Most Common Words

Most Busy Users: It shows the most commonly used word we have used matplotlib to plot the graph and the top frequently used words are displayed. Is shows the busy users and their contribution to chat we have used matplotlib to plot the graph and the users and how frequently the chat is calculated and plotted

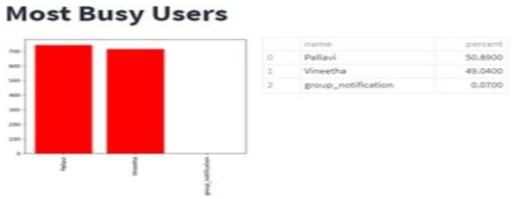


Fig. 6.3.6 Most Busy User

Daily Timeline: It gives the frequency of messages in a day we have used matplotlib to plot the graph and the days are taken and the count of messages are calculated and plotted

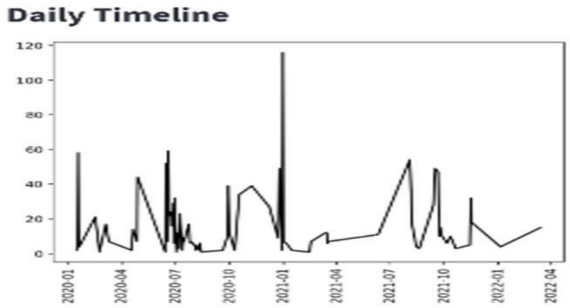


Fig. 6.3.7 Daily Timeline

References

1. *Retrieved from www.statista.com/number-of-monthly-active-whatsapp-users.*
2. *Ravishankara K, Dhanush, Vaisakh, Srajan I S, "International Journal of Engineering Research & Technology(IJERT)", ISSN: 2278-0181, Vol. 9 Issue 05, May-2020*
3. *Dr. D. Lakshminarayanan, S. Prabhakaran, "DogoRangsang Research Journal", UGC Care Group I Journal, Vol-10Issue-07 No. 12 July 2020*
4. *F. Meng Cai, "PubMed Central", PMCID: PMC7944036, PMID: 33732917*
5. *Abdullah, A., Brobst, S, Pervaiz.I., Umer M.,and A.Nisar.2004. Learning dynamics of pesticide abuse through data mining. Proceedings of Australian Workshop on Data mining and Web Intelligence, New Zealand, January, 2004.*
6. *https://www.interaction-design.org/literature/topics/web-design*
7. *Radhika, Narendiran, "Kind of Crops and Small Plants Prediction using IoT with Machine Learning," International Journal of Computer & Mathematical Sciences April 2018, pp. 93-97*
8. *Available from: http://www. statista.com/statistics/260819/numberof-monthly-activeWhatsApp-users. Number of monthly active WhatsApp users worldwide from April 2013 to February 2016(in millions).Journal of Innovative Research in Computer and Communication Engineering January 2017, pp. 318- 323*
9. *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).*
10. *Mike Dickson, "An examination into yahoo messenger 7.0 contact identification", Digital Investigation, ScienceDirect, vol. 3, issue 3, pp. 159-165, 2006.*
11. *Akash Raj N, Balaji Srinivasan, Deepit Abhishek D, SarathJeyavanth J, Vinith Kannan A, "IoT based Agro Automation System using Machine Learning Algorithms", International Journal of Innovative Research in Science, Engineering and Technology November 2016, pp. 19938- 19342*
12. *D Ramesh ,B Vishnu Vardhan. Analysis Of Crop Yield Prediction Using Data Mining Techniques. IJRET: International Journal of Research in Engineering and Technology.(IJERT), 2015.*