Whatsapp Chat Analyzer

[1]Mr. Manvendra Pratap Singh, [2]Nitish Raj, [3]Pushpendra Verma, [4]Ashank Pandey, [5]Anchal Verma

[1]*Faculty of Department of Information Technology, Bansal Institute of Engineering and Technology, Lucknow, India*
[2]*Student of Department of Information Technology, Bansal Institute of Engineering and Technology, Lucknow, India*
[3]*Student of Department of Information Technology, Bansal Institute of Engineering and Technology, Lucknow, India*
[4]*Student of Department of Information Technology, Bansal Institute of Engineering and Technology, Lucknow, India*
[5]*Student of Department of Information Technology, Bansal Institute of Engineering and Technology, Lucknow, India*

—---***---

Abstract - *WhatsApp has emerged as one of the most widely adopted and efficient platforms for digital communication, facilitating a broad spectrum of interactions within individual and group conversations. These chat logs encompass diverse topics and serve as a rich source of unstructured textual data, which holds substantial value for contemporary technologies such as machine learning (ML) and natural language processing (NLP). For machine learning models to perform optimally, the availability of relevant and high-quality training data is paramount, as the learning outcomes are directly influenced by the input data.*

The objective of this research is to develop a tool capable of conducting comprehensive exploratory data analysis (EDA) on WhatsApp chat data. The proposed system is agnostic to the subject matter of the conversation and can be uniformly applied to extract actionable insights from any chat dataset. The implementation leverages widely-usedPython libraries including Pandas, Matplotlib, Seaborn, and NLTK for sentiment analysis. These tools facilitate the transformation of raw textual data into structured formats (e.g., DataFrames), enable statistical visualization, and support sentiment classification. The analytical results are rendered via a Flutter-based front-end application, chosen for its lightweight architecture and efficient resource consumption, thus enabling scalability and applicability to large-scale datasets.

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

INTRODUCTION

The proposed tool is grounded in the principles of data analysis and preprocessing, focusing on extracting structured information from unstructured conversational data. WhatsApp serves as a prolific communication platform, supporting a variety of interactions between individual users and within group settings. A critical prerequisite for developing robust machine learning (ML) algorithms is the availability of relevant and diverse training data, which enables models to learn and generalize effectively. Data preprocessing, therefore, constitutes a foundational stage in the machine learning pipeline.

To enhance model accuracy and generalization, large-scale, real-world data is essential. WhatsApp, a platform owned by Meta (formerly Facebook), presents a valuable opportunity in this regard, with an estimated 55 billion messages exchanged daily. Additionally, users spend approximately 195 minutes per week on the application and often participate in multiple chat groups. Given this enormous volume of communicative data generated continuously, the platform represents a significant and largely untapped source for behavioral and linguistic analysis. This research leverages this opportunity to extract meaningful insights from everyday digital interactions through systematic data mining and analysis techniques.

1.1 PROBLEM STATEMENT

WhatsApp Chat Analyzer is a statistical data analysis tool developed to process and interpret chat logs exported from the WhatsApp platform. The system ingests raw chat files and applies data parsing, transformation, and aggregation techniques to derive actionable insights.Through exploratory data analysis (EDA), the tool generates a range of visual representations, including user interaction metrics such as identifying the most frequently responded-to participants. The primary objective of this tool is to employ advanced data manipulation and analytical methodologies to uncover patterns and trends within user-generated conversation data, thereby facilitating deeper understanding of communication dynamics in WhatsApp environments.

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 versions we couldn’t share images through doc’s format. In this system users are 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 users can send or share or get the chat detail for data analysis through email, Facebook or some messenger application.

1.3 PROPOSED SYSTEM

In the initial phase of the research, data pre-processing involves leveraging various Python libraries to enhance the implementation and functionality of the WhatsApp Chat Analyzer tool. This phase underscores the importance of using established libraries, rather than developing custom functions from scratch, for tasks such as data handling, visualization, and analysis. These libraries facilitate optimized code execution and improve the clarity and interpretability of the analysis for users. The following Python libraries are utilized: NumPy, SciPy, Pandas, CSV, Scikit-learn (sklearn), Matplotlib, sys, re, emoji, NLTK, Seaborn, among others.

As part of exploratory data analysis (EDA), the first step involves applying a sentiment analysis algorithm to classify the chat messages into positive, negative, and neutral categories. The results of the sentiment analysis are then visualized through pie charts based on these sentiment categories. Additionally, multiple data visualizations are generated.

1.4 OBJECTIVE

In the current decade, emerging technologies are increasingly reliant on data. However, this data can only be effectively harnessed through targeted research that addresses the specific needs of the tool being developed. With the growing interest in machine learning, there has been a surge in the development of models designed to tackle a wide range of problems. Consequently, the demand for large-scale, high-quality data has become more critical than ever.

This project aims to enhance the understanding of various types of chat data, specifically WhatsApp conversations.By conducting thorough analysis, the project seeks to provide valuable input for machine learning models that explore and process chat data. These models rely on well-defined learning instances to improve their performance and achieve higher accuracy. Therefore, our research ensures a comprehensive exploratory data analysis (EDA) of different categories of WhatsApp chats, facilitating the development of more robust and accurate machine learning models.

LITERATURE SURVEY

A. Examination of the Utilization and Influence of WhatsApp Messenger through a Demonstration Study [1]:Extensive research has been conducted to assess the usage and societal impact of WhatsApp across diverse demographics. Some studies focus on its influence on students, while others explore its effects on specific local populations. For example, a survey conducted in South India, targeting individuals aged 18 to 23, found that users spend an average of 8 hours per day on WhatsApp and remain online for approximately 16 hours daily. The study emphasized WhatsApp's pivotal role in facilitating peer-to-peer communication, with participants frequently sharing multimedia content such as images, audio, and videos. Notably, WhatsApp has emerged as the most dominant app among the youth in smartphone usage. These research endeavors utilize analyticaltechniques to explore the functionalities of WhatsApp and assess its positive and negative impacts on users.

B. Content Analysis of WhatsApp Chats

[2]: Various research studies evaluate the effectiveness of WhatsApp as a communication tool. This study will contribute to exploring WhatsApp's potential to become the leading mobile communication application. With advancements in technology and the proliferation of smartphones, the way people communicate has drasticallycontains essential information, including their WhatsApp name, status update, and avatar (usually a photo). These profiles are stored in a centralized system, which can be accessed by other users who have added the individual to their contacts. Additionally, the centralized system offers other services such as message relay, user registration, and authentication. Forensic analysis of these interactions can help uncover patterns in user behavior and provide insights into communication practices across the platform.

evolved. Social networking apps,

particularly WhatsApp, have

revolutionized communication, enabling faster and more convenient interactions. Despite potential challenges, individuals may not prioritize basic needs like food or sleep but will still have access to their smartphones for ongoing communication with family, friends, and clients. Messaging platforms like WhatsApp, along with Viber and Skype, have reshaped the global communication landscape, empowering users with instant access to a wide range of features.

2. Forensic Analysis of WhatsApp Messenger [3]: WhatsApp facilitates several types of communication, including user-to-user messaging, broadcast messages, and group chats. Users can exchange simple text messages as well as multimedia content, such as voice, video, and image files, contact cards, and geographical data. Each user’s profile

3. SOFTWARE REQUIREMENT ANALYSIS

Software requirements analysis in the context of systems engineering and software engineering refers to the process of defining, documenting, and managing the requirements for a new or modified product or tool. This process involves identifying and addressing the potentially conflicting needs and expectations of various stakeholders, such as users, developers, and business leaders. It includes tasks such as eliciting requirements, analyzing them for feasibility and consistency, validating them to ensure they align with the intended goals, and managing them throughout the lifecycle of the software or system. The outcome of this analysis is a clear, structured set of requirements that guide the development process and ensure the

product meets the needs of all stakeholders.

3.1 FEASIBILITY STUDY

The primary objective of the feasibility study is to assess the technical, operational, and economic viability of developing the application. Feasibility refers to the process of determining whether a project is worth pursuing based on available resources, time, and other constraints. While theoretically, all systems are feasible with unlimited resources and infinite time, the feasibility study for this project evaluates the practicality of development within real-world limits. The study will be divided into three main areas:

- Technical Feasibility
- Operational Feasibility
- Economic Feasibility

3.1.1 TECHNICAL FEASIBILITY

Technical feasibility refers to evaluating the technical solution's viability and assessing the availability of the necessary resources and expertise to support the system's development. This is one of the first studies conducted after identifying the need for a tool. A technical feasibility study examines the logistical and operational aspects required for the system's development and implementation. It includes specifying the appropriate equipment, software, and technologies that will effectively meet the user’s requirements.

The system's technical requirements may vary, but typically, they include factors such as the ability to produce outputswithin a specified time frame, response times under specific conditions, and the capacity to process a certain volume of transactions at a required speed. These considerations ensure that the system is capable of handling the expected workload efficiently.

In the case of the proposed system for analyzing WhatsApp group chat data, Jupyter software is utilized for its development. Jupyter, a non-profit organization, fosters the creation of open-source software, standards, and services for interactive computing in various programming languages. For this project, Jupyter provides an environment where Python-based data processing code is implemented to analyze and derive meaningful insights from WhatsApp group chat data, ensuring that the system is both technically feasible and capable of handling the required data processing tasks.

3.1.2 OPERATIONAL FEASIBILITY

Operational feasibility primarily addresses concerns related to the system's potential usage, including whether users will adopt the system if developed and implemented, and whether any resistance from users could hinder the system's effectiveness and the realization of its benefits. It focuses on the system's ability to be utilized, supported, and maintained to perform the necessary tasks and functions. Operational feasibility involves all stakeholders, including those who design, operate, and use the system.

This aspect of feasibility assesses how well the proposed system addresses the identified problems and leverages the opportunities discovered during the scope definition and problem analysis phases. In the case of the proposed WhatsApp Chat Analyzer system, operational feasibility is evaluated based on how effectively it meets the users' needs and integrates into their workflow. The system provides valuable insights, such as the number of WhatsApp users and detailed data about their sharing activities, which are visualized in easy-to-understand formats like pie charts and bar charts. These visualizations help users quickly grasp key patterns and trends in the data, contributing to the system's overall operational effectiveness.

3.1.3 ECONOMIC FEASIBILITY

Economic feasibility is one of the most crucial aspects when evaluating the potential effectiveness of a new system. It assesses the cost-effectiveness of the proposed information system solution, ensuring that the benefits outweigh the associated costs. This analysis is particularly important because information systems are often regarded as capital investments for an organization and, as such, should undergo the same investment evaluation as any other capital expenditure. Economic analysis helps determine whether the system justifies its costs through the value it provides.

A key component of economic feasibility is **cost-benefit analysis**, which compares the costs of developing and maintaining the system against the anticipated benefitsit will deliver. This helps stakeholders decide whether the project is a worthwhile investment.

In the case of this project, however, the economic feasibility is considered limited. The system primarily relies on data sharing between two devices, which may not involve significant infrastructure or high operational costs. However, its economic viability might be reduced since the benefits derived from this system are not substantial in terms of financial returns or resource-intensive capabilities. Consequently, this project's cost-benefit analysis may not yield a favorable outcome in the traditional sense, as the benefits are more related to data analysis and user insights rather than a direct economic return.

4. SYSTEM IMPLEMENTATION

Python: Python is an interpreted, high-level, general-purpose programming language created by Guido van Rossum and first released in 1991. Its language constructs and object-oriented approach are designed to support clear and logical code, making it suitable for both small and large-scale applications. Python is widely used in various fields such as web development (server-side), software development, mathematics, data analysis, and machine learning. It can be integrated with other software to create workflows, connect to databases, read and modify files, and handle large datasets for complex computations. Additionally, Python is ideal for rapid prototyping and production-ready software development.

1. **Matplotlib:** Matplotlib is a powerful Python library used for creating static, animated, and interactive visualizations. It offers a wide range of chart types, including pie charts, line charts, bar charts, scatter plots, and histograms. In this project, Matplotlib is used for generating visualizations such as bar charts, line charts, and pie charts to represent the WhatsApp chat data effectively.

2. **Seaborn:** Seaborn is a Python library primarily designed for statistical data visualization. Built on top of Matplotlib, it

6. **NumPy:** NumPy is a general-purpose array-processing library that provides high-performance multidimensional array objects and tools for working with them. It is essential for scientific computing with Python and offers functionality for performing complex mathematical operations on large datasets. In this project, NumPy can be used to handle numerical computations and array manipulations, particularly when dealing with large amounts of chat data.

provides a high-level interface for creating informative and attractive statistical graphics.

In this project, Seaborn can be used for more advanced visualizations, such as heatmaps and distribution plots, that provide deeper insights into the WhatsApp chat data.

3. **Streamlit:** Streamlit is a Python library used to create beautiful, interactive web applications with minimal effort. In this project, Streamlit is utilized to build a web-based interface that presents the WhatsApp chat analytics through various charts and visualizations. It enables real-time, interactive analysis of the chat data, making it accessible and user-friendly.

4. **Pandas:** Pandas is a powerful Python library used for data manipulation and analysis, particularly in data science and machine learning applications. It provides flexible data structures like DataFrames for handling and analyzing structured data. In this project, Pandas is used to manipulate and analyze the WhatsApp chat data, particularly for time series analysis and numerical computations.

5. **Word-Cloud:** A word cloud is a data visualization technique used to display text data, where the size of each word corresponds to its frequency or importance within a dataset. In the context of this project, a word cloud can be used to visualize the most frequently mentioned words in the WhatsApp chat data, providing an intuitive understanding of the key topics and themes in the conversations.

5. WORKING OF PROJECT

1. Data Collection and Preprocessing:

This phase focuses on gathering WhatsApp chat data from the source and preparing it for further analysis. The preprocessing tasks involve extracting and structuring the data into a usable format. This includes identifying and segregating components such as the username, message content, timestamps (day, month, year, hour, minute, and second). The goal is to organize the data into a structured format that is conducive to effective analysis, ensuring that it is clean, consistent, and ready for feature extraction.

2. Feature Extraction: After organizing the data, the next step is to extract relevant features that provide valuable insights into the chat activity. Key metrics include the total number of words, total messages, media shared (such as images, videos, audio files, or other attachments), and links exchanged within the conversation. These features allow a deeper understanding of user engagement and

communication patterns within the WhatsApp group.

3. Timeline Analysis:

In this phase, graphical representations such as line graphs and bar charts are created to visualize the chat activity over different time periods. Weekly timelines reveal trends in messaging behavior across the week, while daily timelines offer a more granular view of activity patterns throughout the day. Analyzing these timelines helps identify peak usage times, periods of low activity, and recurring communication patterns, allowing for a comprehensive understanding of when the group is most active.

4. Identification of the Busiest Users:

Machine learning algorithms are applied to analyze user activity and identify the most active participants in the chat group. Key metrics used to determine user engagement include message frequency, total posts, and overall engagement levels. This analysis helps pinpoint the busiest users and highlights their contribution to the conversation, offering a clearer picture of who is driving the communication within the group.

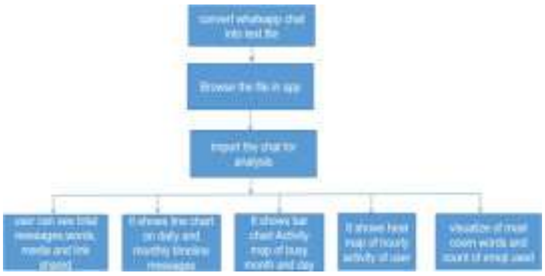
5. Analysis of Emojis and Words:

Textual content within the chat is analyzed to identify the most commonly used emojis and words. Emojis provide emotional context and nuance to the messages, so analyzing their frequency offers insights into the overall tone of the conversation. Word frequency analysis is also conducted to identify recurring themes, topics of interest, and commonlyused phrases, helping to uncover the underlying subjects that dominate the conversations.

6. Visualization:

Insights gained from the analysis are presented visually to enhance user understanding. This includes creating bar charts to show message frequency over time, histograms to illustrate user activity levels, activity maps to identify communication patterns during specific times of the day, and word clouds to display the most frequently used words within the chat. These visualizations simplify the interpretation of data, making it easier for users to gain actionable insights from their WhatsApp conversations and understand key patterns in the data.

FLOWCHART :


convert whatsapp chat into text file
Browse the file in app
import the chat for analysis
user can see total messages,words, media and link shared
It shows line chart on daily and monthly timeline messages
It shows bar chart Activity map of busy month and day
It shows heat map of hourly activity of user
visualize of most coom words and count of emoji used

6. CONCLUSION

The overarching goals established during the initial stages of requirements analysis have been successfully realized. After the system's implementation, it deliversconsistent and reliable results, meeting the expectations outlined in the project requirements. The system is designed with a strong focus on user-friendliness, ensuring that even individuals with limited knowledge of computer environments can easily navigate and operate the developed tool.

Additionally, the system overcomes the limitations of existing manual systems, providing automated processes that reduce human error. It significantly minimizes the risk of incorrect data entry by incorporating built-in validation capabilities, ensuring that only accurate and relevant data is processed. This advancement improves both the efficiency and reliability of the system compared to traditional manual methods.

ACKNOWLEDGEMENT

The authors would like to express their heartfelt appreciation to everyone who contributed to the successful completion of this research study. We extend our special thanks to the faculty members and mentors at Bansal Institute of Engineering and Technology for their continuous encouragement, expert advice, and thoughtful feedback throughout the project. Their guidance played a key role in shaping the research direction and refining the development of the WhatsApp Chat Analyzer. We are equally thankful to all the participants who were involved in the testing phase, whose practical insights and constructive suggestions helped improve the system’s functionality and user experience. Their cooperation andhonest feedback were invaluable in driving this research toward its objectives.

REFERENCES

[1] Marada Pallavi, Meesala Nirmala, Modugaparapu Sravani, Mohammad Shameem. WhatsApp Chat Analysis. International Research Journal of Modernization in Engineering Technology and Science. Volume: 04/Issue:05/May-2022

[2] Shaikh Mohd Saqib. Whatsapp Chat Analyzer. International Research Journal of Modernization in Engineering Technology and Science. Volume: 04/Issue:05/May-2022

[3] K, Ravishankara & Dhanush, & Vaisakh, & S, Srajan. (2020). Whatsapp Chat Analyzer. International Journal of Engineering Research and. V9.10.17577/IJERTV9IS050676.

[4] D.Radha, R. Jayaparvathy, D. Yamini, "Analysis on Social Media Addiction using Data Mining Technique", International Journal of Computer Applications (0975 – 8887).

[5] https://towardsdatascience.com/sentime ntal-analysis-using-vader-a3415fef7664

[6] https://www.analyticsvidhya.com/blog/ 2021/06/build-web-app-instantly-for-mach ine-learning-using-streamlit/

[7] Meng Cai, "PubMed Central", PMCID: PMC7944036, PMID: 33732917