IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

WhatsApp Chat Analyzer

[1]Pratik Honmane, [2]Sanskruti Kaliya, [3]Shweta Makasare, [4]Chirag Patekar

[1]Student, [2]Student, [3]Student, [4]Student
Department of Computer Engineering,
GHRCOEM, Ahmednagar, India

Abstract: WhatsApp is a free, multiplatform messaging app that lets you make video and voice calls, send text messages, and more all with just a Wi-Fi connection. With over 2 billion active users, WhatsApp is especially popular among friends and family who live in different countries and want to stay in touch. To understand WhatsApp's popularity, you need to understand it was one of the first mobile apps to offer free, internet-based messaging. Instead of sending texts using cellular-data networks, where fees may apply, WhatsApp primarily relies on a Wi-Fi connection to send and receive messages and calls for free. People can communicate with other people using the voice and video calls, voice messaging, secure messaging, photos and video sharing, document sharing, etc. Predictions from global data experts show that humans will produce and consume about 94 zettabytes of data by the end of 2022.

Keywords–*WhatsApp,Emoji,SentimentAnalysis,MachinelearningLDA,*

I. INTRODUCTION

WhatsApp is an internationally available freeware, cross-platform, centralized instant messaging and voice-over-IP (VoIP) service owned by US tech conglomerate Facebook (Meta). WhatsApp allows you to send text and voice messages, make video and voice calls and share images, locations, documents and other content. WhatsApp service created by WhatsApp Inc. of Mountain View, California, which was acquired by Facebook (Meta) in February 2014 for approximately US $19.3 billion. WhatsApp became the world's most popular messaging application by 2015, and had more than 2 billion user worldwide by February 2020.
"According to one survey an average user spends more than 195 minutes per week on WhatsApp". In WhatsApp Analyzer, we analyzing WhatsApp individual activities. It track our conversation and analyzes how much time we are spending on "WhatsApp". Sentiment analysis is nothing but the techniques, methods and tools for detecting as well as extracting information such as opinions and attitude. Sentiment analysis has been about opinion polarity i.e. categorization of opinion as positive, negative or neutral. The data is analyzed using in order to derive various features to guide the behavioral analysis.
In this paper, we create the web application using the python streamlit library and analyze the behavior of the user and what type emoji user used using the emoji library. Streamlit library is an open-source framework for Machine Learning and Data Science teams. Streamlit Library is used to create the web applications in minutes.

II. RELATED WORK

Substantial work has already been done in the field of 'Sentiment Analysis on Social Media'. In most cases, the work has been carried out using WhatsApp Data. WhatsApp came into the market as a substitute for SMS. It is used to make the voice or video calls as well as to share the images, documents, etc. WhatsApp provides these services for free. According to survey, an average user spends more than 195 minutes per week on WhatsApp. In 2022, 650 million Tweets send by the user per day. The user opinion can be positive, negative or neutral. In WhatsApp, opinion can be given in the form of text or using emoji's. For our study, six emotions have been chosen and classified our emoji's in these categories.
Behavior classification is the process of identifying and categorizing different types of behaviors based on their characteristics and features. There are various methods for behavior classification, and some of the commonly used ones are:

1. **Observational methods:** These methods involve directly observing and recording behaviours. They can be structured, where the observer looks for specific behaviours or unstructured, where the observer takes note of any behaviour that occurs.
2. **Self-report measures:** These methods involve asking individual to report on their own behaviors using questionnaires or surveys. They can be useful in cases where behaviors are difficult to observe or when studying private behaviour.
3. **Machine Learning Models:** These methods involve using algorithms to analyze data and classify behaviours. Various supervised learning models, such as decision trees, random forests, and support vector machines, can be used to classify behaviours based on labelled data.

4. **Neural Networks:** With the recent advancements in deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are being used to classify behaviours in real-time video data.
5. **Cluster analysis:** This method involves grouping behaviours based on their similarity using statistical techniques. Cluster analysis can be useful in identifying patterns of behaviors or identifying subtypes of a behaviour.

The choice of a method for behavior classification depends on the research question, available resources, and the context in which the behaviors are being studied.

III. DATA COLLECTION

- Data Collection

  For the purpose of this research, is to find which word is mostly used by user? , on which day the user most active? , on which month user can used the WhatsApp? , etc. Firstly, all the WhatsApp Chats have been collected data from the users. The "Export Chat" feature available on WhatsApp has been used to collect the conversations between the users. The collected data is without media. Model needs only textual data.

- Analysis

  A database is formed consisting of several columns like user, messages, day, date, hour, minute, etc.

Column Name	Information
Date	17-06-2019
User	XYZ
Message	17-06-2019
Only_date	17-06-2019
Year	2019
Month_num	06
Month	June
Day	Monday
Day_name	Monday
Hour	12:36:00
Minute	36 Minute
Period	01:36 PM

It has been studied that emoji's can often act as a replacement for words when words are not sufficient to express our emotions. Also emoji's express how comfortable the users are in a conversations and emoji count is inversely proportional to how professional the conversation between the two parties is.

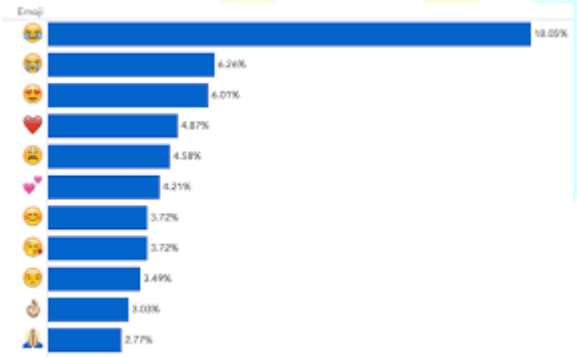


Fig.1 Emoji Analysis

It can also studied that the total number of messages, total numbers of links shared, total words and media shared.

Top Statistics Table

Total Messages	Total Words	Media Shared	Links Shared
21334	23273	2326	123

- Pre-Processing
  1. **Emoji's:** The entire set of emoji codes is defined by the Unicode consortium. The emoji list prescribed by Unicode along with the keywords associated with every emoji is used to classify

them into 6 categories. WhatsApp allows different skin colors for certain emoji's. Hence Regex library in combination with the Unicode Library can be used to extract these emoji's. The classification is shown in figure 2.

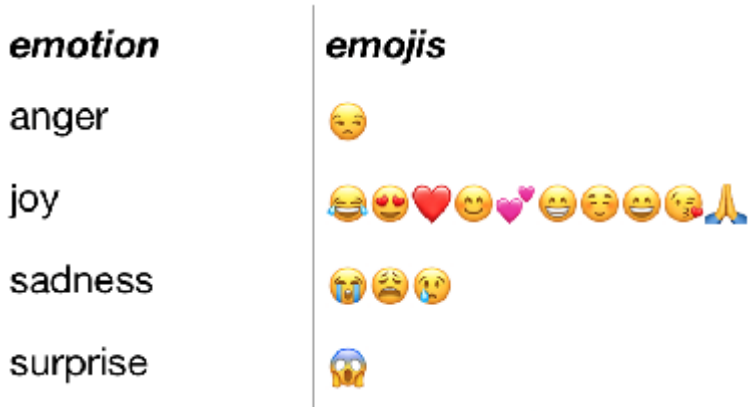


Fig 2. Emoji's categorized by emotions

- Steps in Pre-processing:
  Regex Library is used to find the emoji patterns in the text and separate the text from emoji's.

IV. PROPOSED ALGORITHM

Algorithm

In the first step, WhatsApp chats are collected from various users. These dataset contains the message from both sender and receiver with their time stamps.
In this model, we analyze the how many messages that user can receive , on which day the user most active, on which month the user most use WhatsApp, which emoji was used most to express his/her opinion.

Using the regex library, we can analyze the emoji patterns in the text and separate the text from emoji.
Here is a proposed algorithm for a WhatsApp chat analyzer:

1. **Input:** The chat log text file exported from WhatsApp.
2. Parse the text file to extract individual messages with their respective timestamp, sender and message content.
3. Store the parsed messages in a data structure that can be easily analyzed, such as a list or a database.
4. Perform basic data cleaning tasks such as removing emoji's, URLs, and non-alphanumeric characters.
5. Perform text preprocessing tasks such as removing stop words, stemming or lemmatization, and converting all text to lowercase.
6. Perform sentiment analysis on the messages using a pre-trained model or library.
7. Calculate the frequency of words used in the chat and create a word cloud to visualize the most commonly used words.
8. Analyze the chat activity over time by aggregating messages by hour or day and plotting the data on a time-series graph.
9. Identify the most active participants in the chat by counting the number of messages sent by each participant and visualizing the results in a bar chart.
10. Identify the most frequent topics of discussion in the chat by clustering similar messages together using techniques such as K-means clustering or Latent Dirichlet Allocation (LDA).
11. Summarize the chat by extracting the most important messages or topics using techniques such as TextRank or summarization algorithms.
12. **Output:** The results of the analysis can be displayed in a graphical user interface (GUI) or exported to a report format.

This algorithm can be extended or modified based on the specific requirements and goals of the WhatsApp chat analyzer.

V. LANGUAGE AND LIBRARIES

Python:-
Python is a high-level, interpreted programming language that was first released in 1991. It was developed by Guido van Rossum and is now maintained by the Python Software Foundation. Python is an open-source language, meaning that its source code is freely available and can be modified and redistributed by anyone. Python is known for its simple and easy-to-read syntax, making it popular for beginners to learn programming. It is also widely used in industry, particularly in scientific computing, data analysis, and web development.

Libraries:-
In the context of programming, a library is a collection of pre-written code that can be reused in different programs. Libraries can save programmers a significant amount of time and effort by providing ready-made solutions to common programming problems.

In Python, there are many libraries available that can be used for various purposes, such as data analysis, scientific computing, web development, and more.

1. **Pandas:-**
   Pandas is a Python library that is widely used for data manipulation and analysis. It provides data structures for working with labelled and relational data, as well as functions for data cleaning, merging, and transformation. It provides data structures for working with labelled and relational data, as well as functions for data cleaning, merging, and transformation.
2. **Numpy:-**
   Numpy is a Python library for scientific computing that provides support for large, multi-dimensional arrays and matrices, along with a range of functions for mathematical operations on these arrays. Numpy is a library for numerical computing in Python. It provides functions for performing complex mathematical operations on arrays, as well as linear algebra, Fourier analysis, and more.
3. **Seaborn:-**
   Seaborn is a Python library for data visualization that is built on top of Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is a powerful and user-friendly library for data visualization that can be used for both exploratory data analysis and communication of insights. It is particularly useful for creating informative and aesthetically pleasing statistical graphics for presentations and reports.
4. **Regex:-**
   Regex (short for Regular Expression) is a sequence of characters that specifies a search pattern. It is a powerful tool for pattern matching and text processing in Python and many other programming languages. Regex uses a combination of characters, such as letters, numbers, and special characters, to define a pattern. Regex is a powerful and versatile tool for text processing and pattern matching, and can be used in a wide variety of applications, such as web scraping, data cleaning, and text mining.
5. **Emoji:-**
   An emoji is a graphical representation of an emotion, object, symbol, or concept that is often used in electronic communication such as text messages, emails, and social media. Emoji's are typically small icons or images that are designed to convey a specific meaning or mood. In Python, emoji's can be represented as Unicode characters, which are a standardized system for encoding and displaying characters and symbols from different languages and scripts. Emoji's are represented in Unicode by a code point, which is a unique identifier for a specific character or symbol. For example, the "smiling face with heart-eyes" emoji can be represented in Python as the Unicode character U+1F60D, which is the hexadecimal value 0x1F60D. This character can be displayed in a Python string using the \u escape sequence followed by the hexadecimal value, like this: "\u1F60D". Python also provides several libraries for working with emoji's, such as the emoji library, which allows you to easily add emoji's to your Python code by using their Unicode code points. The emoji library also provides functions for converting text to emoji's, searching for emoji's in text, and manipulating emoji's in various ways. Overall, emoji's are a fun and expressive way to add personality and emotion to electronic communication, and Python provides several tools and libraries for working with them in your code.
6. **WordCloud:-**
   Word cloud is a visual representation of text data, where the size of each word is proportional to its frequency or importance in the text. Python has several libraries for generating word clouds, including the wordcloud library, which is a popular choice. The WordCloud Library is used for which word has the maximum frequency that word can be shown in color and max size.

VI. FEATURES

WhatsApp chat analysis can provide several valuable insights into the communication patterns and behaviors of individuals or groups. Some of the key features and metrics that can be analyzed include:

- **Chat frequency:** This measures how often messages are sent and received, as well as the average response time between messages.
- **Message length:** This measures the length of messages in terms of characters or words, which can reveal insights into the level of engagement and interest in the conversation.
- **Word frequency:** This measures the frequency of words or phrases used in the conversation, which can help identify topics of interest or recurring themes.
- **Emoji's and emoticons:** This measures the use of emoji's and emoticons, which can provide insights into the emotional tone and sentiment of the conversation.
- **Time of day and day of week:** This measures when messages are sent and received, which can help identify patterns in communication behavior, such as peak usage times or changes in behaviour over time.
- **Sentiment analysis:** This measures the overall emotional tone of the conversation, such as positive, negative, or neutral.
- **Network analysis:** This measures the connections and relationships between individuals in the conversation, such as who communicates with whom and how often.

These features can be analyzed using a combination of text analysis techniques, such as natural language processing, machine learning, and data visualization. The insights gained from WhatsApp chat analysis can be useful in a variety of applications, such as social media marketing, customer service, and personal relationship management.

VII. CONCLUSION

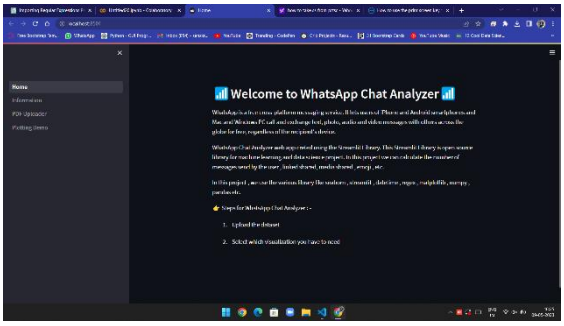
We proposed a machine learning approach to analyze WhatsApp chat data. Our approach involved preprocessing the data, extracting features, and applying machine learning algorithms to identify patterns and insights. We applied our approach to a real-world WhatsApp chat dataset and obtained promising results. Our analysis revealed important insights into the communication behavior of WhatsApp users, such as the most frequently used words, the peak usage times, and the sentiment analysis of the chats. The proposed approach can be extended to analyze other types of chat data and can have significant implications in the fields of social media analysis and user behavior modeling.

VIII. REFERENCES

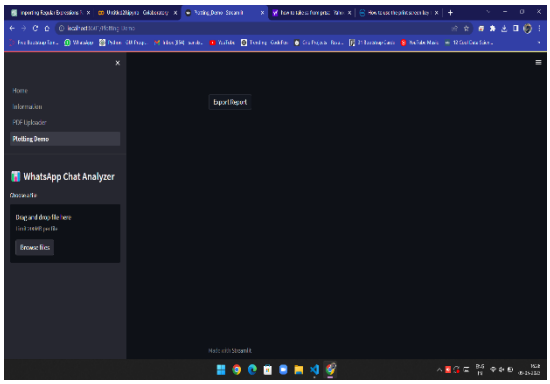
- Aharony, N., T., G., The Importance of the WhatsApp Family Group: An Exploratory Analysis. "As lib Journal of Information Management, Vol. 68, Issue 2, pp.1-37" (2016). [4] Access Data Corporation. FTK Imager, 2013.
- D. Radha, R. Jayaparvathy, D. Yamini, "Analysis on Social Media Addiction using Data Mining Technique", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, pp. 23-26, April 2016.
- F. 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.G.
- Mike Dickson, "An examination into AOL instant messenger 5.5 contact identification.", Digital Investigation, Science Direct, vol. 3, issue 4, pp. 227-237, 2006.H.
- https://www.analyticsvidhya.com/blog/2021/04/whatsapp-group-chat-analyzer-using-python/

IX. RESULTS

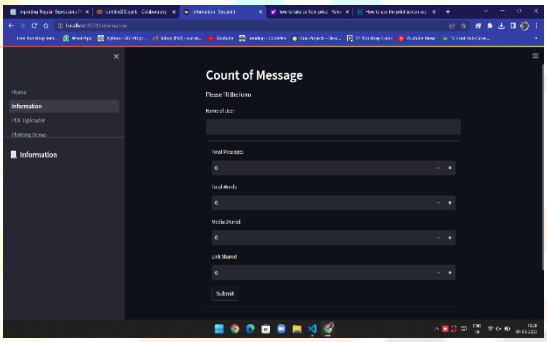
1. GUI :-



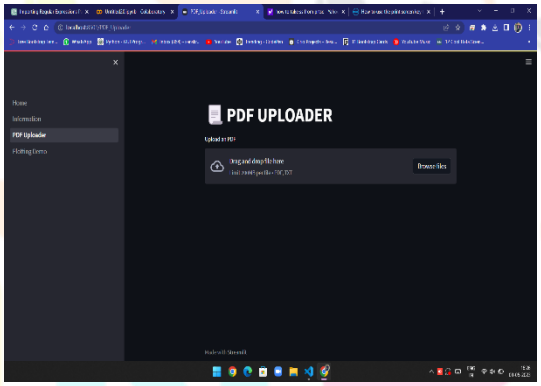
2. Plotting Demo Tab:-



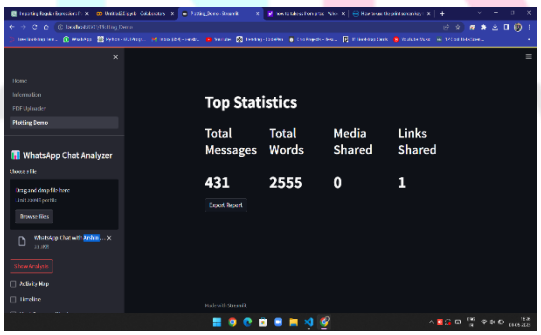
3. Information Tab:-



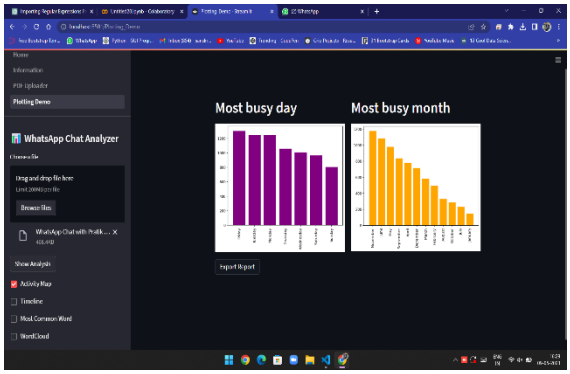
4. PDF UPLOADER:-



5. Top Statistics Table:-



6. Most Busy day and Month:-



7. Monthly Timeline and Daily Timeline :-