WhatsApp Chat and Sentiment Analysis

1. Introduction

In this paper, we propose a WhatsApp Chat Sentiment Analyzer designed to analyze conversations within WhatsApp groups and among individual users. WhatsApp conversations encapsulate diverse forms of communication, reflecting interactions and exchanges among participants. Leveraging the exported chat file, our system facilitates the analysis of such data, offering insights into communication dynamics and sentiment expressions.

Our approach emphasizes the utilization of simple yet effective Python libraries, including seaborn, Streamlit, Numpy, Matplotlib, and Pandas. These libraries enable the creation of data frames and graphical representations, ensuring a user-friendly and efficient analysis process.

Data preprocessing assumes a pivotal role in our methodology, aiming to enhance the quality and relevance of the analysis outcomes. With claims of over 50 billion messages sent daily and an average user spending more than 500 minutes weekly on the application, the significance of robust data preprocessing techniques cannot be overstated.

Overall, our proposed WhatsApp Chat Sentiment Analyzer provides a valuable tool for users seeking to gain insights from their WhatsApp interactions, empowering them to better understand communication trends and sentiment expressions within their digital communities.

2. Problem Statement

The WhatsApp Chat Sentiment Analyzer serves as a statistical analysis tool tailored for WhatsApp conversations. Utilizing exported chat files, the tool facilitates the generation of various analytical plots, such as identifying the most frequent interactions with other group participants. To enhance comprehension of WhatsApp chats on mobile devices, we advocate for the application of record manipulation techniques.

3. Project Concept

Considerable advancements have been made in the evolution of the WhatsApp application. Previous versions lacked features such as status display, document

`

The WhatsApp Chat Sentiment Analyzer project aims to leverage advancements in both the WhatsApp application itself and data analysis technologies to provide users with valuable insights into their chat conversations. By integrating features such as status display, document sharing, and location sharing, the current version of WhatsApp has become a comprehensive communication platform. However, previous versions lacked these features, highlighting the importance of adapting analysis tools to keep pace with application updates.

Development Approach:

The WhatsApp Chat Sentiment Analyzer project takes a comprehensive approach to chat analysis, starting with the preprocessing and formatting of exported chat files. Leveraging the Numpy library for efficient data manipulation and preprocessing tasks ensures that the data is clean and ready for analysis.

Key Components:

1. Cleaning and Formatting:
- Utilizing Numpy for preprocessing tasks ensures the integrity and consistency of the chat data, enabling accurate analysis downstream.

2. Data Structuring with Pandas:
- The Panda's library is instrumental in constructing a structured data frame from the cleaned chat data. This structured format facilitates in-depth data analysis and the extraction of meaningful insights.

3. Sentiment Analysis with NLTK Vader:
- Leveraging the Natural Language Toolkit (NLTK) framework, specifically the Vader library, enables sentiment analysis of group chats or individual conversations. This analysis sheds light on the emotional dynamics within the chat data.

Visualization Dashboard:

The culmination of the analysis process is the visualization dashboard, where various parameters extracted from the chat file are presented in a user-friendly and intuitive manner. Statistical representations enhance the interpretability of sentiment dynamics within the chat data, empowering users to gain valuable insights into their conversations.

`

sharing, and location sharing, which are now integrated into the current version. Additionally, the inability to share images in document format was a limitation of older versions.

In the WhatsApp Chat Sentiment Analyzer, we have developed a visualization dashboard to present various parameters extracted from the exported chat file. Initially, the exported chat undergoes cleaning and formatting using Numpy for preprocessing. Subsequently, leveraging the Panda's library, a structured data frame is constructed, facilitating data analysis and the derivation of meaningful insights.

Furthermore, we employ the NLTK framework, specifically utilizing the Vader library, to conduct sentiment analysis on the group chat or individual conversations. The results are then visualized through statistical representations, enhancing the interpretability of sentiment dynamics within the chat data.

4. Implementation

4.1 Technical Feasibility

Python: Python is a high-level, general-purpose programming language. The following libraries of python are used like Numpy, Scipy Pandas, CSV, Matplotlib, sys, re, emoji, NLTK seaborn, etc.

Regex (Regular Expression): A regular expression (regex) is a sequence of characters that define a search pattern. It can be used for "find" or "find and replace" operations on strings, or for input validation.

4.2 Operational Feasibility

Matplotlib: Matplotlib is a comprehensive plotting library for Python. It can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code.

Streamlit: In this project, we will use this library to create beautiful web elements and objects to represent our WhatsApp chat analytics using different kinds of charts and visualizations with Streamlit as a web application.

Pandas: Pandas is a Python library providing high-performance, easy-to-use data structures and data analysis tools. It's an indispensable tool in the world of data analysis and data science because it allows for efficient data cleaning, transformation, and analysis.

`

CSV: The CSV module facilitates reading and writing CSV files, which are commonly used for storing tabular data.

Matplotlib: Matplotlib is a versatile plotting library that enables the creation of a wide range of visualizations, including line plots, bar charts, scatter plots, and more. Its ease of use and extensive customization options make it well-suited for visualizing the results of WhatsApp chat analysis.

re: The re module enables the use of regular expressions for text processing tasks such as pattern matching and string manipulation.

Emoji: The emoji library allows for handling emojis within text data, which is essential for analyzing chat conversations containing emojis.

Data Visualization Techniques: Discuss various data visualization techniques employed in the project, such as bar charts for message frequency, time series plots for message distribution over time, and word clouds for visualizing frequently used words or emojis in the chat data.

Scalability Considerations: Address potential scalability issues and discuss strategies for handling large volumes of chat data efficiently, such as parallel processing, data streaming, or distributed computing frameworks. By addressing these scalability considerations and implementing appropriate strategies, the WhatsApp Sentiment and Chat Analysis project can effectively handle large volumes of chat data while maintaining optimal performance and usability.

User Interface Design: Provide insights into the design considerations for the user interface, including layout, interactivity, and accessibility features, to ensure a seamless user experience while interacting with the chat analysis tool. By incorporating these design considerations into the user interface of the chat analysis tool, you can ensure a seamless and engaging user experience, allowing users to effectively analyze WhatsApp conversations and derive valuable insights from the data.

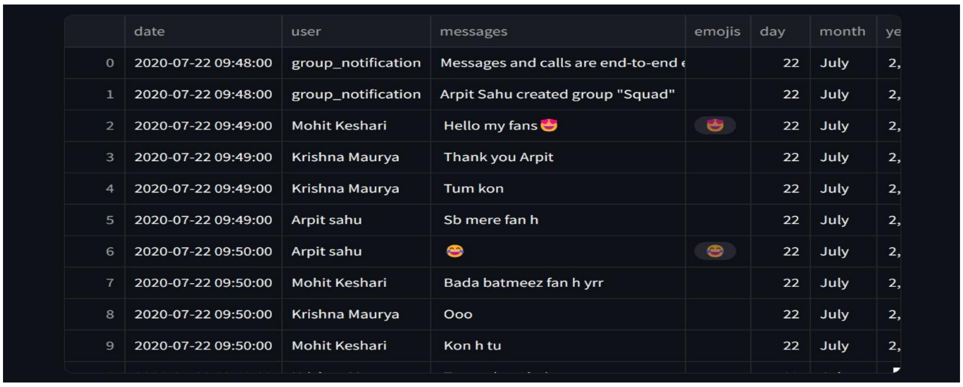
NLTK: NLTK (Natural Language Toolkit) offers a suite of libraries and programs for natural language processing tasks such as tokenization, stemming, tagging, parsing, and more. Seaborn: Seaborn is built on top of Matplotlib and provides a higher-level interface for creating aesthetically pleasing statistical graphics.

`

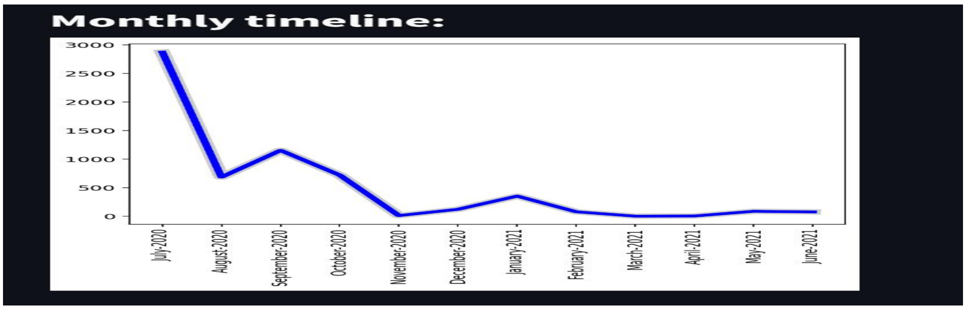
VADER: VADER (Valence Aware Dictionary for Sentiment Reasoning) is a rule-based sentiment analyzer. It contains a list of lexical features which are generally labelled as per the semantic orientation as positive or negative It is included in the NLTK package and can be applied directly to unlabelled text data. VADER's SentimentIntensityAnalyzer() takes a string and returns a dictionary of three categories as Positive, Negative, Neutral.

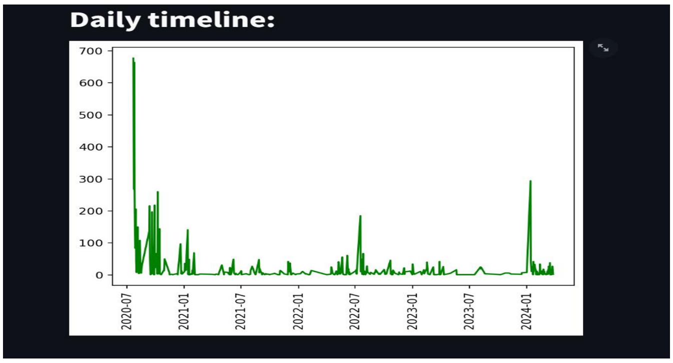
5. Working

1. Go to sidebar and click on browse files.
2. Select the WhatsApp chat file for which analysis has to be done.
3. Here you can see the entire group chats in data frame and Monthly and Daily Timeline of the chats plotted on the graph.



	date	user	messages	emojis	day	month	ye
0	2020-07-22 09:48:00	group_notification	Messages and calls are end-to-end e		22	July	2,
1	2020-07-22 09:48:00	group_notification	Arpit Sahu created group "Squad"		22	July	2,
2	2020-07-22 09:49:00	Mohit Keshari	Hello my fans 😍	😍	22	July	2,
3	2020-07-22 09:49:00	Krishna Maurya	Thank you Arpit		22	July	2,
4	2020-07-22 09:49:00	Krishna Maurya	Tum kon		22	July	2,
5	2020-07-22 09:49:00	Arpit sahu	Sb mere fan h		22	July	2,
6	2020-07-22 09:50:00	Arpit sahu	😂	😂	22	July	2,
7	2020-07-22 09:50:00	Mohit Keshari	Bada batmeez fan h yrr		22	July	2,
8	2020-07-22 09:50:00	Krishna Maurya	Ooo		22	July	2,
9	2020-07-22 09:50:00	Mohit Keshari	Kon h tu		22	July	2,





4. The initial segment presents comprehensive statistics encompassing the total number of messages, words, media files, URL counts, and emojis utilized within the chat.

5. Following this, a sentiment analysis is conducted. Initially, it computes the aggregate monthly data concerning positive, neutral, and negative activity within the chat.

6. Subsequently, a weekly report detailing chat activity is generated and visually depicted through bar graphs.

7. Individual contributions towards positive, negative, and neutral sentiments are quantified, and visual representations, such as pie charts and data frames, are employed for clarity.

8. Utilizing value counts, the frequency of sentiments is analyzed, and separate graphs are constructed to highlight the most positive, negative, and neutral users.

9. An examination of chat dynamics without sentiment analysis is conducted.

10. Identification of the top five most active users within the group is executed, illustrated through pie charts, and detailed in a sorted data frame based on message volume per individual.

11. A Word Cloud visualization is utilized to showcase the prevalent words employed in the group chat.

12. The analysis extends to identifying the most frequently utilized words in the chat, visualizing their frequency on a graph, and documenting the results in a corresponding data frame.

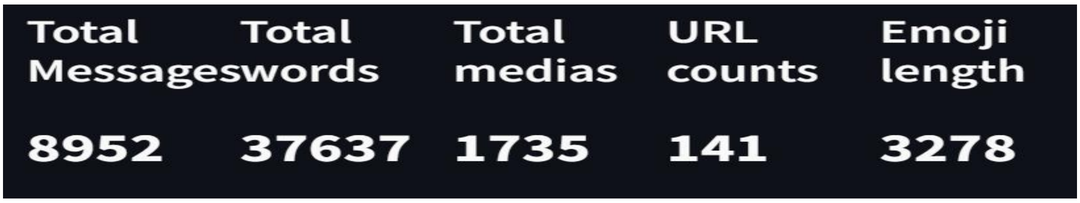
13. Subsequently, the analysis identifies the most frequently utilized emoji, and the top five used emojis are visually represented through a pie chart.

14. The messages undergo filtration to remove instances containing Hinglish words. Following this, a count of messages per user is computed and presented in a structured data frame. Additionally, the user with the highest word count is highlighted through a plotted visualization.

15. An evaluation of the most active week is conducted, resulting in the creation of a corresponding data frame detailing the pertinent statistics.

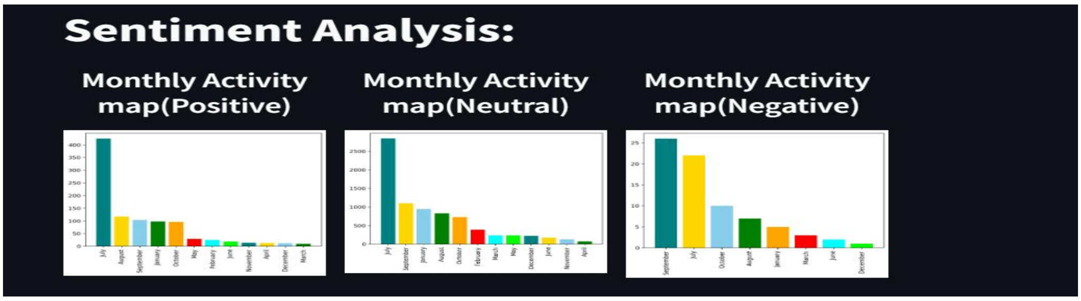
16. Similarly, an examination is carried out to determine the most active month, with a corresponding data frame providing a comprehensive overview of the activity levels.

6. Output



Total Messages	Total words	Total medias	URL counts	Emoji length
8952	37637	1735	141	3278

Analysis



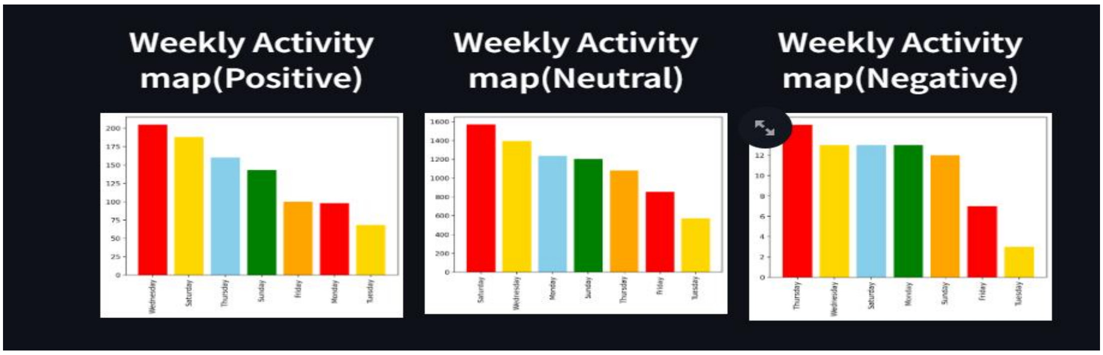
Sentiment Analysis:

Monthly Activity map(Positive)

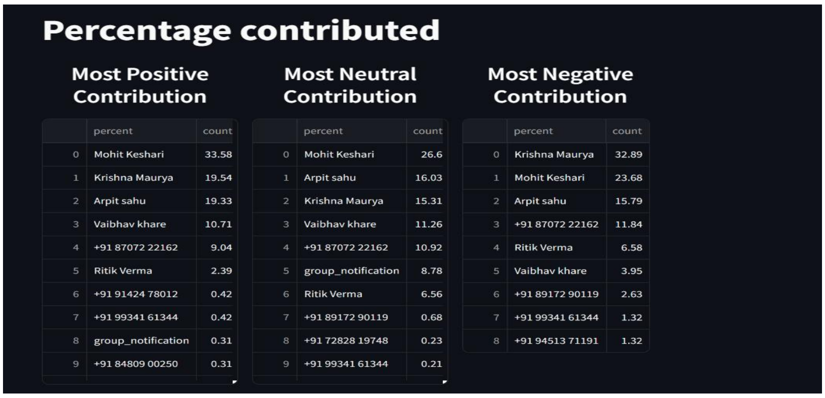
Monthly Activity map(Neutral)

Monthly Activity map(Negative)

Monthly Sentiments



Weekly Sentiments



Percentage contributed

Most Positive Contribution

	percent	count
0	Mohit Keshari	33.58
1	Krishna Maurya	19.54
2	Arpit sahu	19.33
3	Vaibhav khare	10.71
4	+91 87072 22162	9.04
5	Ritik Verma	2.39
6	+91 91424 78012	0.42
7	+91 99341 61344	0.42
8	group_notification	0.31
9	+91 84809 00250	0.31

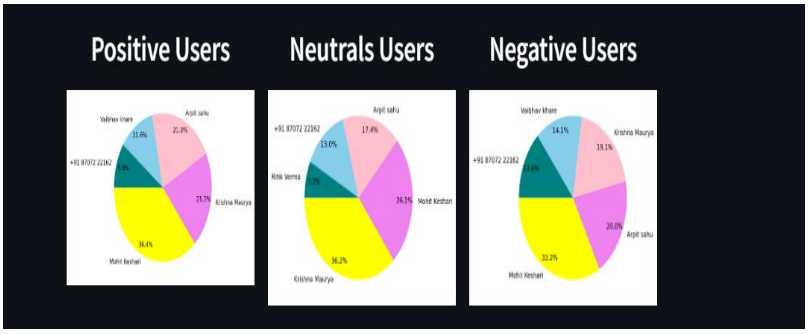
Most Neutral Contribution

	percent	count
0	Mohit Keshari	26.6
1	Arpit sahu	16.03
2	Krishna Maurya	15.31
3	Vaibhav khare	11.26
4	+91 87072 22162	10.92
5	group_notification	8.78
6	Ritik Verma	6.56
7	+91 89172 90119	0.68
8	+91 72828 19748	0.23
9	+91 99341 61344	0.21

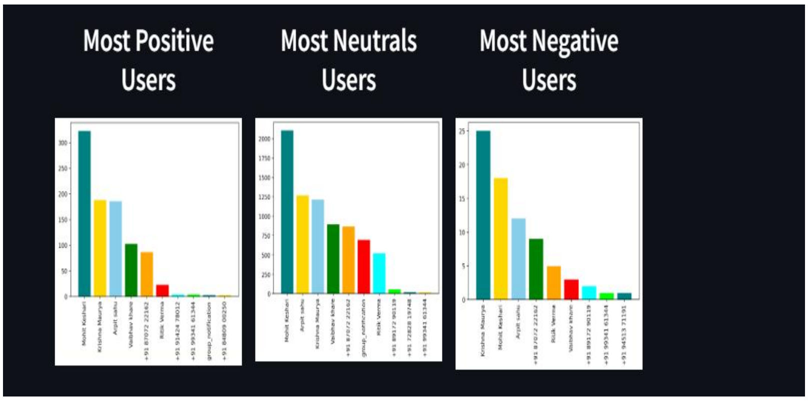
Most Negative Contribution

	percent	count
0	Krishna Maurya	32.89
1	Mohit Keshari	23.68
2	Arpit sahu	15.79
3	+91 87072 22162	11.84
4	Ritik Verma	6.58
5	Vaibhav khare	3.95
6	+91 89172 90119	2.63
7	+91 99341 61344	1.32
8	+91 94513 71191	1.32

Data Frame for percentage contributed by each user for each sentiment

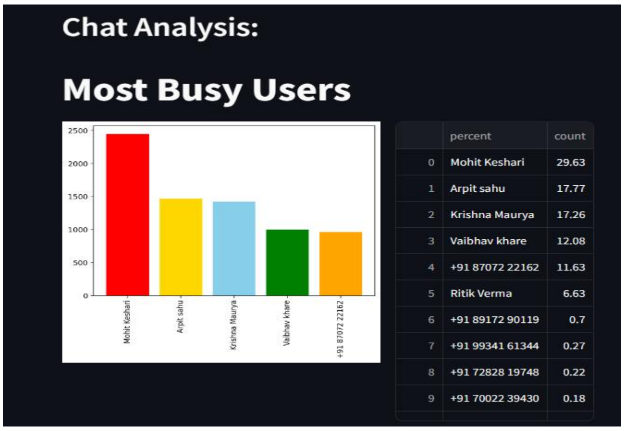


Visualisation of Percentage Distribution on Pie Chart



Visualisation of Most Positive, Neutral and Negative Users

`

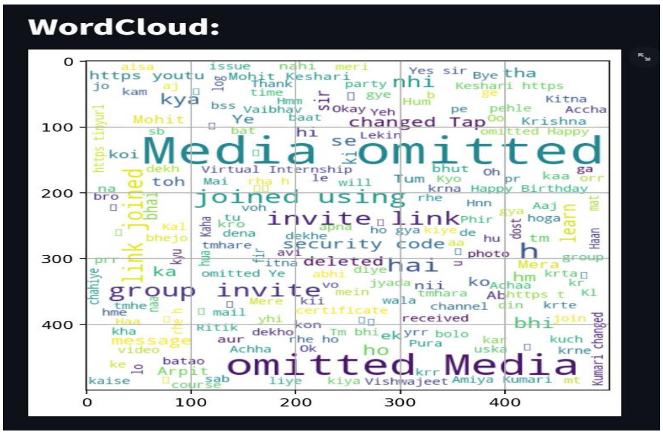


Chat Analysis:

Most Busy Users

	percent	count
0	Mohit Keshari	29.63
1	Arpit sahu	17.77
2	Krishna Maurya	17.26
3	Vaibhav khare	12.08
4	+91 87072 22162	11.63
5	Ritik Verma	6.63
6	+91 89172 90119	0.7
7	+91 99341 61344	0.27
8	+91 72828 19748	0.22
9	+91 70022 39430	0.18

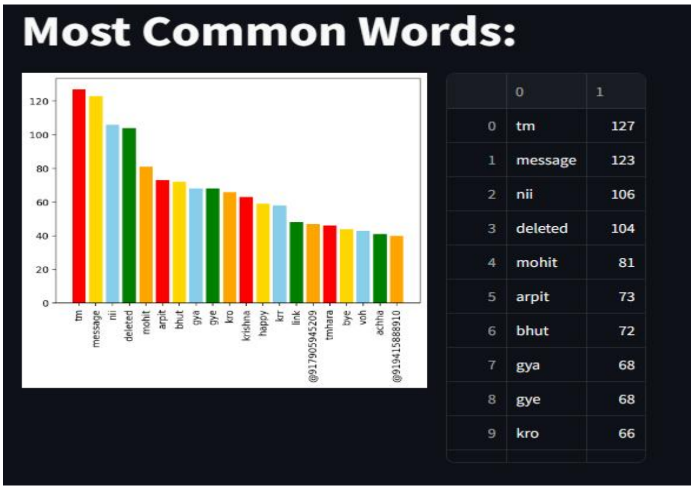
Most Busy User Among the Group



WordCloud:

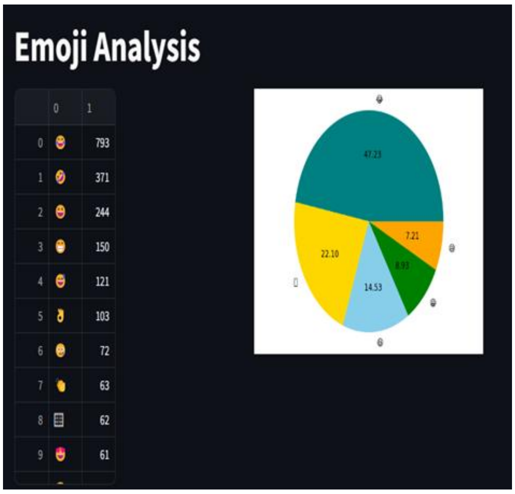
0
100
200
300
400
aisa issue nahi meri Yes sir Bye tha
https youtu log Mohit Keshari Thank party nhi Keshari https
jo kam aj time Hmm gye b ge Kitna
kya bss Vaibhav sir Okay Yeh Hum pe pehle Accha
Mohit Ye baat changed Tap Oo Krishna
https tinyurl sb bat hi Lekin omitted Happy
koi Media omitted se ki
dekh Virtual Internship bhut Oh ga
toh Mai rha h le will Tum Kyo pr kaa orr
na bro joined using rhe krna Happy Birthday
link joined bhai voh Hnn Aaj mat
Kal tu kro invite link apna Phir gya hoga learn
bhejo Kaha dena dekhe ho gya kiye de hu dost tm Haan
tmhare security code aa photo h
prr kyu hua fir avi itna deleted hai u Mera group
ka omitted Ye abhi diye hm krta
chahiye group invite vo jyada nii ko Achaa kr
tmhe naa Mere mein tmhara Ab https t Kl
hme mail kii wala channel din krte
Haa rhe h yhi certificate received join
kha Ritik kon bhi
message aur dekho rhe ho Tm bhi ek yrr bolo kar kuch
video Achha Ok ho Pura uska krne
ke batao lo omitted Media krr
Arpit kaise course sab liye kiya Vishwajeet Amiya Kumari mt Kumari changed
0
100
200
300
400

Word Cloud



Most Common Words:

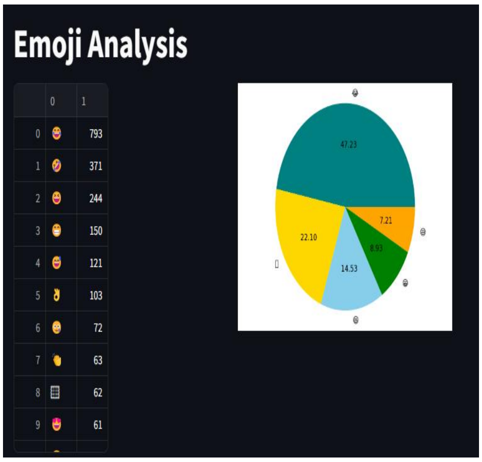
	0	1
0	tm	127
1	message	123
2	nii	106
3	deleted	104
4	mohit	81
5	arpit	73
6	bhut	72
7	gya	68
8	gye	68
9	kro	66



Emoji Analysis

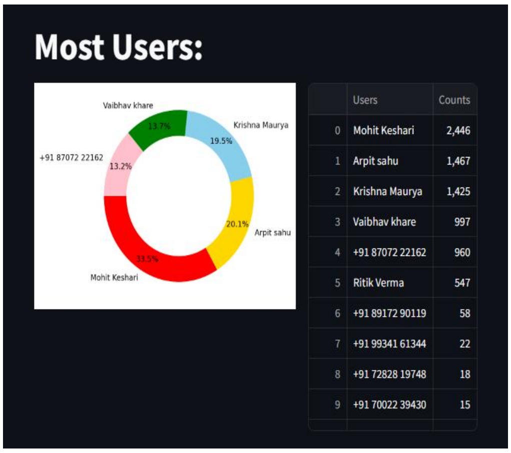
	0	1
0	😂	793
1	🤣	371
2	😄	244
3	😁	150
4	😅	121
5	👌	103
6	😊	72
7	👏	63
8	📄	62
9	😍	61

Most Common Word



Emoji Analysis

	0	1
0	😂	793
1	🤣	371
2	😄	244
3	😁	150
4	😅	121
5	👌	103
6	😊	72
7	👏	63
8	📄	62
9	😍	61



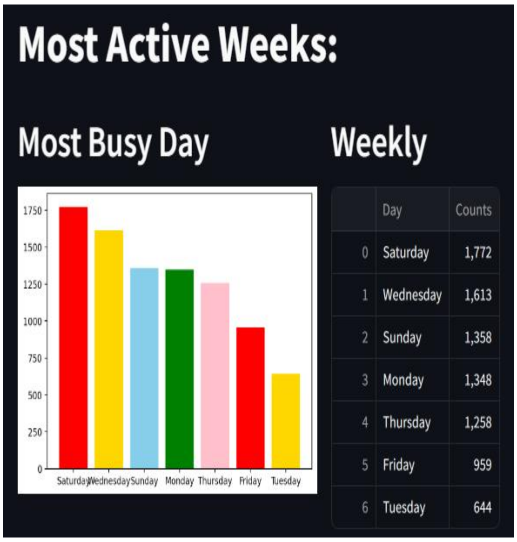
Most Users:

	Users	Counts
0	Mohit Keshari	2,446
1	Arpit sahu	1,467
2	Krishna Maurya	1,425
3	Vaibhav khare	997
4	+91 87072 22162	960
5	Ritik Verma	547
6	+91 89172 90119	58
7	+91 99341 61344	22
8	+91 72828 19748	18
9	+91 70022 39430	15

Most Frequently Used Emoji

Users with Most Word Count

`



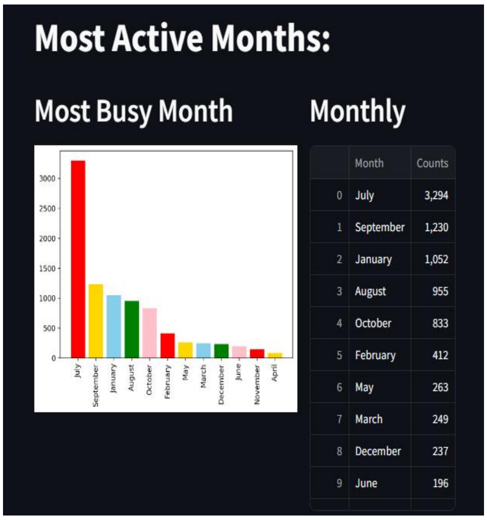
Most Active Weeks:

Most Busy Day

Weekly

	Day	Counts
0	Saturday	1,772
1	Wednesday	1,613
2	Sunday	1,358
3	Monday	1,348
4	Thursday	1,258
5	Friday	959
6	Tuesday	644

Most Active Weeks



Most Active Months:

Most Busy Month

Monthly

	Month	Counts
0	July	3,294
1	September	1,230
2	January	1,052
3	August	955
4	October	833
5	February	412
6	May	263
7	March	249
8	December	237
9	June	196

Most Active Months

`

Results and Analysis:

The WhatsApp Chat Sentiment Analyzer successfully fulfills its intended purpose as a statistical analysis tool tailored for WhatsApp conversations. Utilizing exported chat files, the tool seamlessly processes and analyzes the data, providing valuable insights into the dynamics of group interactions. The implementation of record manipulation techniques involves preprocessing the raw chat data to extract relevant information and organizing it into a structured format. This structured data can then be used to generate analytical plots and perform statistical analysis using the tool's functionalities.

1. Analytical Plots:

- The tool generates various analytical plots to visualize key aspects of the WhatsApp conversations. These plots include:
  - Frequency of Interactions: A graphical representation showcasing the most frequent interactions among group participants. This plot highlights the individuals who are actively engaged in the conversation, as well as the distribution of interactions across the group.
  - Sentiment Trends: Graphs illustrating the sentiment trends over time within the chat. These plots help identify fluctuations in sentiment and pinpoint specific events or topics that trigger emotional responses among participants.
  - Word Clouds: Visual representations of the most commonly used words in the chat, with word size indicating frequency. Word clouds offer a quick overview of the main themes and topics discussed within the conversation.
  - Emoji Analysis: Analysis of emoji usage within the chat, including the most frequently used emojis and their associated sentiments. This provides insights into the emotional tone and dynamics of the conversation.

2. Advocacy for Record Manipulation Techniques:

- The tool advocates for the application of record manipulation techniques to enhance the comprehension of WhatsApp chats on mobile devices. By utilizing these techniques, users can efficiently navigate through large chat records, filter relevant information, and extract actionable insights.
- Record manipulation techniques include features such as search functionality, filtering by date or participant, and summarization tools. These features empower users to effectively manage and analyze WhatsApp conversations, ultimately improving comprehension and decision-making.

.

Conclusion

In conclusion, the project aimed to analyze WhatsApp chat data, focusing particularly on sentiment analysis. Through the utilization of Python libraries such as Pandas, NLTK, and Matplotlib, as well as techniques like regular expressions and data visualization, valuable insights were extracted from the chat data.

The project began with data preprocessing steps, including data cleaning, parsing, and formatting. Various features of the chat data, such as message frequency, active users, and message lengths, were analyzed to gain a comprehensive understanding of the communication patterns within the chat group.

Subsequently, sentiment analysis was conducted to assess the overall sentiment expressed in the chat messages. By employing sentiment analysis techniques, including the VADER sentiment analyzer, the polarity of each message was determined, allowing for the categorization of messages into positive, negative, and neutral sentiments. Visualizations, such as sentiment distribution plots and word clouds, were employed to provide intuitive representations of the sentiment analysis results.

Overall, the project demonstrated the effectiveness of python techniques in analyzing WhatsApp chat data and extracting meaningful insights. The insights gained from sentiment analysis can be utilized for various purposes, including improving communication strategies, detecting sentiment trends, and enhancing the overall user experience of the chat group. Additionally, the project highlighted the importance of data-driven approaches in understanding human communication patterns and behaviours in digital communication platforms.

`