IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

ChatProbe Profiling WhatsApp Conversations Using Machine Learning Approaches

Sravanthi D[1], Lepakshi Reddy S[2], Vittal Sai C[3], Jahnavi S[4], Vamsi C[5]

Assistant Professor, Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyala, 518501.[1]

Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyala, 518501. [2-5]

Abstract: WhatsApp has become integral to modern communication, yet managing unwanted messages and group notifications presents challenges for effective analysis. The model present in paper will address this issue by developing a robust chat analyzer capable of handling such content. Utilizing Python libraries like pandas, seaborn, and matplotlib, alongside advanced natural language processing, the analyzer identifies and filters out irrelevant messages and group notifications. A preprocessing module ensures that subsequent analysis focuses on meaningful conversations, while sentiment analysis provides insights into user interactions. Deployed as a user-friendly application, the analyzer offers comprehensive visualization and statistical analysis of chat data. Through interactive features, users gain valuable insights into conversation dynamics while efficiently managing unwanted content. Incorporating machine learning and sentiment analysis, this project presents a versatile solution for WhatsApp conversation analysis, empowering users to extract meaningful information while mitigating the impact of unwanted content. It is deployed on the Heroku web platform, utilizes a combination of Python libraries such as matplotlib, streamlit, seaborn, re, pandas, and concepts of natural language processing. This amalgamation of machine learning and NLP techniques enables the tool to import WhatsApp chat files, analyze them, and generate various visualizations, enhancing comprehension of the data.

Keywords: WhatsApp, Chat Analysis, Emoji Analysis, Emotion Analysis, Sentiment Analysis, Preprocessing, Natural Language Processing, Data Visualization, Machine Learning, Python Libraries, Pandas, Seaborn, Matplotlib, Text Analysis, Behavioural Analysis, Group Notifications, Unwanted Messages, User Interaction.

I. INTRODUCTION

In today's digital era, WhatsApp stands out as the premier platform for communication, connecting billions of users worldwide. Within this expansive network, diverse conversations unfold, encompassing individual chats and group discussions on a myriad of topics. The richness of these interactions holds invaluable insights into human behaviour, [2]emotions, and social dynamics. ChatProbe emerges as a powerful tool for delving into the depths of WhatsApp conversations, offering a nuanced perspective on the intricacies of human communication. With a focus on profiling and analysis, chatprobe transcends mere data examination, striving to uncover hidden patterns, [4]sentiments, and behavioural traits embedded within WhatsApp chats.

The proliferation of multimedia elements, including emojis, images, and videos, adds layers of complexity to the analysis process. Through sophisticated techniques such as sentiment analysis and emotion classification, ChatProbe endeavours to decode the underlying sentiments conveyed through text and multimedia content. By harnessing the power of natural language processing (NLP) and machine learning, the tool aims to discern emotions ranging from joy and sadness to anger and excitement, painting a holistic picture of the emotional landscape within [1]WhatsApp conversations.

Moreover, ChatProbe addresses practical challenges encountered in managing WhatsApp interactions, including the influx of unwanted messages, spam, and group notifications. By integrating advanced filtering mechanisms, the tool seeks to streamline the analysis process, enabling users to focus on meaningful interactions while eliminating noise and distractions. At its core, ChatProbe is built upon a foundation of robust data preprocessing and analysis.

Leveraging custom preprocessing scripts and Python libraries such as pandas, numpy, and seaborn, the tool streamlines the data preparation process, paving the way for insightful analytics. Through interactive visualizations and intuitive dashboards, users gain a comprehensive understanding of their WhatsApp interactions, empowering them to make informed decisions and glean actionable insights.

 This work is licensed under a Creative Commons Attribution 4.0 International License

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13423

This paper provides a comprehensive overview of ChatProbe, detailing its methodology, core functionalities, and the underlying principles driving its development. By synthesizing insights from previous research on emotion classification, sentiment analysis, and texting psychology, ChatProbe aims to push the boundaries of WhatsApp conversation analysis, offering novel perspectives on human communication in the digital age.

The paper is organised as follows: Section 2 highlights the Literature Survey, Section 3 highlights the proposed work that shows how our proposed model works, Section 4 includes the results and discussions of whatsapp chat data which creates various visualisation graphs using various python libraries, section 5 includes conclusion of the project, Section 6 highlights the future scope, section 7 includes the references which I take for my project.

II. LITERATURE SURVEY

1. Literature Review on WhatsApp Chat Analysis:

WhatsApp Messenger has emerged as a ubiquitous platform for communication, shaping social interactions and digital behaviours across diverse demographics. A survey analysis on the usage and impact of WhatsApp Messenger reveals compelling insights into the pervasive influence of this messaging app, particularly among young users in Southern India.

The survey highlights the staggering amount of time spent on WhatsApp, with users between the ages of 18 to 23 dedicating approximately 8 hours daily to the platform, sometimes even being online for up to 12-16 hours a day. This extensive usage underscores WhatsApp's status as the most widely utilized app on smartphones, surpassing other social media platforms in terms of user engagement.

The survey further illuminates the multifaceted nature of WhatsApp interactions, encompassing the exchange of images, audios, and videos, thereby enriching the communication experience. However, alongside its widespread adoption, the survey also seeks to uncover the positive and negative impacts of WhatsApp usage, prompting the exploration of tools and methodologies aimed at providing users with insights into their chat interactions and revealing previously unknown facets of their digital conversations.

❖ **Transformative Solutions:** The WhatsApp Chat Analyzer revolutionizes the process of analyzing WhatsApp chats by providing a user-friendly interface that streamlines analysis tasks. This departure from traditional methods of dealing with raw, unstructured text data marks a significant advancement in chat analysis.

❖ **Accessibility and Convenience:** One of the key advantages of the WhatsApp Chat Analyzer is its universal accessibility, running seamlessly on all devices. This ensures convenience for users across diverse platforms, enhancing the usability and reach of the platform.

❖ **Overcoming Data Limitations:** By leveraging WhatsApp's export functionality to obtain chat data in text format, the analyzer effectively addresses the limitations of raw data. This enables comprehensive analysis of chat interactions, leading to more accurate and insightful findings.

❖ **Rich Visualizations and Insights:** The WhatsApp Chat Analyzer offers users a rich array of visualizations and insights, providing valuable metrics such as total messages, total words, media shared, links shared, monthly timelines, and peak activity days. These features empower users to gain deeper insights into their WhatsApp usage patterns and communication dynamics.

❖ **Evolution of Chat Analysis:** The emergence of the WhatsApp Chat Analyzer reflects the evolving landscape of WhatsApp chat analysis, transitioning from manual methods to modern, analytics-driven platforms. This evolution highlights the growing importance of leveraging advanced analytics techniques to unlock the full potential of WhatsApp conversations.

2. Literature Review on Python:

Python has emerged as a versatile and powerful programming language, offering a range of benefits for scientific research and data analysis. With its easily understandable syntax and rich set of features, Python has become a preferred choice for researchers and professionals alike. Its statistics-oriented capabilities and strong support for data visualization make it particularly well-suited for tasks involving theoretical calculations and data analysis. Unlike other proprietary languages like MATLAB or LabView, Python's open-access nature ensures that researchers have access to a wide range of tools and libraries necessary for high-quality scientific work. The flexibility and efficiency of Python empower researchers to tackle complex data analysis tasks with ease, enhancing productivity and accelerating research outcomes.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13423

Python's suitability for educational purposes has been highlighted in various studies. Deploying Python as a first programming language has been shown to significantly reduce the learning curve for students and improve their problem-solving efficiency. Researchers have observed that tasks that took hours to solve in languages like C++ could be accomplished in significantly less time using Python. This accessibility and ease of use make Python an invaluable resource for professors and students alike, facilitating learning and enabling rapid prototyping of ideas. Additionally, Python's broad compatibility with various operating systems further enhances its appeal, ensuring that researchers and students can seamlessly utilize Python across different computing environments for their scientific and educational endeavours.

3. Literature Review on Modules:

❖ **Streamlit:** Streamlit is a free and open-source python framework.We can quickly develop web apps for Machine Learning and Data Science by using Streamlit. Streamlit can easily integrates with other popular python packages such as NumPy, Pandas, Matplotlib, Seaborn. Streamlit provides fastest way to develop and deploy web apps.

❖ **UrlExtract:** This library provides functionality for extracting URLs from text data. It simplifies the process of identifying and extracting URLs, which is useful for tasks such as link analysis and content categorization.

❖ **Word Cloud:** Word Cloud is a data visualization library used for representing most frequently used words within a given text. Most frequent and important words are represented in bigger and bolder size

❖ **Pandas:** Pandas is a powerful data manipulation and analysis library in Python, primarily used for working with structured data. It provides convenient data structures (DataFrame and Series) and functions for tasks such as data cleaning, transformation, and analysis.

❖ **Collections:** It provides specialized data structures beyond the built-in data types like lists and dictionaries. The 'counter' class is utilized for efficient counting of elements, such as word frequencies in text data.

❖ **TextBlob:** It is a Python library for processing textual data, offering simple API methods for common natural language processing (NLP) tasks. It provides functionalities such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and language translation.

❖ **Scikit-Learn(sklearn):** Scikit-learn is a popular machine learning library in Python, offering a wide range of tools and algorithms for data mining and analysis. It uses **LatentDirichletAllocation** Algorithm for topic modelling and text feature extraction.

❖ **Networkx:** It is a Python library for creating, manipulating, and analyzing complex networks or graphs. It provides functions and algorithms for tasks such as graph visualization, centrality analysis, and community detection.

❖ **VaderSentiment:** It is a Python library for sentiment analysis, specifically designed for analyzing sentiments expressed in social media texts. It provides a pre-trained model that assigns sentiment scores to text data, capturing both polarity (positive, negative) and intensity.

❖ **Matplotlib:** It is a plotting library in Python used for creating static, interactive, and animated visualizations. It offers a wide range of functions for creating various types of plots, including line plots, bar charts, histograms, scatter plots, and more.

❖ **Seaborn:** Seaborn is the data visualization library. It is used for making statistical graphs. Visualization is the central part of seaborn. Seaborn provides exploration and better understanding of data. Seaborn closely integrates into the data structures from python.

III. PROPOSED WORK

1. Data Analysis: It is a process of cleaning, transforming, inspecting and modelling data with the goal of discovering some useful information and finally indicating some conclusions. Analysis means it breaks a whole component into its separate components for individual examination. [3]Data analysis is a process for acquiring raw data and transforming it into useful information for decision-making by users. This project provides a basic statistical analysis WhatsApp chat.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

Following are the analysis made:

- ❖ To find total messages, total words, total media, total links shared and total deleted messages in the WhatsApp chat
- ❖ To find daily, weekly and monthly timeline in the form of graphs.
- ❖ To find most busy users in the group
- ❖ To find most busy day and busy month
- ❖ To find weekly activity from hour to hour
- ❖ To find sentiment trends overtime
- ❖ To find the most frequently and commonly used words in the group.
- ❖ To find topic modelling and its visualization
- ❖ To calculate network centrality measures
- ❖ To find whether emotions are positive or negative or neutral
- ❖ To find the most used emojis in the group.

2. Proposed System

Data pre-processing is the initial part of the project, it is to understand the implementation and usage of various python inbuilt modules. These various modules provide better user understandability and code representation. The following libraries are used such as NumPy, pandas, matplotlib, sys, re, emoji, seaborn etc.

It analyses the data and gives [8] top statistics like total messages, total media, links, images shared, graphs showing the activity map weekly and monthly, monthly timeline, daily timeline, mostly busy users, chart most common words used, emojis used.

The working of proposed system is as follows:

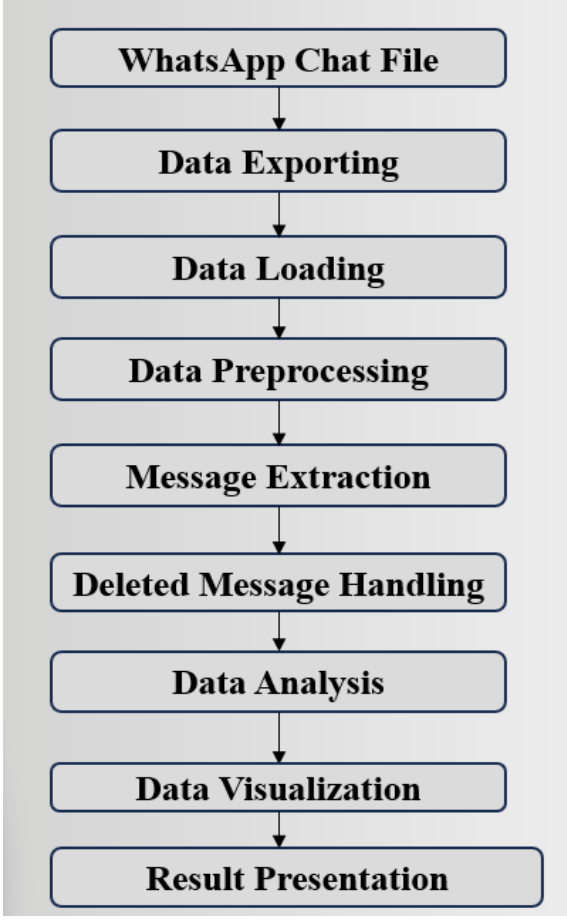


Fig-1: Flow chart of Proposed System

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

3. Working:

A. WhatsApp Chat File: A WhatsApp chat file typically includes messages, media files (such as images and videos), emojis and symbols used in messages, links shared within the conversation, and participant information. Deleted messages may also be present but marked as inaccessible. It contains a wealth of data that can be analyzed to gain insights into communication patterns, sentiment and content sharing.

B. Data Exporting: Open WhatsApp chat for a group ->click on the menu ->click on more- ->select export chat->choose without media. Working of WhatsApp chat analysis.

C. Data Loading: The exported chat file is loaded into the streamlit framework.

D. Data Preprocessing: After loading the data, it will automatically converts the time format of chat into 24hours at the initial stage of data pre-processing. It also removes the group notifications which are present in the exported file before converting data into dataframes using pandas.

E. Message Extraction: The messages in the file will be separated based on a pattern using [7] regular expressions.

F. Deletion of Unwanted Messages: In the loaded chat file we have the following unwanted messages, which are shown in Fig-2.

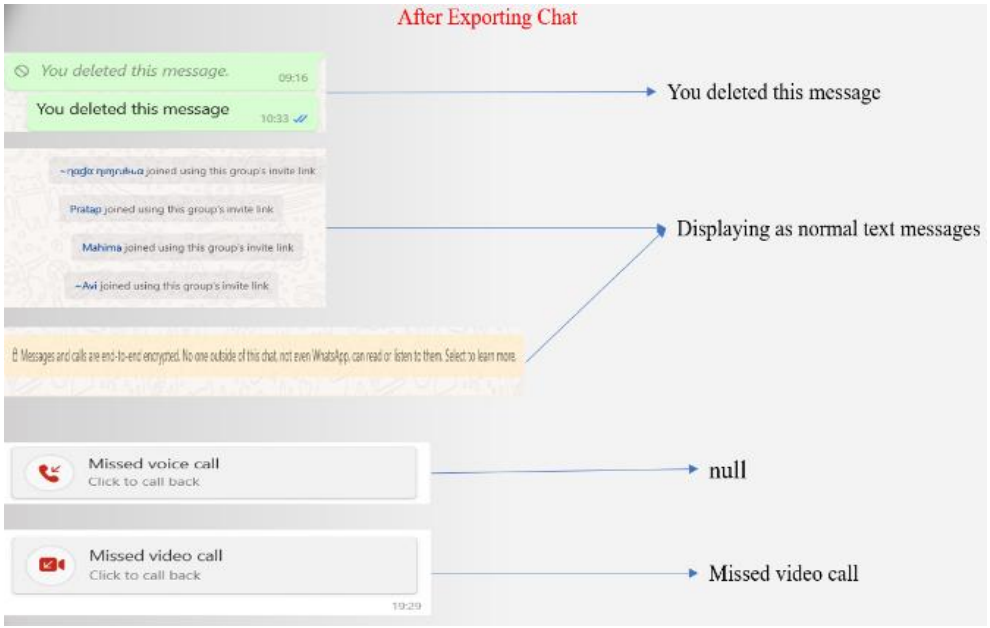


Fig-2: Unwanted Messages in chat file

G. Data Analysis and Visualization: After removing the unwanted messages we can do the following analysis-

❖ **[1]Basic Statistics:** It includes the count of total messages, total words, total media files, total links shared, total deleted messages, message frequency overtime (daily, weekly, monthly) and determines the most active participants.

❖ **[5]Sentiment Analysis:** Analyze the sentiment of messages (positive, negative, neutral) by using the library called TextBlob. It can also plots the sentiment trends overtime.

❖ **[9]Word Frequency Analysis:** It can count the frequency of words and phrases and creates a word cloud for visualization.

❖ **Topic Modelling:** It identifies the topics of conversation using Latent Dirichiet Allocation and extract the keywords associated with each topic and creates visualization.

❖ **Network Analysis:** It constructs a network of interactions between participants and analyze the network centrality measures (degree centrality, betweenness centrality). It visualizes the network graph using networkx library.

❖ **[10]Emotion Analysis:** It detects the emotions expressed in messages (happiness, sadness, anger etc.,)

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

IV. RESULTS AND DISCUSSIONS

1. Dataframe:

	date	user	message	only_date	year	month_num	month	day	day_name	hour	minute	period	receiver
0	2022-05-21 17:15:00	R T	['<Media omitted>\n']	2022-05-21	2,022	5	May	21	Saturday	17	15	17-18	group_notificatior
1	2023-07-07 11:52:00	+91 99851 97806	['<Media omitted>\n']	2023-07-07	2,023	7	July	7	Friday	11	52	11-12	Naveen P
2	2023-07-11 13:47:00	Naveen P	['<Media omitted>\n']	2023-07-11	2,023	7	July	11	Tuesday	13	47	13-14	+91 99851 97806
3	2023-07-13 15:41:00	+91 99851 97806	['<Media omitted>\n']	2023-07-13	2,023	7	July	13	Thursday	15	41	15-16	+91 6305 504 540
4	2023-08-08 12:45:00	+91 6305 504 540	['<Media omitted>\n']	2023-08-08	2,023	8	August	8	Tuesday	12	45	12-13	+91 6305 504 540
5	2023-08-08 12:45:00	+91 6305 504 540	['<Media omitted>\n']	2023-08-08	2,023	8	August	8	Tuesday	12	45	12-13	+91 6305 504 540
6	2023-08-08 12:45:00	+91 6305 504 540	['<Media omitted>\n']	2023-08-08	2,023	8	August	8	Tuesday	12	45	12-13	+91 97878 90752
7	2023-08-09 08:26:00	+91 97878 90752	['<Media omitted>\n']	2023-08-09	2,023	8	August	9	Wednesday	8	26	8-9	+91 97878 90752
8	2023-08-09 08:26:00	+91 97878 90752	['<Media omitted>\n']	2023-08-09	2,023	8	August	9	Wednesday	8	26	8-9	+91 97878 90752
9	2023-08-10 09:19:00	+91 97878 90752	['<Media omitted>\n']	2023-08-10	2,023	8	August	10	Thursday	9	19	9-10	group_notificatior
10	2023-09-07 12:18:00	+91 99851 97806	['<Media omitted>\n']	2023-09-07	2,023	9	Septembe	7	Thursday	12	18	12-13	+91 99851 97806

Fig-3: Text chat into Dataframe

This Fig-3 shows the[8] dataframe which organizes the chat data with columns like ‘user’, indicating message senders, and ‘message’, storing message content, facilitating text analysis. ‘Date’ stores message timestamps.

The ‘period’ columns categorizes messages temporarily, supporting analysis of message frequency and activity patterns overtime by excluding group notifications at the initial stage of data preprocessing.

2. Basic Statistics:

It can analyze the dataframe and plot them in the form of graphs by excluding the unwanted messages.

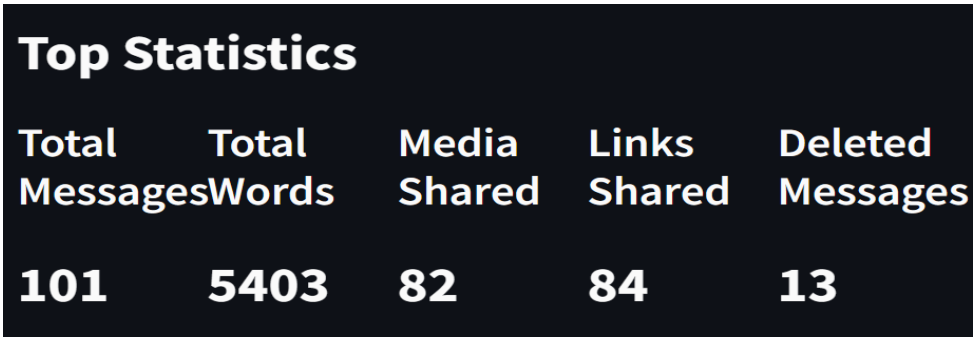


Fig- 4: Various Statistics

This Fig-4 shows the count of total messages, words, media files and links shared by excluding the deleted messages which will cause false analysis. It can also displays the count of deleted messages.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

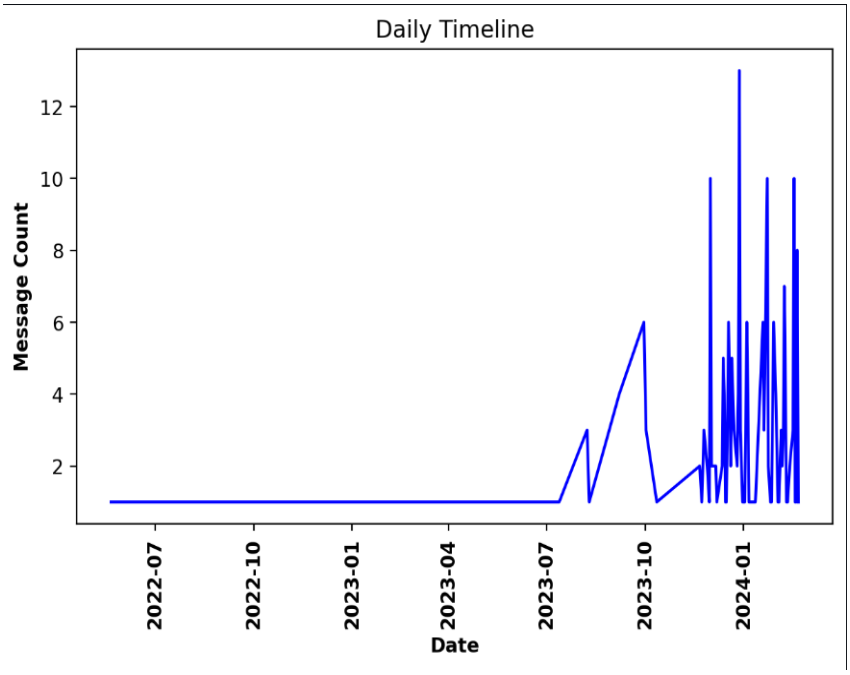


Fig-5: Daily Timeline

This Fig-5 gives the[2] 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.

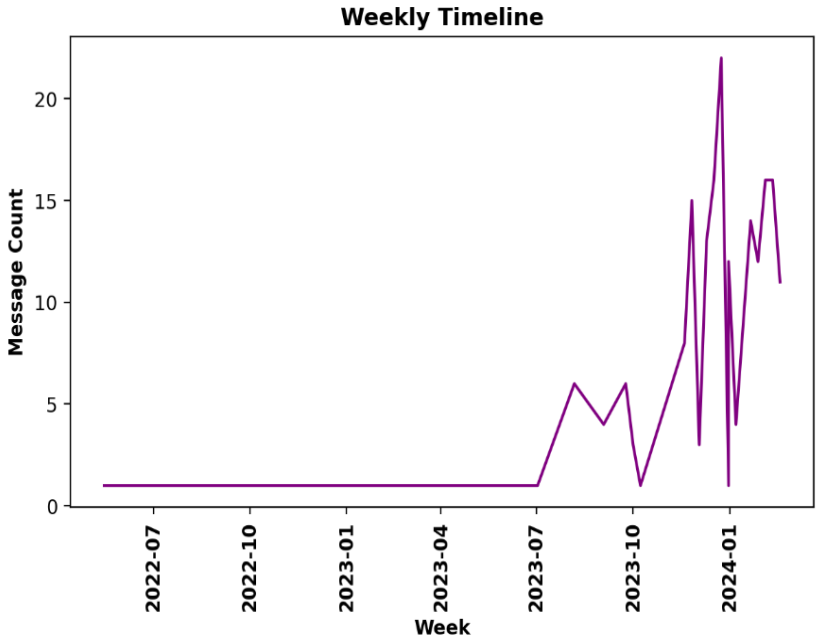


Fig-6: Weekly Timeline

This Fig-6 gives the frequency of messages in a week We have used matplotlib to plot the graph and the weeks are taken and the count of messages are calculated and plotted.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

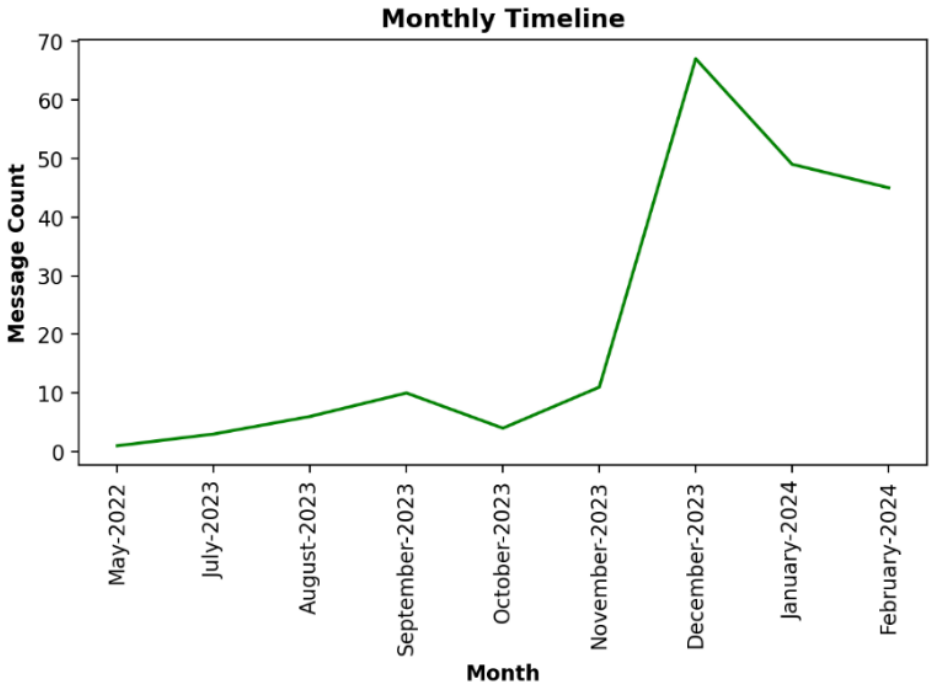
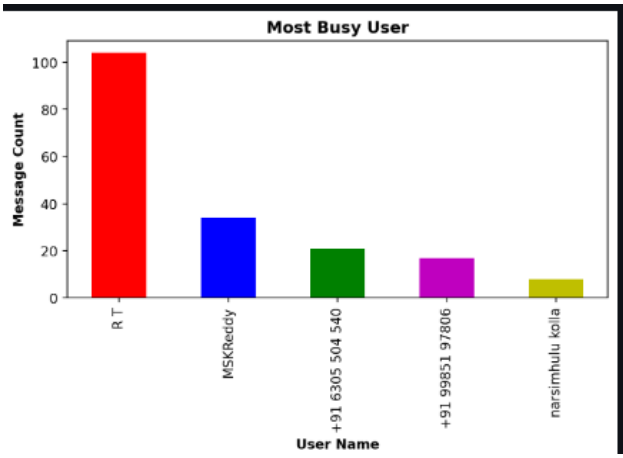


Fig-7: Monthly Timeline

This Fig-7 gives the frequency of messages in a month We have used matplotlib to plot the graph and the months are taken and the count of messages are calculated and plotted.



	name	percent
0	R T	53.06
1	MSKReddy	17.35
2	+91 6305 504 540	10.71
3	+91 99851 97806	8.67
4	narsimhulu kolla	4.08
5	+91 97878 90752	2.55

Fig-8: Most Active users

This Fig-8 visualizes [4]message activity trends over months, facilitating temporal analysis. It provides a timeline of message counts aggregated by year and month, aiding in understanding communication patterns.

For identifying the busiest users, the code presents a bar chart illustrating message counts and a dataframe with user statistics. This dual presentation enhances comprehension of user engagement dynamics within the group.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

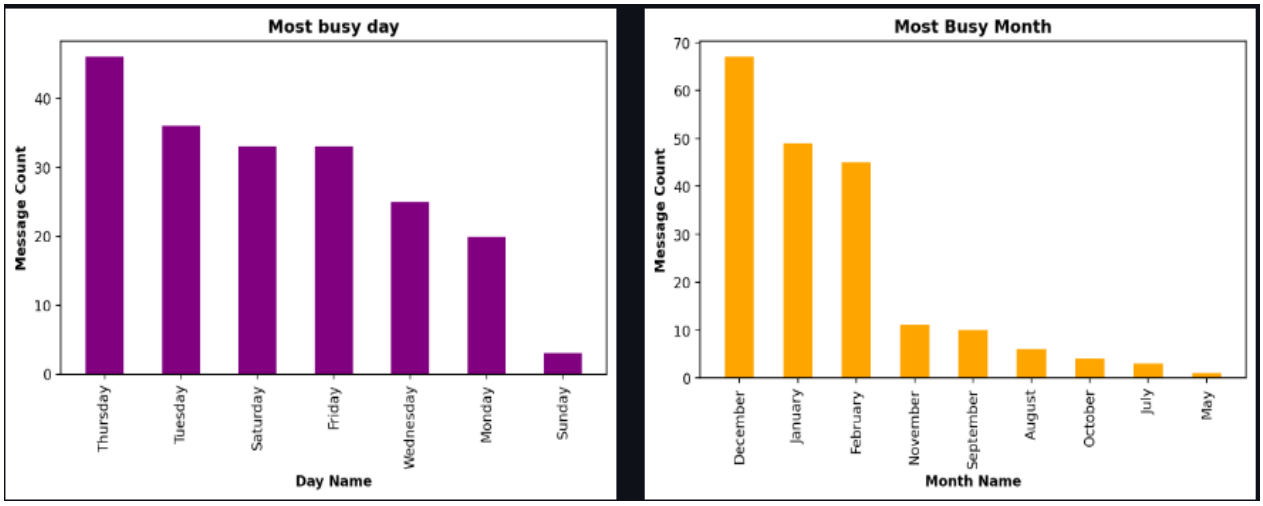


Fig-9: Most Busy Day and Month

This Fig-9 generate visualizations illustrating message activity trends over days and months, respectively. They utilize the Pandas library for data manipulation and [2]Streamlit for interactive visualization.

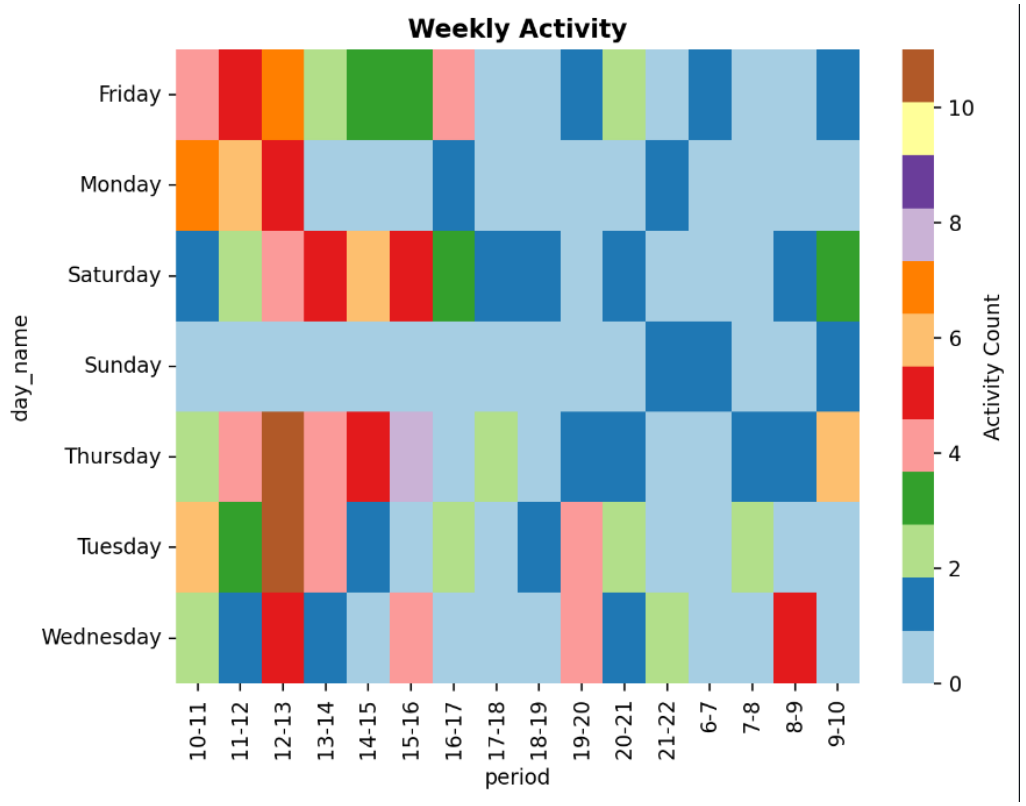


Fig-10: Weekly Activity Map

This Fig-10 generates a[9] heatmap illustrating message activity patterns over days and time periods within a week. It leverages Pandas for data manipulation and Seaborn for heatmap visualization.

It visualizes message counts across days and time periods, providing insights into peak communication times. The colour intensity represents the message activity level, with darker shades indicating higher activity.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

3. Sentiment Analysis:

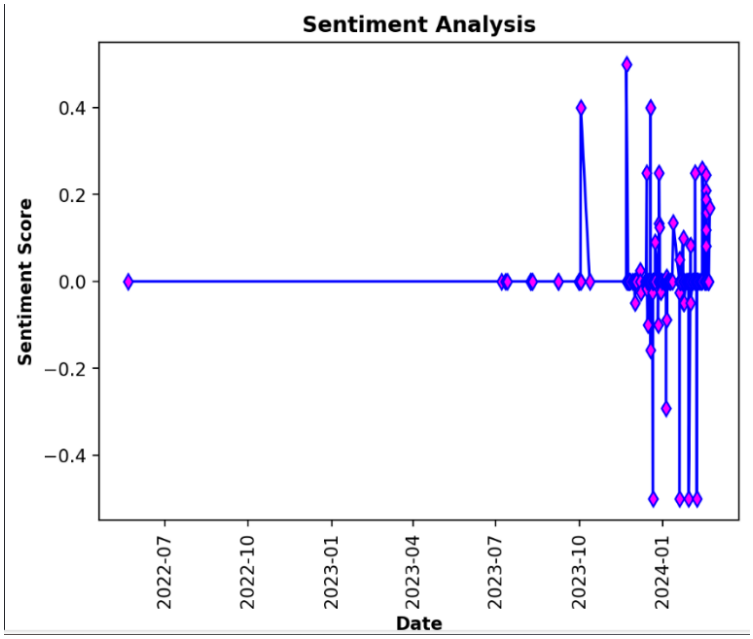


Fig-11: Sentiment Trends Over Time

This Fig-11 analyzes the sentiment of messages in the DataFrame, assigning sentiment scores based on polarity using [10]TextBlob. Matplotlib is utilized to visualize the sentiment trends over-time. This plot illustrates the sentiment scores of messages over the timeline, providing insights into the overall sentiment dynamics. Positive sentiment scores indicate positive sentiment, while negative scores suggest negative sentiment.

4. Word Frequency Analysis:

This Fig-12 generates a word cloud visualization based on the messages from the selected user in the DataFrame. It preprocesses the messages by removing stop words and filtering out non-string messages before constructing the word cloud using the[4] WordCloud library.

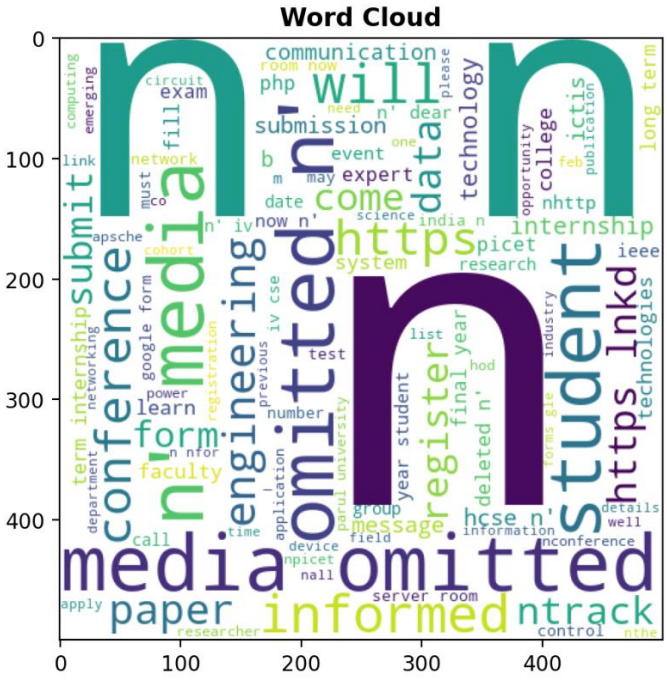


Fig-12 Word Cloud Frequency

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

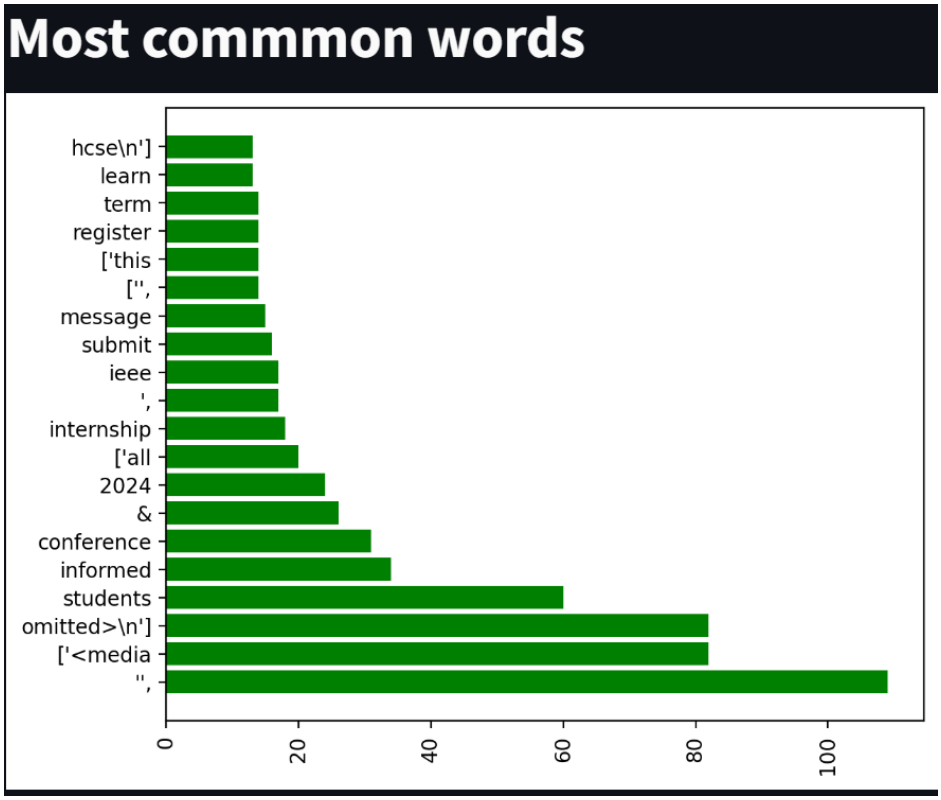


Fig-13: Most Common Words

This Fig-13 shows the most commonly used words in the chat by using WordCloud library.

5. Topic Modelling:

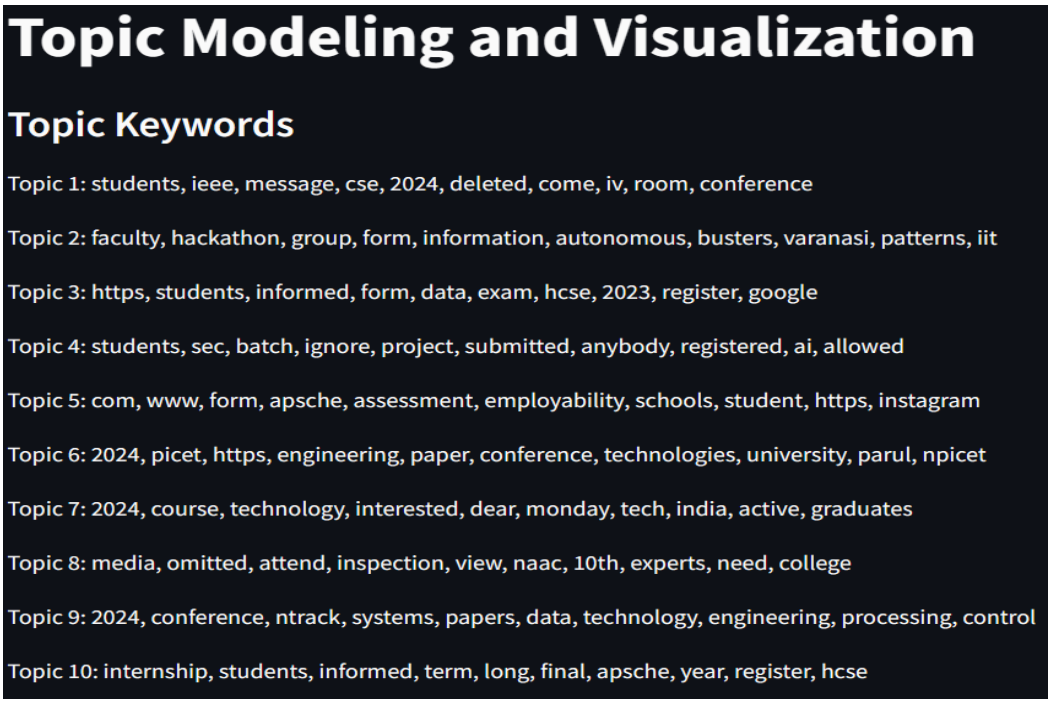


Fig-14: Topic Keywords

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

Topic Distributions

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
1	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
2	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
3	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
4	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
5	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
6	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
7	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
8	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333
9	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.7	0.0333	0.0333

Fig-15: Topic Distributions

This Fig-14&15 utilizes [6]Latent Dirichlet Allocation (LDA), a technique for topic modelling, to analyze the textual content of messages in the DataFrame. It preprocesses the messages, converts them into a matrix representation using [6]CountVectorizer, and then applies LDA to identify latent topics within the messages.

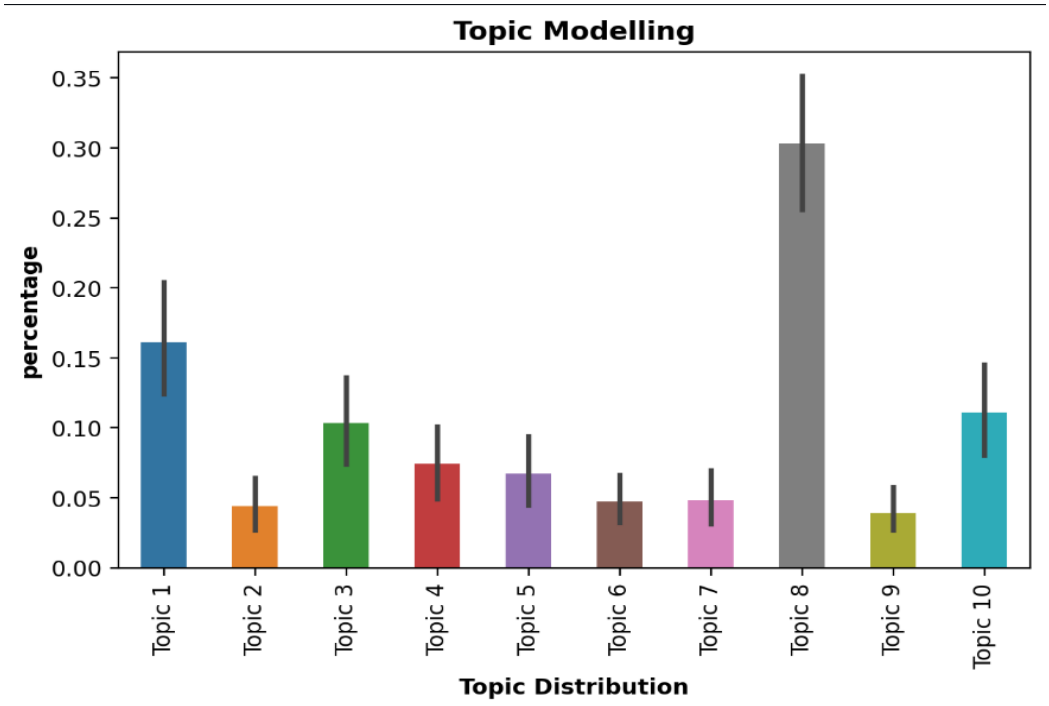


Fig-16: Topic Modelling

This Fig-16 displays the results of the topic modelling process. It presents the keywords associated with each topic, the distribution of topics across the messages, and a bar plot visualizing the distribution of topics.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

6. Network Analysis:

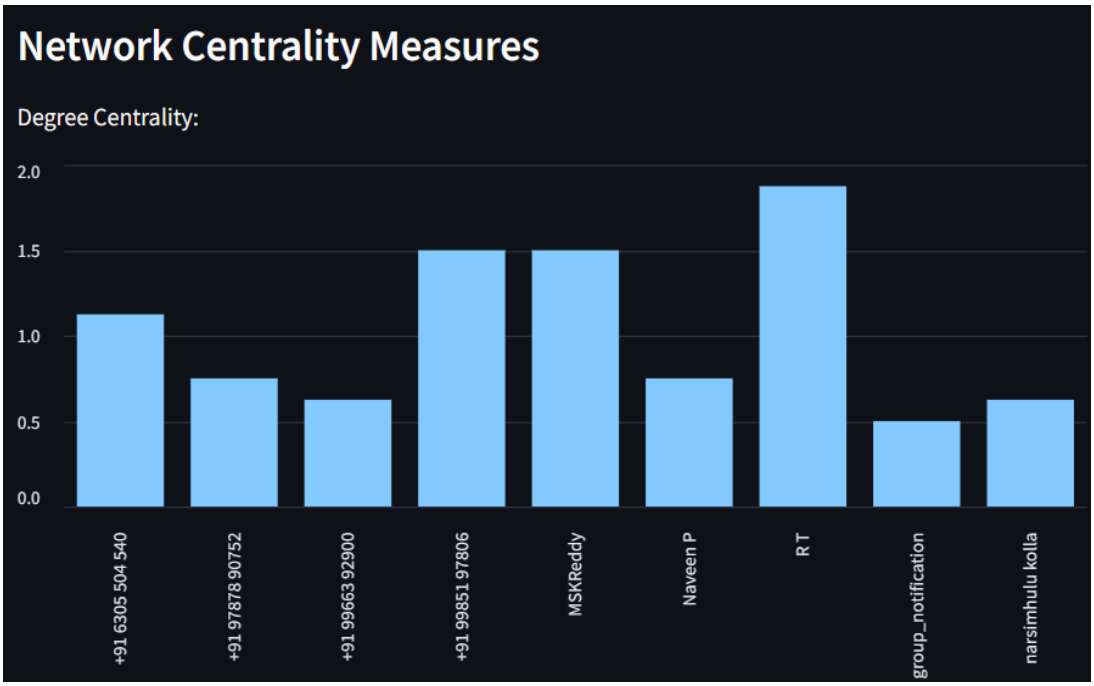


Fig-17: Degree Centrality

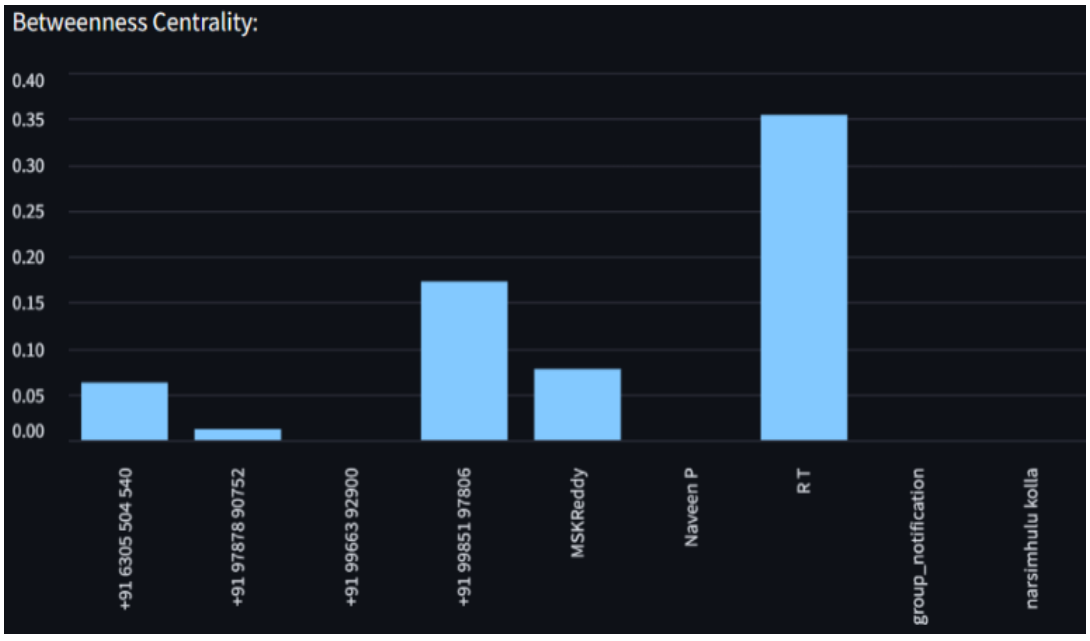


Fig-18: Betweenness Centrality

This Fig-17&18 constructs a directed graph representing participant interactions. It creates edges between senders and receivers based on message data, forming a network model of communication dynamics.

The "Network Centrality Measures" section analyzes degree and betweenness [1]centrality within the graph. Degree centrality quantifies participants' connections, while betweenness identifies influential intermediaries.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13423

7. Emotion Analysis:

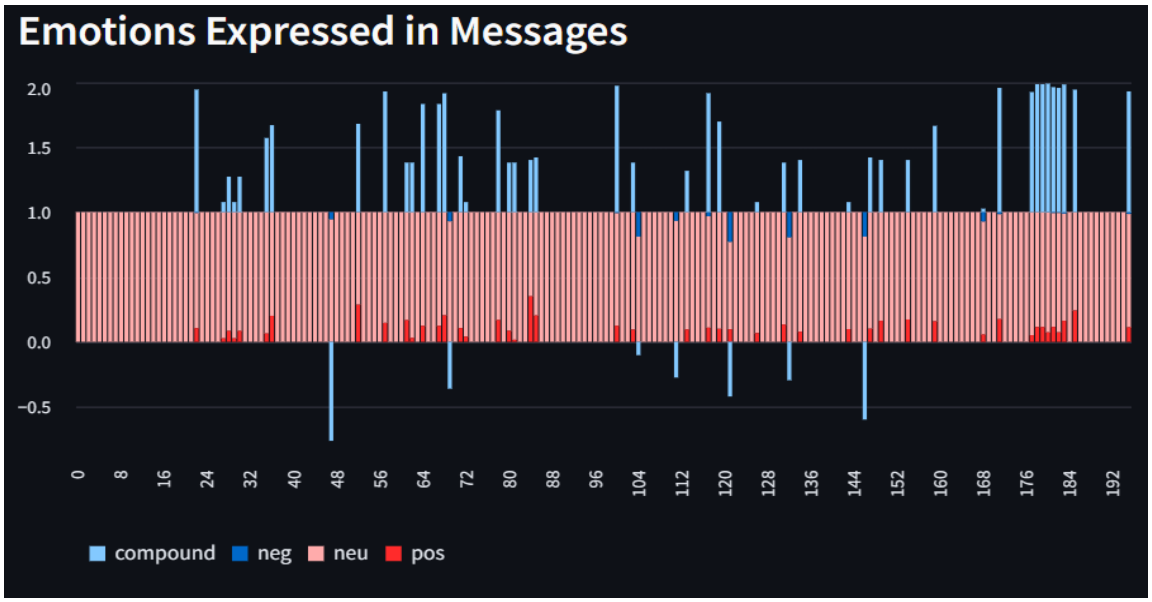


Fig-19: Emotions Analysis

This Fig-19 utilizes the [8] sentiment analysis tool to assess emotions expressed in messages. It computes sentiment scores for each message, capturing emotional nuances such as positivity, negativity, and neutrality. The resulting emotion scores are aggregated into a dataframe for visualization.

8. Emoji Analysis:

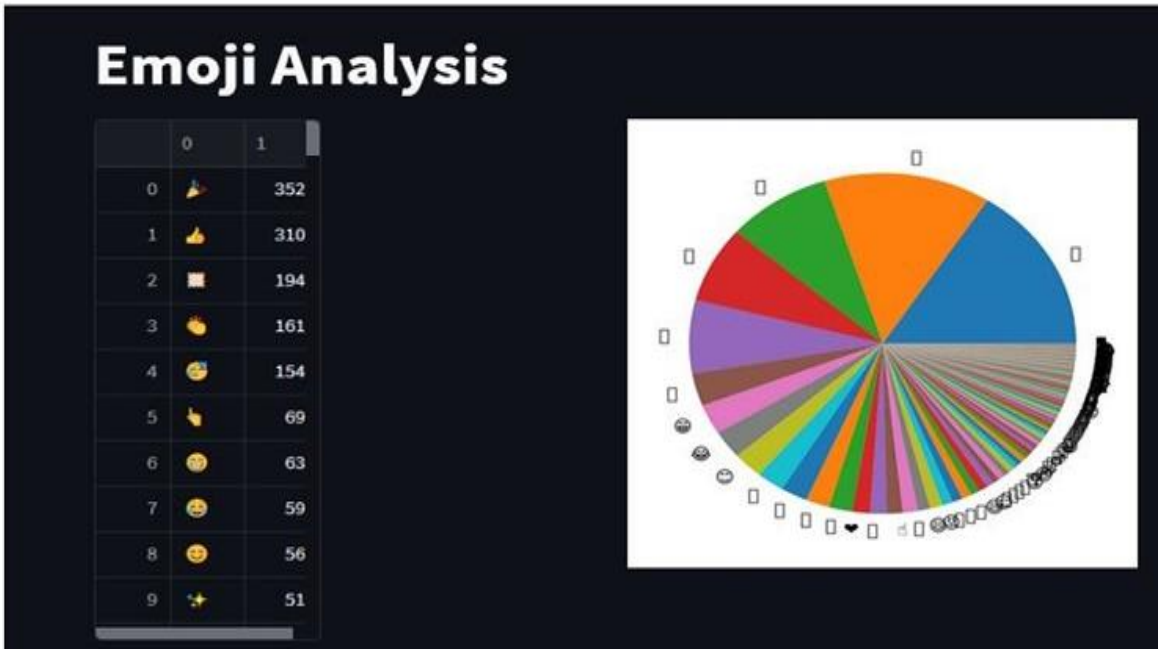


Fig-20: Emoji Analysis

This Fig-20 shows the most commonly used emojis We have used the [1]Emoji library to select or distinguish the emojis from the messages and plotted the pie chart using matplotlib

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13423

V. CONCLUSIONS

In conclusion, Python's versatility in network data analysis, coupled with the robust features of the WhatsApp application, demonstrates the potential for comprehensive chat data exploration and interpretation. By employing Python libraries and adhering to a menu-based and user-friendly design, the system ensures accurate and reliable outcomes, catering to users with varying levels of technical proficiency. The system's ability to analyze any WhatsApp group data and its validation feature to prevent erroneous data entry underscore its effectiveness in meeting the objectives outlined during the requirement analysis phase, thus offering a valuable tool for understanding and managing WhatsApp conversations.

VI. FUTURE SCOPE

Future iterations of this project could focus on refining the preprocessing stage to handle a wider range of message formats and languages more effectively. Enhancing the user interface with more intuitive controls and visualizations could improve the overall user experience and accessibility. Incorporating advanced machine learning techniques, such as deep learning models, could yield more accurate sentiment analysis and topic modelling results.

REFERENCES

[1]. D.Radha, R. Jayaparvathy, D. Yamini, “Analysis on Social Media Addiction using Data Mining Technique”, International Journal of Computer Applications (0975 – 8887).
[2]. Python for Everybody: Exploring Data in Python 3 by Dr. Charles Russell Severance.
[3]. Storytelling with Data: A Data Visualization Guide for Business Professionals by Cole Nussbaumer Knaflic.
[4]. Ravishankara K, Dhanush, Vaisakh, Srajan I S, “International Journal of Engineering Research & Technology (IJERT)”, ISSN: 2278-0181, Vol. 9 Issue 05, May-2020
[5]. Meng Cai, “PubMed Central”, PMCID: PMC7944036, PMID: 33732917
[6]. Dr. D. Lakshminarayanan, S. Prabhakaran, “Dogo Rangsang Research Journal”, UGC Care Group I Journal, Vol-10 Issue-07 No. 12 July 2020
[7]. Ahmed, I., Fiaz, T., “Mobile phone to youngsters: Necessity or addiction”, African Journal of Business Management Vol.5 (32), pp. 12512-12519, Aijaz, K. (2019).
[8]. 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” (2022).
[9]. 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 2021.
[10]. Jessica Ho, Ping Ji, Weifang Chen, Raymond Hsieh, “Identifying google talk”, IEEE International Conference on Intelligence and Security Informatics, ISI ‘09, pp. 285-290, 2019.
[11]. Mike Dickson, “An examination into AOL instant messenger 5.5 contact identification.”, Digital Investigation, ScienceDirect, vol. 3, issue 4, pp. 227-237, 2021.
[12]. Mike Dickson, “An examination into yahoo messenger 7.0 contact identification”, Digital Investigation, ScienceDirect, vol. 3, issue 3, pp. 159-165, 2022.

 This work is licensed under a Creative Commons Attribution 4.0 International License