ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 547-557, 2026

AI-ASSISTED ANALYSIS AND VISUALIZATION OF WHATSAPP CONVERSATIONS USING NATURAL LANGUAGE PROCESSING

B Renuja[1*], Rahul Sahu[2], Shankar Sharan Tripathi[3]

[2]*Department of Computer Science, Shri Shankaracharya Technical Campus, Junwani.*

[3]*Assisant Professor, Computer Science and Engineering, Shri Shankaracharya Technical Campus*

Abstract

The rapid expansion of digital communication platforms has resulted in a significant increase in informal textual data, providing new opportunities to analyze behavioral patterns, language usage, and social interaction dynamics. This study introduces a **WhatsApp Chat Analyzer**, an interactive web-based application developed using Python and Streamlit, designed to process and analyze exported WhatsApp conversations from both group and individual chats. By applying Natural Language Processing (NLP) techniques along with statistical analysis, the system extracts important communication metrics such as message frequency, peak activity hours, frequently used words, emoji distribution, and user engagement trends. The platform also generates various visual representations, including temporal activity charts, word clouds, and heatmaps, enabling users to easily interpret communication patterns. The proposed system aims to provide an accessible analytical tool for researchers, educators, marketers, and social scientists while maintaining a modular design that allows future extensions and research applications.

Keywords: *WhatsApp Analytics, Social Media, Mining, Streamlit, Data Visualization, Chat, Behavior, NLP, Text Mining.*

* *Corresponding author*

1. Introduction

The rapid expansion of digital communication platforms has significantly changed the way individuals and communities communicate, generating vast amounts of unstructured conversational data. Among these platforms, WhatsApp has emerged as one of the most widely used messaging applications worldwide, supporting personal, academic, and professional interactions across diverse environments. The analysis of WhatsApp chat logs can provide valuable insights into social dynamics, communication behaviors, engagement levels, and discussion trends within groups or between individuals. However, extracting meaningful and reliable insights from raw chat exports is challenging due to factors such as inconsistent message formats, the presence of media placeholders and system-generated notifications, multilingual text content, and variations in timestamp formats across different regions. Addressing these challenges requires a well-designed preprocessing pipeline, effective feature extraction methods, and an intuitive analytical framework capable of supporting both exploratory and structured data analysis.

This research proposes a lightweight, end-to-end framework for WhatsApp chat analysis that integrates data preprocessing, feature extraction, and interactive visualization to help users understand communication behavior at multiple levels of detail. The system processes raw WhatsApp chat exports and converts them into a structured tabular format by extracting essential attributes such as sender identity, timestamp, date components, and message content, while effectively filtering system-generated notifications and media placeholders. Based on this structured dataset, the framework computes several descriptive statistics, including total message counts, word frequency, media sharing activity, and links shared within the conversation. It also incorporates temporal analysis through daily and monthly timelines to illustrate activity trends over time, along with weekly heatmaps that reveal variations in messaging patterns across different days and hours. To analyze participation levels, the system identifies the most active users and calculates their contributions in terms of both message counts and percentage share. For semantic exploration, the framework generates word cloud visualizations and ranks frequently used terms after removing domain-specific stop words. Additionally, recognizing the importance of emojis in modern digital communication, the system records emoji usage frequencies and visualizes their distribution to better understand emotional expression within conversations.

A key aspect of this study is the implementation of a reproducible preprocessing pipeline designed to accommodate common variations found in WhatsApp chat exports while maintaining accurate temporal representation of messages. The preprocessing stage standardizes timestamps, extracts relevant calendar-based attributes such as year, month, day name, and hour, and categorizes system-generated lines into a unified "group notification" class to prevent analytical bias caused by non-conversational messages. The analytical component of the framework consists of modular functions that can be applied either to the entire conversation dataset or to messages from individual participants. This flexibility enables comparative analysis among users without altering the underlying data structure. To facilitate meaningful interpretation of results, the system incorporates multiple visualization techniques, including timelines to illustrate activity trends, bar charts to display user participation and message distribution, heatmaps to highlight hourly and weekly communication patterns, and pie charts to represent emoji usage. Together, these elements form a structured methodology for transforming raw conversational data into interpretable analytical insights.

This research is relevant across several domains where understanding digital communication behavior is important, including education, organizational studies, digital sociology, mental health research, and community management. By offering an open and extensible analytical framework built using publicly available technologies, the proposed system reduces the technical barriers for researchers and practitioners who wish to study chat-based communication without relying on proprietary software. Furthermore, the modular design of the workflow allows for future enhancements such as sentiment analysis, topic modeling, toxicity detection, interaction network analysis, and multilingual communication analysis. Overall, this framework acts as a bridge between raw WhatsApp chat exports and meaningful analytical insights by providing a practical and validated workflow that supports both exploratory visualization and quantitative analysis, thereby enabling deeper behavioral and linguistic studies of conversational media.

2. Literature Review

In recent years, the analysis of WhatsApp chat data has gained considerable attention across multiple disciplines, including computational social science, digital communication studies, and behavioral research. Several studies have highlighted the potential of conversational data from messaging platforms to reveal complex patterns of social interaction, emotional expression, and information exchange within groups. Early research on WhatsApp chat analysis primarily focused on descriptive statistics, such as message counts, user participation, and media-sharing frequency. These studies established foundational methodologies for preprocessing and cleaning chat data, while also identifying common challenges faced by researchers, including inconsistent timestamp formats, the presence of system-generated messages, and difficulties in accurately identifying participants within conversations.

More recent research has incorporated advanced Natural Language Processing (NLP) techniques to extract deeper insights from conversational data. These approaches extend beyond basic quantitative summaries to enable qualitative analysis of themes, sentiment, and user intent within digital communication. Researchers have utilized tools such as word clouds, frequency analysis, and emoji sentiment evaluation to better capture the informal and expressive nature of instant messaging. Additionally, temporal visualization techniques—including timelines, activity heatmaps, and interaction maps—have been employed to analyze communication rhythms and behavioral cycles. Such analyses help connect messaging patterns with broader social, cultural, and psychological dynamics within digital communities.

From a methodological perspective, several studies have proposed preprocessing pipelines for WhatsApp chat exports in order to standardize datasets and improve the comparability and reproducibility of research findings. These pipelines address challenges such as multilingual text processing, normalization of abbreviations and informal language, and the removal of system-generated or notification messages to reduce analytical noise. Researchers have also explored techniques for segmenting conversations based on participants, enabling individual-level analysis and comparative studies within group interactions. In addition, there has been a growing emphasis on the use of interactive visualization tools that provide intuitive and user-friendly interfaces for exploring complex conversational datasets and extracting meaningful insights.

Despite these developments, several gaps remain in the existing literature. Many previous studies rely on relatively small datasets, limiting the generalizability of their findings and sometimes analyzing only partial segments of conversations rather than the complete dataset. Furthermore, there is a lack of standardized open-source frameworks that integrate preprocessing, statistical analysis, and visualization components specifically designed for WhatsApp chat data. Research focusing on emoji semantics and other multimodal aspects of communication remains relatively limited. Ethical considerations such as privacy protection, data anonymization, and informed consent are also addressed inconsistently across studies. In particular, few frameworks provide comprehensive guidance on incorporating ethical considerations throughout the entire analytical workflow.

This research contributes to the existing body of work by proposing a comprehensive and modular system that integrates automated preprocessing, multi-level statistical analysis, and interactive visualization within a single framework. Designed with accessibility and reproducibility in mind, the system builds upon existing computational tools while addressing the limitations identified in previous studies. By offering an integrated and extensible platform for WhatsApp chat analysis, this work supports both academic research and practical applications in the study of digital communication behavior.

3. Analysis and Methods

This study adopts a structured, step-by-step methodology for analyzing WhatsApp chat data through a flexible framework designed to support scalability and reproducibility of results. The process begins with importing raw WhatsApp chat export files, typically provided in plain text format containing hundreds of messages. A customized preprocessing stage is then applied to clean and standardize the varying message formats. This stage addresses common issues such as media placeholders, system-generated notifications, and inconsistencies in timestamp formats. Using regular expression techniques, the system extracts key attributes including date, time, sender identity, and message content, while also converting timestamps into a standardized format. Group notifications and media-related references are either filtered out or categorized appropriately to ensure that non-conversational elements do not introduce bias into the analytical process.

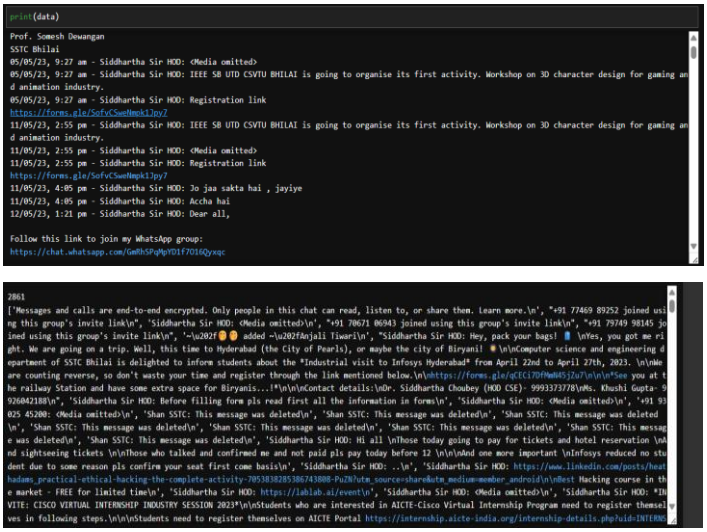

Fig 1: Data of 300+ messages

Following the initial preprocessing stage, the cleaned dataset serves as the foundation for extracting key attributes such as sender identity, message content, date, day of the week, and hour of communication. Additional calendar-based variables—including year, month, and hourly intervals—are also derived to facilitate analysis across different temporal scales. The dataset can then be analyzed either as a complete conversation or segmented by individual participants, enabling the identification of both overall group communication patterns and user-specific behavioral trends.

Subsequently, the system computes several descriptive statistics, including total message counts, word frequencies, and the number of shared media files and links. These metrics are visualized through bar charts that categorize activity by day, month, and participant, allowing users to easily identify peak communication periods and the most active contributors. To further examine temporal trends, the framework generates daily and monthly timelines that illustrate variations in messaging activity over time and highlight changes in engagement levels throughout the conversation.

After calculating weekly activity statistics, the system converts the aggregated data into a heatmap visualization that illustrates messaging intensity across different days of the week and hours of the day. This visualization helps identify peak communication periods within the group. In addition, the system ranks participants based on the number of messages they contribute and calculates each user's percentage share of the total conversation. To identify dominant themes within the discussion, the cleaned textual data is processed to generate a word cloud. During this process, media placeholders, system-generated messages, and domain-specific stop words are removed so that only meaningful conversational content is analyzed. A frequency analysis is then performed to identify the most commonly used words, providing insights into recurring topics and key ideas discussed in the conversation.

Emoji analysis is conducted using Unicode recognition techniques to detect, count, and visualize emoji usage within the chat dataset. This analysis helps capture emotional expressions and social cues that are often conveyed through emojis in digital conversations. The analytical framework is designed to remain flexible, allowing both group-level and user-specific analyses across all statistical and visualization components. Furthermore, consistent visualization design—such as color-coded charts, clearly labeled axes, and well-positioned legends—enhances readability and supports accurate interpretation of the analytical results.

Fig 2: Steps involved in data Collection

The analytical framework is built using several open-source Python libraries that support data processing and visualization. The **pandas** library is used for data manipulation and structured dataset handling, while **matplotlib** and **seaborn** are utilized for generating visual representations of communication patterns. The **wordcloud** library is employed for lexical analysis to visualize frequently occurring terms, and **urlextract** is used to detect and

analyze shared links within the chat dataset. The framework is implemented as an interactive web application using **Streamlit**, allowing users to upload WhatsApp chat export files, configure analysis parameters, and explore insights in real time without requiring specialized software or programming expertise. This transparent and systematic approach enhances the reproducibility of the research and provides a flexible foundation for further studies on messaging behavior and conversational data analysis.

	date	user	message	year	month	day	hour	minute	am_pm
0	28/03/2023, 06:11 PM	group_notification	Messages and calls are end-to-end encrypted. O...	2023	March	28	06	11	PM
1	05/04/2023, 10:58 AM	group_notification	+91 77469 89252 joined using this group's invi...	2023	April	5	10	58	AM
2	05/04/2023, 10:56 AM	Siddhartha Sir HOD	<Media omitted>\n	2023	April	5	10	56	AM
3	05/04/2023, 11:41 AM	group_notification	+91 70671 06943 joined using this group's invi...	2023	April	5	11	41	AM
4	05/04/2023, 11:42 AM	group_notification	+91 79749 98145 joined using this group's invi...	2023	April	5	11	42	AM

Fig 2.1: Cleaned Dataset

4. Results and Discussion

The analysis of the WhatsApp chat dataset revealed several notable patterns in group communication behavior. The results indicate that **Friday emerged as the most active day** in terms of message volume, with a significantly higher number of messages compared to other days of the week (Fig. 3). This pattern suggests a time-based clustering of communication activity, which may be influenced by weekly work schedules, academic commitments, or social interaction routines within the group. The corresponding weekly activity heatmap further highlights these patterns, showing concentrated peaks not only on Fridays but also during specific times of the day, such as lunch breaks or late evening hours. These peaks may reflect periods when group members are more available or prefer to engage in conversations, providing insights into the temporal dynamics of group communication.

To prepare the dataset for analysis, the raw WhatsApp chat export was first uploaded into the system and processed to remove system-generated messages, such as notifications indicating that messages are secured with end-to-end encryption, as well as media placeholders. Each remaining message was then annotated with relevant attributes, including the sender identity and timestamp. Using this structured dataset, the total number of messages contributed by each participant was calculated. **Figure 4** presents a ranked bar chart visualization of participants based on their message counts. This visualization highlights variations in conversational engagement, showing that a significant portion of the discussion activity is concentrated among a small number of highly active participants.

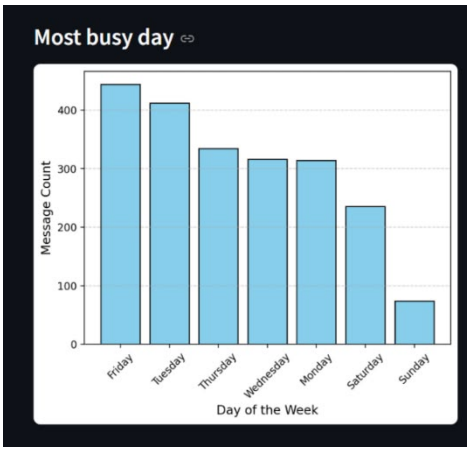

Fig 3: Most busy day

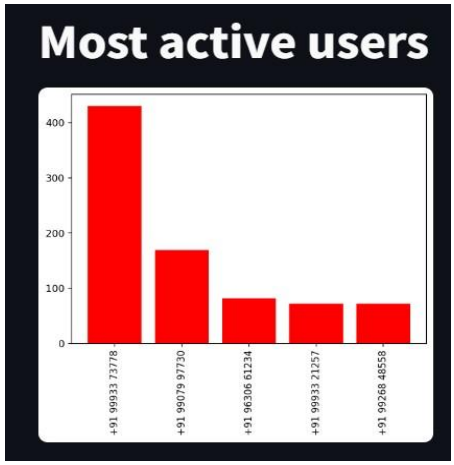

Fig 4: Most active users graph

Quantitative analysis further clarified participation patterns within the conversation. Metrics such as message count, word count, media shares, and link shares were calculated to provide an overview of both overall and user-specific activity (Fig. 5). The results indicate that the majority of messages were text-based, while media files and shared links constituted a smaller but still notable portion of the communication. This distribution reflects the varied nature of interaction within group chats, where different forms of content are exchanged among participants. Additionally, the analysis revealed that a small number of users contributed a significant proportion of the messages, a common pattern observed in group conversations. These findings are supported by both absolute counts and percentage-based contributions, which are illustrated through bar charts and tabular summaries.

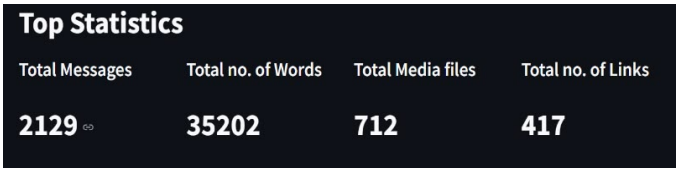

Fig 5: Top Statistics

All relevant textual data from the message dataset was first cleaned and tokenized for analysis. During this process, irrelevant symbols, common stop words, URLs, and numerical values were removed to focus on meaningful vocabulary within the conversation. The remaining text was then analyzed to calculate word frequencies, which

were subsequently visualized using a word cloud, as shown in **Figure 6**. In this visualization, the size of each word corresponds to its frequency of occurrence in the conversation, with larger words representing higher usage. This representation provides a simple yet effective way to highlight the most frequently discussed topics and recurring themes within the group, offering insights into the shared interests and common discussion points among participants.

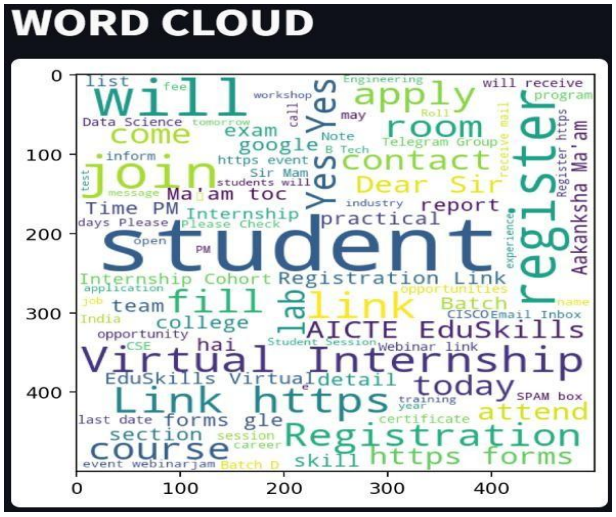

Fig 6: Word Cloud

To further enhance the analytical capabilities of the system, an **AI-based conversation summarization feature** has been integrated into the WhatsApp Chat Analyzer. This feature automatically generates a concise textual overview of the entire chat conversation using Natural Language Processing techniques. After the chat data undergoes preprocessing and cleaning, the processed message corpus is analyzed by an AI summarization module that identifies the most relevant information, key topics, and overall context of the discussion. Instead of manually reviewing hundreds or thousands of messages, users can quickly understand the main purpose and tone of the conversation through the generated summary. As shown in **Fig. 9**, the system provides an interface where users can generate the summary by clicking a dedicated "Generate AI Summary" button.

The AI-generated summary helps highlight the dominant themes, recurring discussions, and the general sentiment present in the chat. In group conversations, where messages often contain mixed information such as announcements, academic discussions, and informal interactions, this feature provides a clear and structured interpretation of the chat content. The summary enables researchers and users to quickly grasp the overall communication context without manually reading the entire dataset. As illustrated in **Fig. 7**, the summarization output presents a short descriptive paragraph capturing important aspects of the conversation such as key topics, communication tone, and informational context, thereby complementing the statistical and visual analysis provided by the system.

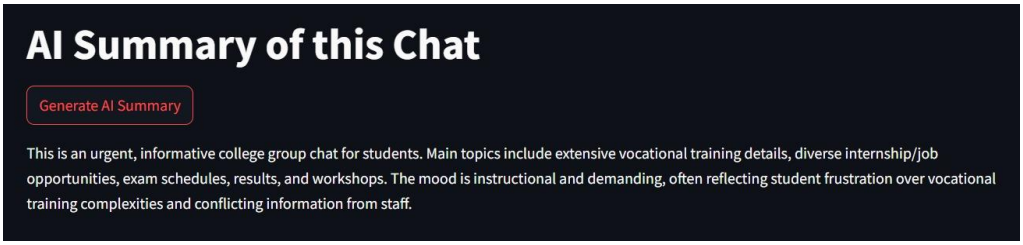

Fig.7: AI-Generated Chat Summary Interface showing automated summarization of WhatsApp conversation

To analyze group engagement patterns, the timestamp information from each message was further decomposed into two components: the day of the week and the hour of the day. By combining these two dimensions, messages were grouped and counted to observe communication intensity across different time intervals. This approach enabled the creation of a heatmap visualization (**Figure 8**), where variations in color intensity represent differences in messaging activity. Darker shades indicate a higher volume of messages, while lighter shades represent lower activity levels. This visualization helps identify the most active periods of communication, highlighting both the days of the week and the specific hours when interaction is most frequent. It also makes it easier to detect patterns such as increased activity during weekends or spikes in conversation during late evening hours.

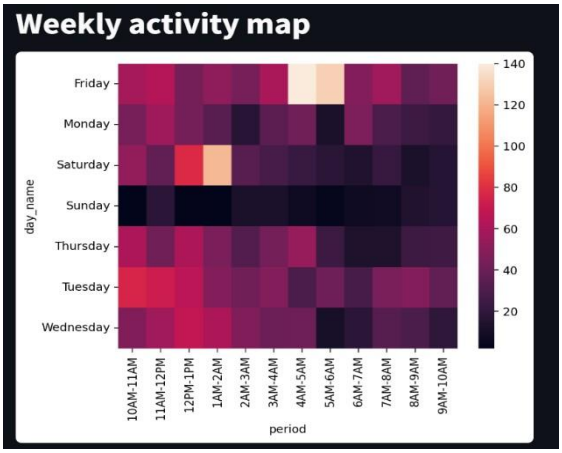

Fig 8: Heatmap

During the preprocessing stage, each message was examined to detect and extract emoji characters present in the chat data. These emojis were then counted to determine the frequency of each unique emoji used throughout the dataset. A pie chart visualization (**Figure 9**) was generated to represent the proportion of each emoji relative to the total number of emojis used in the conversation. This graphical representation highlights the diversity of visual expressions within the chat and provides an indication of the overall emotional tone of the group, whether predominantly positive, humorous, affectionate, or otherwise.

The emoji analysis further emphasized the role of visual cues in digital communication within group interactions. Emojis often convey emotions and reactions that may not be explicitly expressed through text alone, thereby enriching the communication process. Their usage reflects subtle emotional signals and social responses among participants. These observations are consistent with previous research suggesting that emojis enhance the expressiveness of digital conversations and contribute to strengthening social interactions and relationships within online communication environments.

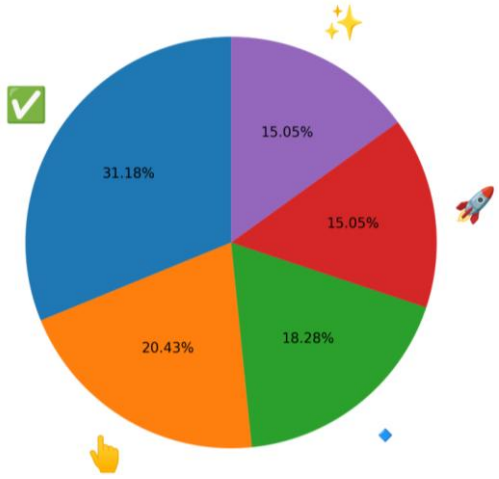

Fig 9 : Pie Chart for most used emojis

In summary, the findings highlight the effectiveness of the proposed analytical framework in transforming unstructured chat data into meaningful and actionable insights. The visualizations generated by the system simplify the interpretation of complex conversational datasets, enabling users to quickly identify patterns and formulate hypotheses regarding group dynamics and communication effectiveness. The modular and adaptable design of the framework allows analyses to be conducted at both the group level and the individual participant level, providing comprehensive perspectives for researchers and practitioners interested in studying engagement patterns, participation behavior, and content trends. These results demonstrate the practical value of automated WhatsApp chat analysis in educational, professional, and community-oriented environments, where understanding patterns of informal digital communication is increasingly important.

5. Conclusion

This research presents a systematic approach for developing an integrated and modular framework for analyzing WhatsApp chat data. The proposed system transforms raw, unstructured text exports into structured and analyzable datasets while providing statistical tools and visualizations to explore communication patterns within group conversations. The findings demonstrate the effectiveness of combining preprocessing, feature extraction, and interactive visualization techniques to identify meaningful patterns such as peak communication periods, dominant participants, and recurring themes within the conversation. Visual tools such as bar charts, timelines, and heatmaps enhance interpretability by enabling users to quickly recognize communication dynamics that would otherwise remain difficult to detect in raw chat logs.

A key strength of the framework lies in its modular and adaptable design, which allows both group-level and individual-level analyses to be performed efficiently. The preprocessing pipeline addresses the inherent inconsistencies present in WhatsApp chat exports, thereby improving the reliability and reproducibility of the analytical results. Furthermore, the use of open-source technologies makes advanced chat analytics accessible to a broader audience, including educators, organizational researchers, management scholars, and digital sociologists.

The study demonstrates that digital conversations can provide valuable insights into social behavior, communication patterns, and interaction dynamics. The framework captures not only temporal participation patterns but also participant engagement levels, thematic discussions, and expressive elements such as emoji usage within conversations. By integrating multiple analytical perspectives, the system effectively bridges the gap between raw conversational data and meaningful insights, supporting both exploratory and targeted analyses within modern digital communication environments.

In summary, the proposed framework provides a practical and scalable workflow for analyzing chat-based communication data while reducing the technical barriers associated with conversational data analysis. The approach opens opportunities for further research into behavioral, linguistic, and organizational aspects of messaging platforms, while also encouraging continued discussion regarding ethical considerations and privacy issues related to chat data analytics.

References

[1] Available from: http://www. statista.com/statistics/260819/numberof-

[2] monthly-active-WhatsApp-users. Number of monthly active

[3] WhatsApp users worldwide from April 2013 to February 2016(in millions).

[4] Ahmed, I., Fiaz, T., “Mobile phone to youngsters: Necessity or

[5] addiction”, African Journal of Business Management Vol.5 (32), pp. 12512-12519, Aijaz, K. (2011).

[6] Radha, D., Jayaparvathy, R., & Yamini, D. (2016). "Analysis on Social Media Addiction using Data Mining

[7] Technique." International Journal of Computer Applications, 139(7), 23-26.

[8] Rizvi, S., & Sherdil, K. (2014). "Content Analysis of WhatsApp Conversation." Karachi Study Report.

[9] Patel, R., & Gupta, S. (2021). "Streamlit: A User-Friendly Framework for Data Analysis Applications." Proceedings of the International Conference on Data Science, 128-135.

[10] Python Software Foundation. (2022). "Python Language Reference, version 3.10." Available: https://www.python.org/doc/versions/.

[11] Kumar, S., & Agarwal, A. (2019). Data Science and Machine Learning with Python. Wiley. This book provides insights into data analysis and machine learning techniques, which were essential for implementing data processing and machine learning algorithms for WhatsApp chat analysis