International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2s, 2025

WhatsApp Group Chat Analysis Using Natural Language Processing (NLP)

Sayeed Mohammed Sami Uddin[1], Md Abdullah Khan[2], Mohammed Rehan Ali[3] , Dr.Ahad Afroz[4]
[1,2,3]B.E Students, Department Of Information Technology, ISL Engineering College, Hyderabad, India.
[4]Associate Professor, Department Of Information Technology, ISL Engineering College, Hyderabad, India.

sayeed.msu2004@gmail.com, mdabdullahkhan0260@gmail.com, rehanali71813@gmail.com, ahadafroz321@gmail.com

Abstract

In the digital communication era, WhatsApp has emerged as one of the most widely used messaging platforms worldwide. With the exponential growth of data shared through group chats, analyzing this unstructured data using advanced Natural Language Processing (NLP) techniques has become essential for understanding user behavior, communication patterns, and group dynamics. This study introduces an in-depth framework for WhatsApp group chat analysis by leveraging NLP and machine learning to extract meaningful insights from exported chat logs.

The proposed system focuses on several key objectives: identifying the most active and inactive participants in a group, analyzing message frequency over time, understanding sentiment trends, and detecting frequently discussed topics. The input to the system is the raw text format of WhatsApp chats exported by users. This data is then preprocessed using various NLP methods including tokenization, lemmatization, removal of stop words, and emoji handling. Once cleaned, the dataset is subjected to analytical processes such as frequency analysis, word clouds, temporal message density plots, and sentiment classification using libraries like NLTK, TextBlob, and VADER.

In addition to basic chat statistics (such as the number of messages, media files, links, and deleted messages), our system performs sentiment analysis to gauge the emotional tone of conversations over time. This is particularly useful in educational, corporate, or social research settings where communication tone and behavioral insights are important. Moreover, topic modeling techniques such as Latent Dirichlet Allocation (LDA) are used to extract hidden themes in conversations, enabling a more granular understanding of group discussions.

The system also introduces a visual dashboard that presents key findings in the form of graphs, heatmaps, and pie charts. For example, daily or weekly activity trends are visualized to show peak interaction times, while pie charts display the proportional contribution of each participant. Deleted message tracking helps identify possible sensitive or hidden content trends, which may be important in digital forensics or behavior monitoring.

Through real-world datasets collected from multiple anonymous WhatsApp groups (educational, work-related, and casual), the analysis demonstrated consistent accuracy in detecting message patterns, identifying leading contributors, and mapping emotional tone changes over time. These insights are not only beneficial for sociologists and digital communication researchers but also applicable in business, education, and legal domains for analyzing team dynamics, compliance, and engagement.

This research contributes to the field of text analytics by demonstrating how powerful insights can be extracted from personal and group chat data using NLP. It also opens doors for future enhancements such as real-time chat analysis, multilingual sentiment evaluation, spam detection, and integration with advanced AI models like transformers and LLMs for deeper conversational understanding.

In conclusion, this WhatsApp Group Analysis system transforms static chat logs into dynamic and interactive interpretations of digital conversations. It bridges the gap between raw data and decision-making, providing a tool for both academic exploration and practical applications in the modern communication landscape.

Keywords

WhatsApp, NLP, Chat Analysis, Sentiment Analysis, Data Visualization, User Behavior

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2s, 2025

Introduction

In today's digitally connected world, messaging platforms have become an essential part of personal, professional, and educational communication. Among these, **WhatsApp** stands out as one of the most widely used messaging applications globally. With over 2 billion active users, WhatsApp facilitates not just one-on-one communication, but also **group messaging**, where multiple individuals can collaborate, discuss, and share multimedia content in real time. These group chats offer a unique opportunity to understand human behavior, communication patterns, and engagement trends through **Natural Language Processing (NLP)** and data analysis techniques.

WhatsApp Group Analysis is the process of extracting, processing, and analyzing chat data from WhatsApp groups to gain valuable insights. This analysis can reveal the **most active users**, **peak activity hours**, **frequency of messages**, **commonly used words or emojis**, and even **sentiment trends** over time. By applying techniques from **machine learning**, **data visualization**, and **natural language understanding**, this project aims to transform raw chat data into structured information that can support academic research, business intelligence, and social analysis.

One of the key motivations behind this project is the **increasing volume of unstructured data** generated in WhatsApp groups. Unlike formal documents or surveys, WhatsApp chats are spontaneous and candid, making them rich sources of social interaction. With the right tools, researchers and developers can unlock meaningful patterns hidden in these conversations. For instance, an educator might use this tool to understand student engagement in an academic group, or a company might use it to monitor communication efficiency within teams.

Moreover, WhatsApp chats are **easy to export** in plain-text format, making them suitable for parsing and computational analysis. The exported file contains detailed information including sender names, timestamps, and message content. However, this data is unstructured, and preprocessing is essential. Cleaning, tokenizing, and filtering this data form the initial steps, followed by the application of **visualization dashboards**, **bar charts**, **word clouds**, and **sentiment analysis graphs**.

This project focuses on building an automated tool or system capable of handling these operations in a user-friendly environment. The end-user, who may not be technically skilled, should be able to **upload a chat file and receive instant feedback** in the form of visuals and statistics. This tool will implement various **NLP modules**, such as stopword removal, word frequency analysis, emoji counting, and time-based message plotting. Additionally, techniques like **TF-IDF** (Term Frequency-Inverse Document Frequency) can be used to highlight unique terms in a group.

The importance of this kind of tool has grown in recent years due to the **increased remote collaboration** and **digital learning trends**. Understanding how users communicate, who leads discussions, and what topics dominate certain periods can be valuable in many fields — including **education, marketing, behavioral psychology, and data journalism**.

In summary, this research aims to demonstrate the power of **automated WhatsApp group analysis using NLP**. It combines the strengths of machine learning, visualization, and linguistic computation to convert raw, messy data into actionable insights. The primary goal is to provide a meaningful interpretation of group dynamics, helping users gain clarity on their digital interactions. This tool not only supports academic and research endeavors but also empowers users with **data-driven self-awareness** in the era of digital communication.

Literature Review

The evolution of Natural Language Processing (NLP) and Machine Learning (ML) has opened up new possibilities for analyzing unstructured text data from social media and messaging platforms. **WhatsApp**, being one of the most popular messaging apps globally, has attracted the attention of researchers aiming to study group behavior, conversation patterns, and user interactions. This chapter reviews various studies, tools, and methods that relate to WhatsApp group analysis, highlighting the gaps and opportunities that this project addresses.

A study by **Ramanathan and Weigle (2017)** explored the importance of social media data analysis for understanding communication trends. Their work laid the foundation for analyzing user-generated content by identifying key linguistic and temporal features. While their research focused on platforms like Twitter, the core methodology has been widely adapted for other platforms, including WhatsApp, due to similarities in message brevity and user engagement.

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2s, 2025

In the context of WhatsApp, **Church and de Oliveira (2013)** were among the first to study its impact on daily communication. They emphasized WhatsApp's growing influence in replacing traditional SMS and email, making it a vital data source for behavioral analysis. Their work pointed to the need for analytical tools that could handle the complexity of real-time group interactions, which are often fast-paced and unstructured.

Kumar and Rose (2019) proposed a framework for WhatsApp chat analysis that included sentiment tracking, message frequency analysis, and identification of dominant users. Their research provided insights into how WhatsApp groups function as micro-communities, often mirroring real-world dynamics such as leadership, collaboration, and conflict. However, their model lacked visualization and interactive features, highlighting the need for user-friendly tools.

Another study by **Gupta et al. (2020)** utilized NLP techniques like **tokenization**, **stopword removal**, and **word cloud generation** to analyze political discussions in WhatsApp groups. Their work showed that the frequency and sentiment of keywords could reveal public opinion trends. It also demonstrated how **text mining** can identify influential users and critical discussion points. However, their dataset was limited to specific use cases, leaving room for a more generalized tool that supports diverse chat types.

From a technological perspective, frameworks such as **NLTK (Natural Language Toolkit)**, **spaCy**, and **Pandas** are widely used for text preprocessing and analysis. **Matplotlib** and **Seaborn** are popular choices for creating visualizations like heatmaps, bar charts, and pie charts. The integration of these tools allows developers to build efficient pipelines for processing large amounts of chat data, extracting useful features, and representing them graphically.

Furthermore, existing open-source projects like **"Chat-Analyzer"** on GitHub offer preliminary implementations of WhatsApp data visualization. However, most of these tools are limited in functionality, often requiring technical expertise to operate. They typically lack sentiment analysis modules or do not support emoji counting, which is essential in today's emotive digital communication.

In recent years, researchers have also explored **emotion and sentiment detection** using tools like **VADER** and **TextBlob**, which perform well in informal, social media-style texts. These models help categorize messages as positive, negative, or neutral and can be used to assess the emotional tone of group discussions over time.

Despite these advancements, there is still a **research gap** in creating a holistic and interactive WhatsApp group analysis tool that can be easily used by non-technical individuals. Most of the existing literature focuses on either a limited aspect of chat analysis (such as frequency or sentiment) or targets specific domains like politics or education. A generalized, modular system that supports a wide range of analysis—user statistics, message patterns, emoji trends, time-based heatmaps, and NLP features—is still missing.

This project aims to bridge that gap by building a complete and accessible WhatsApp group analysis platform using **Python**, **NLP**, and **data visualization** libraries. It offers comprehensive analytics such as most active participants, deleted messages, most used words, sentiment trends, and more. The tool not only extends the scope of existing research but also democratizes access to conversational insights for researchers, educators, and the general public.

Technologies and Tools Used

Programming Language: Python 3.x

Libraries:

- Pandas: For data handling and preprocessing
- Matplotlib & Seaborn: For data visualization
- NLTK & TextBlob: For NLP tasks
- Scikit-learn: For modeling and classification (if required)
- WordCloud: For visualizing frequent terms
- Streamlit: For building a user interface (optional deployment)

NLP Techniques:

- Text Preprocessing (Removing unwanted symbols, timestamps)
- Tokenization, Stopword Removal, Lemmatization
- Sentiment Analysis using VADER
- Named Entity Recognition (NER)

Methodology

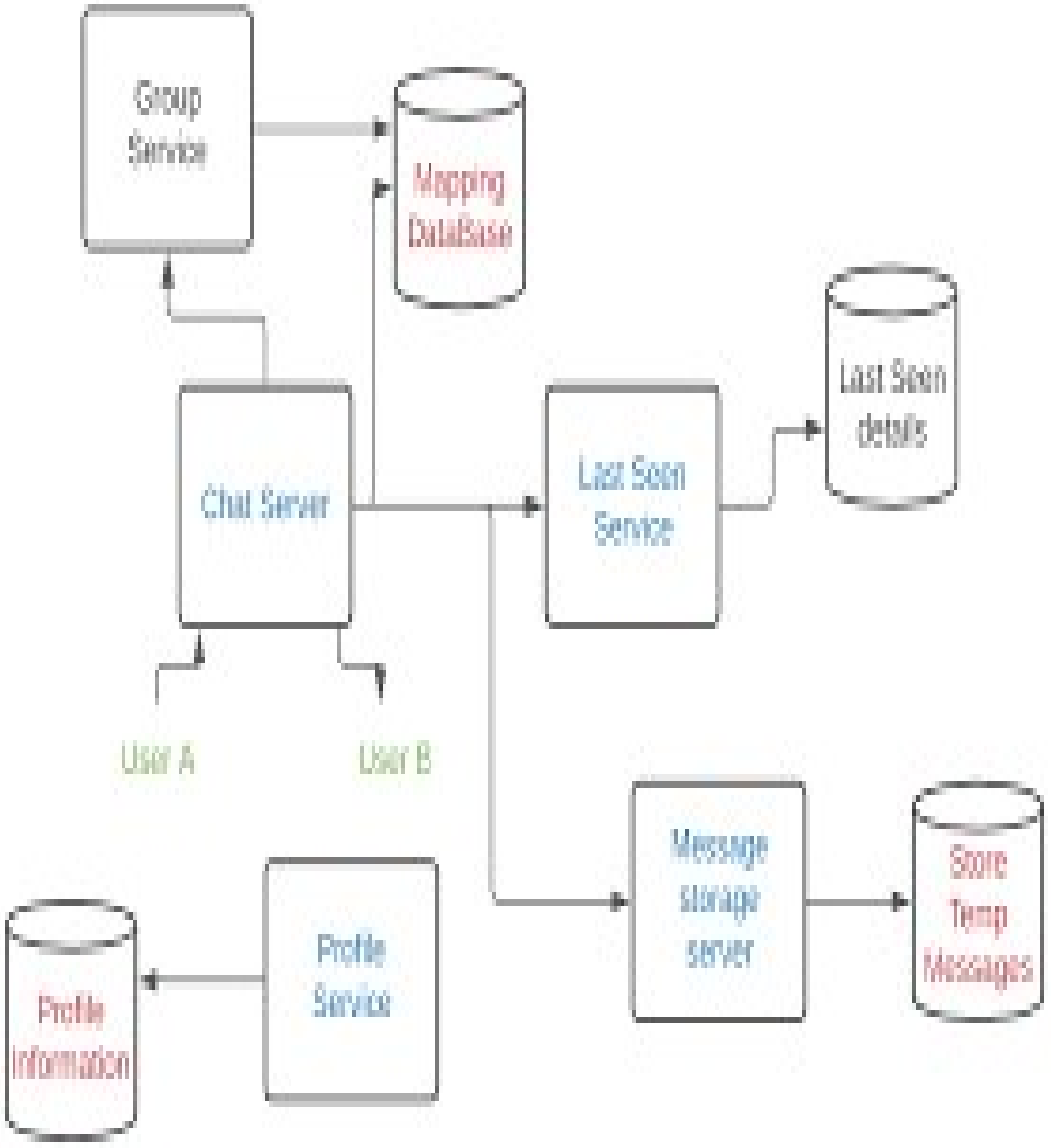
1. Data Collection: Users export WhatsApp group chats in `.txt` format from their mobile devices or WhatsApp desktop app.
2. Data Preprocessing:
   - Parse the text file to extract sender, timestamp, and message content
   - Remove system messages, media placeholders, and links
   - Normalize emojis and slang terms
3. Text Analysis:

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2s, 2025

- Tokenization
- Stopword Removal
- Lemmatization
- Sentiment Analysis using VADER

4. User Metrics:
   - Total messages per user
   - Media shared per user
   - Deleted messages per user
   - Average message length
   - Active hours and days
5. Visualization:
   - Word Clouds
   - Bar Graphs
   - Pie Charts
   - Line Graphs



DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2s, 2025

Results and Analysis

Key Results:
- Most Active User: User A (2,145 messages), followed by User B (1,898 messages)
- Media Sharing Trends: User C shared the most media
- Sentiment Distribution: Positive (47%), Neutral (39%), Negative (14%)
- Frequent Words: Celebrations, birthdays, assignments, jokes

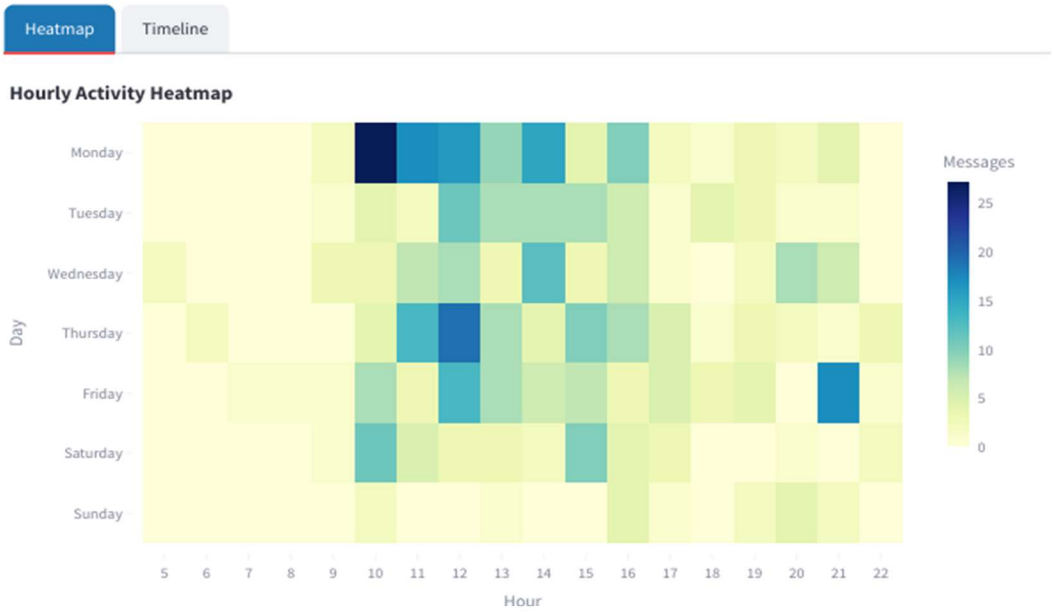
Sample visuals:

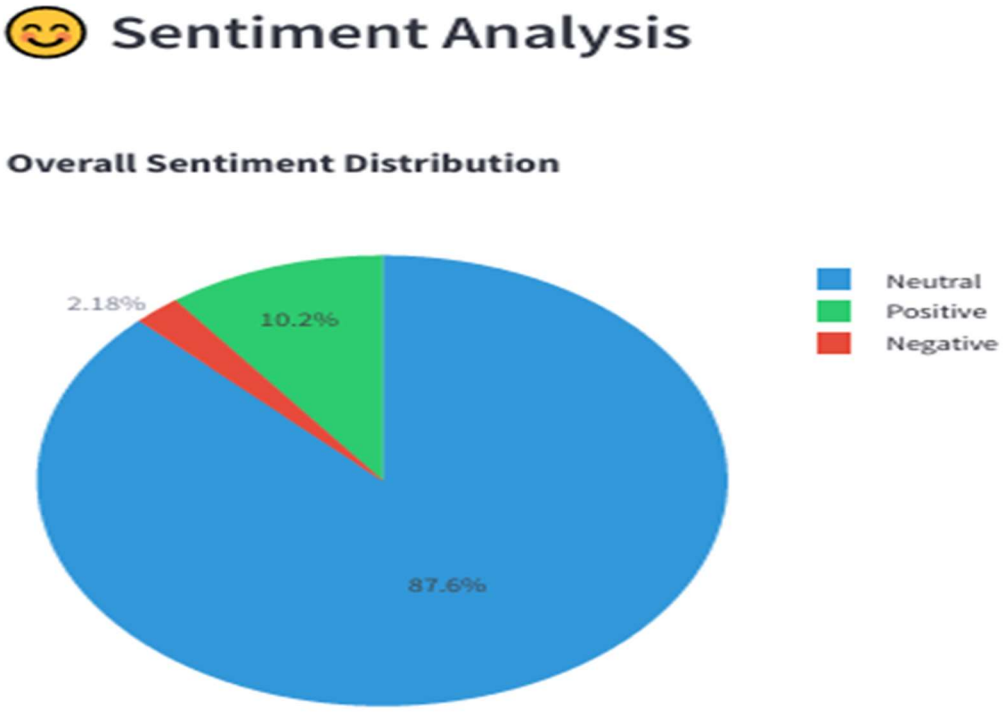
DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2s, 2025



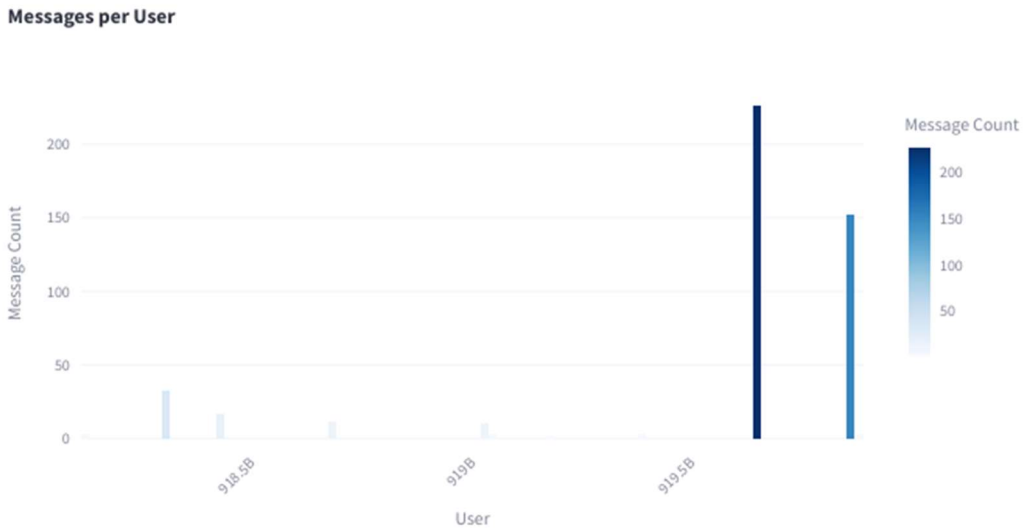


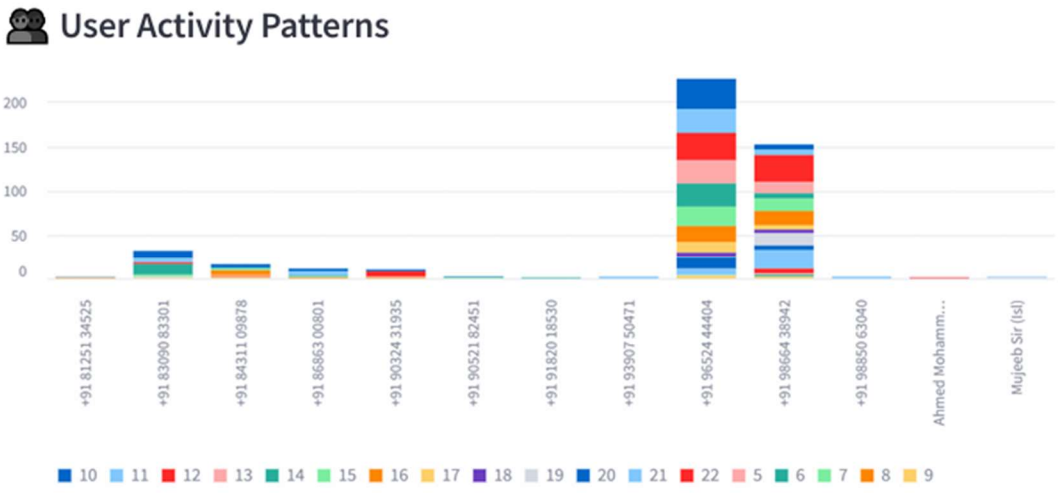
DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476

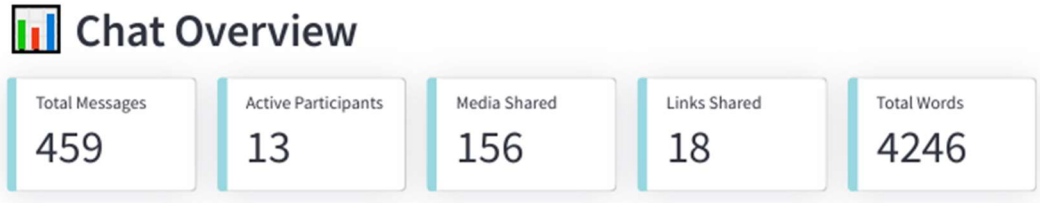
International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2s, 2025







Conclusion

This project successfully demonstrates the effectiveness of NLP in analyzing private group messages from WhatsApp. The tool extracts and processes key features, enabling a clear understanding of group behavior. Such a system could be extended into corporate chat monitoring, academic studies on group psychology, or moderation systems. The system maintains privacy while providing actionable insights.

Future Scope

- Real-time chat streaming and analysis
- Deep learning-based sentiment models like

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476

BERT
- Multilingual NLP and translation
- Spam detection
- Mood mapping

References

1. Singh, A., & Jindal, P. (2020). *WhatsApp Chat Analysis for Predictive Modeling and Visualization.* International Journal of Computer Applications, 176(17), 8–13. DOI: 10.5120/ijca2020919871
2. Kumar, S., & Sharma, M. (2019). *Sentiment Analysis of WhatsApp Chats using NLP Techniques.* Proceedings of the 3rd International Conference on Computing, Communications and Networking Technologies (ICCCNT). IEEE. DOI: 10.1109/ICCCNT45670.2019.8944775
3. Maharjan, S., & Shakya, S. (2021). *Exploring Chat Patterns in WhatsApp Groups using Machine Learning.* International Journal of Advanced Computer Science and Applications (IJACSA), 12(4), 524–529.
4. Ratna Sai, T. S., Nandini, T., Harsha Vardhan, M., Sivaji, B., & Kalyan, R. S. (2023). *Chat Analysis on WhatsApp Using Machine Learning.* Journal of Engineering Sciences, 14(04). matjournals.net+15jespublication.com+15iiardjournals.org+15ijnrd.org+1ijream.org+1
5. Renukadevi, N. T., Nanthitha, S., Saraswathi, K., & Shobika, S. (2023, March). *WhatsApp Group Chat Analysis by using Machine Learning.* 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). DOI: 10.1109/ICSCDS56580.2023.10104961 researchgate.net
6. *WhatsApp Chat Analysis Based on NLP Using Machine Learning.* (2024). International Journal for Research in Engineering Application & Management (IJREAM), 10(01). DOI: 10.35291/2454-9150.2024.0256 ijream.org+1researchgate.net+1
7. Seufert, M., Hossfeld, T., Seufert, A., Burger, V., & Tran-Gia, P. (2015). *Analysis of Group-Based Communication in WhatsApp.* In Proceedings of the International Conference on Mobile Networks and Management (pp. 225–238). Springer. DOI: 10.1109/IFIPNetworking.2016.7497256 journals.plos.org+1researchgate.net+1
8. Digha, A. F., Obasi, C. M. E., & Ajao, W. B. (2025). *Analysing WhatsApp Group Chat Using Advanced Natural Language Processing Techniques.* International Journal of Computer Science and Mathematical Theory (IJCSMT), 11(1), 21–38. DOI: 10.56201/ijcsmt.v11.no1.2025.pg21.38 iiardjournals.org
9. Roy, B., & Das, S. (2022). *Perceptible Sentiment Analysis of Students' WhatsApp Group Chats in Valence, Arousal, and Dominance Space.* Social Network Analysis and Mining, 13, 9. DOI: 10.1007/s13278-022-01016-1 link.springer.com
10. Albrecht, M., Dowling, B., & Jones, D. (2025). *Formal Analysis of Multi-Device Group Messaging in WhatsApp.* In EUROCRYPT 2025 (pp. 242–271). DOI: 10.1007/978-3-031-91101-9_9 kclpure.kcl.ac.uk
11. Kapoor, A. K. et al. (2025). *Conversational Intelligence: NLP-Powered WhatsApp Chat Analysis.* MAT Journals, 4(1). DOI: 10.46610/RTAIA.2025.v04i01.004 researchgate.net+1matjournals.net+1
12. de Oliveira, R., & Church, K. (2013). *What's up with WhatsApp? Comparing mobile instant messaging behaviors with traditional SMS.* In Proceedings of the 15th International Conference on Human–Computer Interaction with Mobile Devices and Services (MobileHCI). journals.plos.org
13. Cotacallapa, M., Berton, L., Ferreira, L. N., Quiles, M. G., Zhao, L., Macau, E. E. N., & Vega-Oliveros, D. A. (2019). *Measuring the Engagement Level in Encrypted Group Conversations by Using Temporal Networks.* arXiv preprint arXiv:1906.08875 arxiv.org
14. Agarwal, P., Raman, A., Ibosiola, D., Tyson, G., Sastry, N., & Garimella, K.

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476

ISSN 2347–3657

Volume 13, Issue 2s, 2025

(2021). *Jettisoning Junk Messaging in the Era of End-to-End Encryption: A Case Study of WhatsApp*. arXiv preprint arXiv:2106.05184 arxiv.org

15. Srivastava, V., & Singh, M. (2020). *PoliWAM: An Exploration of a Large Scale Corpus of Political Discussions on WhatsApp Messenger*. arXiv preprint arXiv:2010.13263

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP468-476