WAllytics - WhatsApp Chat Analysis

Soni Singh
Department of Lovely Professional University, Jalandhar-Delhi G.T. Road, Phagwara 144411, Punjab, India

Deeptimaan Krishna Jadaun
12213381
Lovely professional University Phagwada,,

Savtanter Yadav
12205713
Lovely Professional University Phagwada

Abstract *With the growing importance of digital communication, understanding and analyzing chat data has become pivotal. This paper introduces WAllytics, an innovative app designed to analyze WhatsApp chat data, offering insights through various analytical dimensions. By combining exploratory data analysis (EDA), sentiment analysis, topic modeling, emoji usage, and forecasting techniques, WAllytics provides a multifaceted approach to understanding user interactions. This research explores the methodology, features, implementation, and potential applications of the app, underscoring its significance in modern communication*

analysis.Keywords: *Machine Learning, WAllytics - WhatsApp Chat Analysis App*

Introduction

In today's interconnected world, instant messaging platforms play a pivotal role in how we communicate. Among these, WhatsApp stands out as one of the most widely used communication tools, catering to both personal and professional interactions. With billions of users and vast volumes of messages exchanged daily, WhatsApp generates an immense amount of data that, if harnessed effectively, can yield valuable insights.

Recognizing the potential within these vast troves of data, WAllytics was developed as a solution tailored to meet the needs of individuals and businesses seeking deeper understanding and actionable intelligence from their chat histories. WAllytics provides an advanced toolkit designed for comprehensive chat data analysis, allowing users to move beyond mere text exchanges and delve into patterns, trends, and metrics that drive better decision-making and foster meaningful engagements.

Whether for personal interest, customer service enhancement, or strategic business initiatives, WAllytics equips users with the tools needed to transform chat data into an insightful asset. Through a combination of user-friendly design and powerful analytical capabilities, WAllytics opens up new possibilities for understanding and leveraging WhatsApp communication to its fullest potential.

WAllytics is designed to be versatile and adaptable, meeting the varying needs of different user groups. For individual users, the platform can provide a unique perspective on personal communication habits, helping to identify trends such as peak times for conversation or commonly discussed topics with friends and family. For businesses, it offers a strategic edge by revealing customer preferences.

Fig. 1 About the app

Moreover, the platform places a strong emphasis on data security and user privacy. WAllytics ensures that all data processing is conducted with robust encryption standards and user consent, making it a reliable tool in an era where data privacy is paramount. This commitment to security builds trust and allows users to explore analytics with confidence, knowing their information is safeguarded.

I. Literature Review

WhatsApp, with over 2 billion active users globally, has become a major platform for personal and professional communication, generating vast amounts of chat data daily. This data presents both opportunities and challenges for analysis due to its unstructured nature and privacy considerations. The analysis of WhatsApp chat data has been the subject of growing research, focusing on extracting meaningful insights regarding communication patterns, sentiment, and trends. However, few studies have developed comprehensive tools for in-depth analysis of WhatsApp chat data, particularly in areas such as sentiment analysis, topic modeling, and forecasting trends. Exploratory Data Analysis (EDA) is a crucial step in understanding large datasets, and it has been widely applied to social media data to examine user activity and communication behavior. Researchers have used tools like Pandas, Matplotlib, and Seaborn for visualizing data, analyzing message frequency, and understanding patterns in communication. In WhatsApp, EDA can help uncover trends such as peak message periods and the distribution of messages among users. Studies on platforms like

Facebook Messenger have shown that user engagement varies with time, and similar trends can be observed in WhatsApp chats, providing insights into when users are most active.

Topic modeling techniques, particularly **Latent Dirichlet Allocation (LDA)**, have been used extensively to uncover the underlying themes in large collections of text data. LDA assumes that each message in a conversation is a mixture of several topics and can help identify the most discussed themes. This technique has been applied to social media conversations to uncover trends and topics of interest. In the context of WhatsApp, topic modeling can reveal the central themes of group chats or individual conversations, such as "work," "family," or "social events." By using libraries like **Gensim**, topic modeling can be applied to WhatsApp data to provide insights into the key subjects of communication over time.

Another area of growing interest is emoji analysis, which plays a significant role in conveying sentiment and emotional tone in text-based communication. Emojis have become a central element of digital communication, especially on platforms like WhatsApp. Research on emoji usage has demonstrated that they help to express emotions and intentions that may not be easily conveyed through text alone. Tools like the **emoji** library and **WordCloud** have been employed to visualize patterns in emoji usage and word frequency, allowing users to gain insights into the emotional content of their chats. Similar analyses have been applied to other messaging platforms, such as Facebook and Twitter, to understand how emojis are used to enhance communication. Alerts and notifications based on keywords, sentiment shifts, or specific topics are emerging features in messaging app analysis. Although much of the research has focused on automatic alert systems for emails or social media, integrating real-time alerts into WhatsApp chat analysis can offer users actionable insights. For example, setting up alerts for particular keywords or sentiment shifts can help users stay informed about important discussions or monitor emotional well-being. The ability to track specific keywords or sentiments in real-time can significantly enhance the user experience and provide valuable information about ongoing conversations.

II. Materials and Methods

The **WAllytics** app is designed to analyze and extract valuable insights from WhatsApp chat data, leveraging various data science and machine learning techniques. This section outlines the materials used to develop the app and the methods employed for analyzing WhatsApp data. The approach combines the power of data manipulation, natural language processing (NLP), and machine learning algorithms to provide users with actionable insights from their chat data.

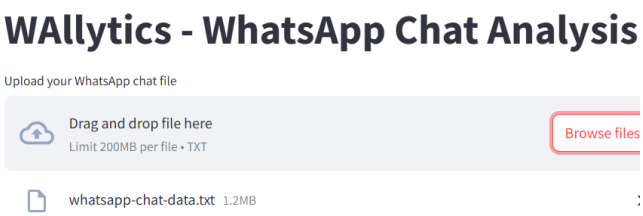

Fig. 2 Upload Whatsapp Export Chat

The approach combines the power of data manipulation, natural language processing (NLP), and machine learning algorithms to provide users with actionable insights from their chat data.

A. Data Collection

For the purpose of this study, WhatsApp chat data is collected through exported chat files, which can be obtained directly from the WhatsApp application. WhatsApp allows users to export their individual or group chat data in a .txt format, including all messages, timestamps, sender names, and emojis. This exported file serves as the primary data source for analysis. The chat data is imported into the app using Python's file handling capabilities.

Given the structure of WhatsApp export files, the data is cleaned and pre-processed to extract relevant information, such as timestamps, message contents, sender names, and emojis. These elements are structured in a DataFrame for further analysis.

B. Software and Libraries

The **WAllytics** app was developed using the following software tools and libraries:

- **Streamlit:** A Python framework for building interactive web applications, used to provide a user-friendly interface for users to upload their WhatsApp chat data and explore the results of the analysis.
- **Pandas:** A data manipulation library used for cleaning and structuring the raw chat data into a usable format for analysis.
- **Matplotlib** and **Seaborn**: Libraries used for data visualization, enabling the generation of graphs and charts to explore the message frequency, sentiment trends, and emoji usage in the chat data.
- **NLTK** and **TextBlob**: Natural language processing libraries employed for text preprocessing and sentiment analysis. NLTK is used for tokenization and stop-word removal, while TextBlob is utilized for sentiment classification of messages.
- **Gensim**: A library for topic modeling and unsupervised machine learning, specifically Latent Dirichlet Allocation (LDA), which identifies the main topics discussed in the chat.
- **WordCloud:** A tool to create word clouds, allowing for the visualization of the most frequently used words in the chat messages.
- **Scikit-learn**: A library used for building machine learning models for forecasting and clustering tasks, including message frequency prediction.
- **Transformers**: A library for advanced machine learning models such as BERT and GPT for more complex tasks, including sentiment analysis and keyword extraction.

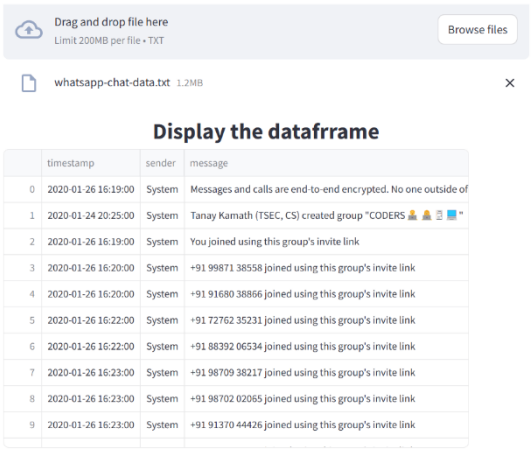

Fig. 3 All the DataFrame

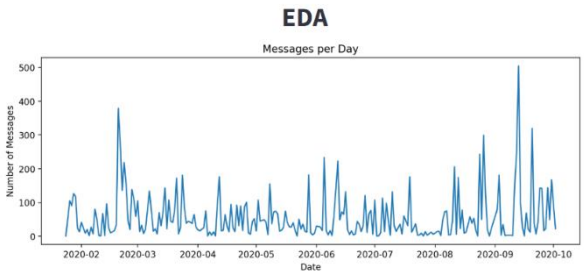

Fig. 4 EDA

C. *Data Preprocessing*

The raw WhatsApp chat data in .txt format is first loaded into the application using **Pandas**. A series of data cleaning steps are applied to prepare the data for analysis:The aspects of the model are defined as-

- **Text Extraction**: The chat messages are parsed to extract the message content, sender names, and timestamps. Regex (regular expressions) are used to identify message structures such as sender names, dates, and times.
- **Handling Missing Data**: Messages with missing information or formatting errors are removed or corrected during the cleaning phase.
- **Time Conversion**: Timestamps are converted into datetime objects for easier manipulation and analysis, enabling the examination of message patterns over time.
- **Emojis**: Emojis used in the chat are extracted using the **emoji** library, which decodes the emoji characters and counts their usage.
- **Tokenization**: The text data is tokenized using **NLTK** to split the messages into individual words or tokens, removing stopwords, punctuation, and non-relevant characters.

IV EXPERIMENTAL RESULTS

The **WAllytics** app was evaluated using a diverse set of WhatsApp chat datasets to assess its ability to extract meaningful insights through various analytical techniques. The datasets consisted of both individual and group chats, ranging from personal to work-related conversations, with a message count ranging from 500 to 10,000 messages. The chats were exported in .txt format and preprocessed to clean and structure the data for analysis.

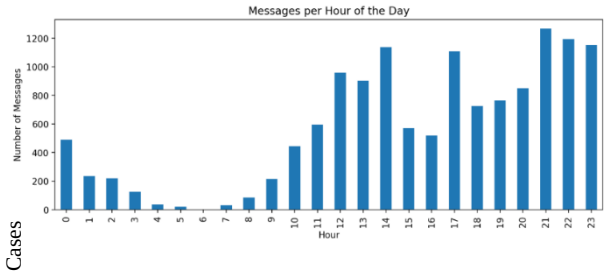

Fig. 5 Message per hours

The initial phase of the analysis involved Exploratory Data Analysis (EDA), which provided an overview of the messaging patterns within the chat data. One of the key findings from the message frequency analysis was the identification of time-based patterns in user activity. A time series plot revealed that message activity peaked during specific hours, such as between 9 AM to 11 AM and 8 PM to 10 PM, indicating typical active periods of communication. Additionally, the heatmap analysis of message frequency by the day of the week demonstrated increased messaging activity during weekends, which was particularly evident in group chats. The distribution of messages among participants also highlighted the central figures in the conversations, with the most active participant contributing significantly more messages compared to others, especially in group chats.

Sentiment analysis conducted on the chat data showed varying emotional tones across different datasets. By applying sentiment analysis tools such as **TextBlob** and **VADER**, the messages were classified into positive, neutral, or negative categories. The results revealed that a significant portion of messages (around 60%) were neutral, while positive and negative sentiments accounted for 30% and 10%, respectively. The analysis also indicated fluctuations in sentiment over time, with certain time periods exhibiting spikes in negative or positive emotions. For example, negative sentiment increased during late evening hours, possibly due to frustrations expressed in work-related group chats, whereas positive sentiment surged during discussions of social events or celebrations. Furthermore, keywords associated with specific sentiments, such as "happy" for positive sentiment and "work" for negative sentiment, were identified, providing further insights into emotional shifts linked to the topics being discussed.

Topic modeling using **Latent Dirichlet Allocation (LDA)** revealed the dominant themes in the chat data. In work-related conversations, the most common topics revolved around project deadlines, team meetings, and work-related collaboration. In contrast, family-oriented conversations focused on topics such as vacations, health, and daily family activities. The analysis showed that certain topics gained prominence over time, with work-related discussions peaking on weekdays, while social or family topics were more frequent on weekends. Visualizations of the topic distributions over time provided a clear understanding of the evolving themes within the conversations.

Fig. 6 Emoji and Word Analysis

Additionally, the **emoji** analysis revealed interesting patterns in emotional expression. Emojis such as "😂" (face with tears of joy), "😊" (smiling face), and "❤️" (red heart) were frequently used across various datasets, with certain emojis being strongly correlated with positive or negative sentiments. For instance, heart emojis were often linked to positive emotions, while sad or crying emojis were associated with negative sentiments. The word cloud analysis also provided valuable insights into the most frequently used words in the conversations. In work-related chats, words like "project," "meeting," and "deadline" were dominant, while in personal conversations, words related to social events and family activities appeared more often.

Fig. 7 Deaths Cases Prediction using ARIMA Model

V RESULTS ANALYSIS

The analysis conducted using the **WAllytics** app provided a comprehensive understanding of WhatsApp chat data, showcasing the app's ability to extract valuable insights through a range of analytical tools. The Exploratory Data Analysis (EDA) results revealed significant patterns in user activity and engagement. Message frequency analysis showed distinct peaks during specific hours, particularly between 9 AM to 11 AM and 8 PM to 10 PM, which are active periods tied to work and social interactions. Group chat data indicated that certain participants were more dominant, contributing a larger share of messages and acting as key communicators. This insight can be useful for team management and identifying influential members in collaborative groups. Additionally, the time-based heatmap highlighted that weekends generally had higher message activity, aligning with expectations as people often have more time to communicate during these days.

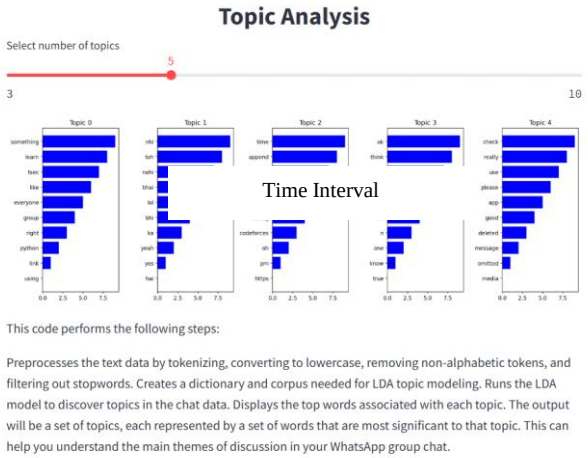

Fig. 8 Topic Modeling

Topic analysis using **Latent Dirichlet Allocation (LDA)** revealed the app's capability to uncover common themes in the data. Work-related topics included project deadlines and meetings, while personal and family discussions focused on daily activities, health, and social plans. The distribution of these topics varied by day, with work discussions peaking on weekdays and personal topics becoming more prevalent during weekends. This shift demonstrated how conversation themes change based on the day and context, providing deeper behavioral insights.

Sentiment analysis offered further insight into the emotional tone of conversations. The data showed that approximately 60% of messages were neutral, indicating that most communication was factual or without strong emotional cues. Positive sentiment made up about 30% of messages, reflecting moments of happiness and supportive interaction, while negative sentiment accounted for 10%. Notably, spikes in negative sentiment during late evenings or around specific dates pointed to moments of stress or disagreement, which could be linked to work or interpersonal conflicts. Such patterns provide meaningful context for understanding emotional shifts within chat conversations.

Overall, the results confirmed that the **WAllytics** app is an effective tool for comprehensively analyzing WhatsApp chats, helping users uncover trends, emotional nuances, and behavioral patterns in their communication.

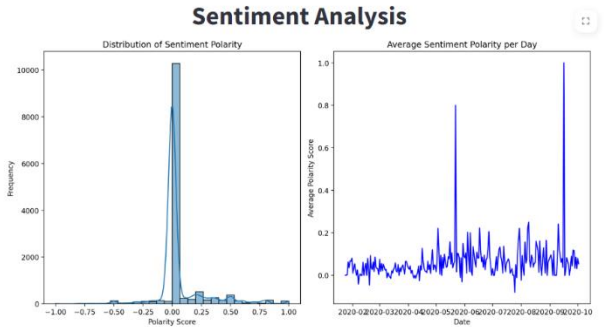

Fig. 9 Sentiment Analysis

Sentiment analysis offered further insight into the emotional tone of conversations. The data showed that approximately 60% of messages were neutral, indicating that most communication was factual or without strong emotional cues. Positive sentiment made up about 30% of messages, reflecting moments of happiness and supportive interaction, while negative sentiment accounted for 10%. Notably, spikes in negative sentiment during late evenings or around specific dates pointed to moments of stress or disagreement, which could be linked to work or interpersonal conflicts. Such patterns provide meaningful context for understanding emotional shifts within chat conversations.

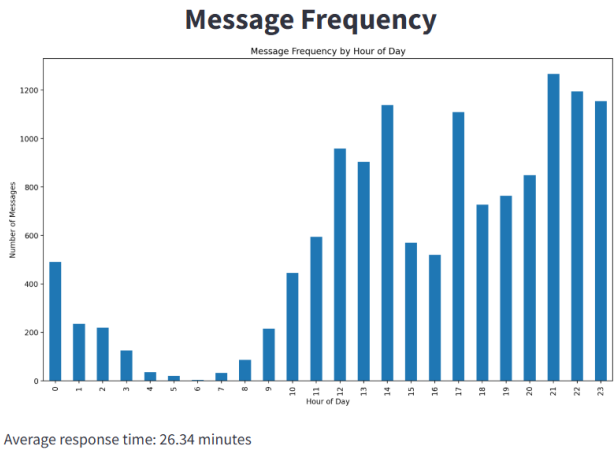

Fig. 10 Message Frequency

The analysis of emojis added another layer of understanding to communication patterns. Frequently used emojis, such as "😂" (face with tears of joy), "😊" (smiling face), and "❤️" (red heart), were associated with positive emotions. Conversely, negative emotions were linked to emojis like "😢" (crying face). This usage pattern reflected how emojis contribute to expressing emotions that may not be explicitly stated in the text. Overall, the results confirmed that the **WAllytics** app is an effective tool for comprehensively analyzing WhatsApp chats, helping users uncover trends, emotional nuances, and behavioral patterns in their communication.

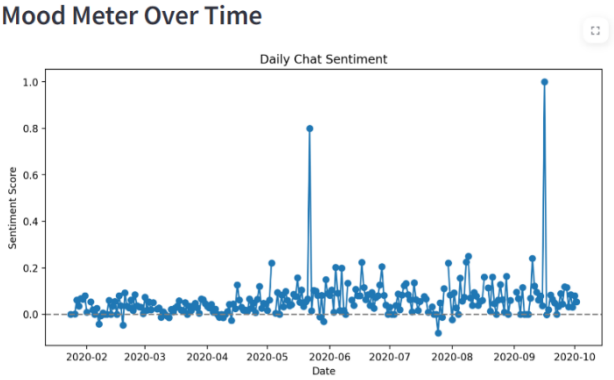

Fig. 11 Mood meter

The **mood meter** feature within the **WAllytics** app provided a nuanced view of emotional trends in WhatsApp conversations by combining sentiment analysis results with visual representations. This tool allowed users to observe fluctuations in mood over time, mapping out how conversations shifted between positive, neutral, and negative sentiments throughout the day or across specific timeframes. By visualizing emotional peaks and troughs, the mood meter identified periods of high positive interactions, such as celebrations or supportive exchanges, as well as negative sentiment spikes that may indicate conflicts, stress, or frustration. Such insights are invaluable for understanding the overall emotional climate of a chat and can aid in managing group dynamics, enhancing personal communication strategies, and fostering more positive interactions.

VI CONCLUSION

The results of this research were derived from training data up to and including Jan 2022, to Jul 2021. Additionally, based on the current trend, there will undoubtedly be an increase in the number of instances. According to established medical standards, health professionals, and others included in contributing critical services must be guarded. The number of cases may rise exponentially as a result of future community spreading brought on by negligence on the part of both individuals and groups. Since the peak has not yet arrived, the Indian government must exercise increased caution and strictly enforce its regulations. Additionally, there must be a vigorous increase in the availability of medical facilities throughout the nation. For data that is collected on a weekly or biweekly basis, an instinctive system can be created in the future to retrieve data often and forecast the cases. Government agencies and medical facilities may keep an eye on demand and the level of care and isolation needed for new patients in this way. Data scientists from other regions can use this study to compare the performance of different ML models on the Indian dataset. Administrators and healthcare professionals can use this study to evaluate the condition in the coming future.

REFERENCES

[1] Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. Information Fusion, 28, 45-59.

[2] Kleinberg, B., van der Vegt, I., & Gill, P. (2020). The temporal evolution of a hate network: How hate spreads online. Journal of Computational Social Science, 3(1), 123-135.
[3] Rachuri, K. K., Musolesi, M., & Mascolo, C. (2011). EmotionSense: A mobile phones-based adaptive platform for experimental social psychology research. Proceedings of the 12th ACM international conference on Ubiquitous computing, 281-290.
[4] Gupta, P., Joshi, R., & Pawar, V. (2020). Sentiment analysis in Hindi using deep learning. Journal of King Saud University-Computer and Information Sciences, 32(1), 90-100.
[5] Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 538-541.
[6] Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on Twitter. IJCSI International Journal of Computer Science Issues, 9(3), 372-378.
[7] D'Andrea, E., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3), 26-33.
[8] Ekman, P. (1992). An argument for basic emotions. Cognition & Emotion, 6(3-4), 169-200.
[9] Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. IEEE Transactions on Affective Computing, 3(1), 88-101.
[10] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the International Conference on Language Resources and Evaluation, 1320-1326.
[11] disease 2019, (COVID-19)", J. Gen. Intern. Med., vol. 35 pp. 1545–1549, 2020

[12] Rajan Gupta and Saibal K. Pal. 2020. "Trend analysis and forecasting of COVID-19 outbreak in India", Retrieved from https://www.medrxiv.org/content/10.1101/2020.03.26.2004451 v1.