iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

TIME-SERIES DATA DRIVEN CONVERSATIONAL CHAT ANALYTICS

[1]Pradyumna Maydeo, [2]Avishkar Rode, [3]Kunal Jadhav and [4]Parvati Bhadre
Department of Information Technology,
D. Y. Patil College of Engineering, Akurdi

ABSTRACT
As of 2024, WhatsApp has over 2 billion monthly active users, making it one of the most widely used messaging platforms globally. In addition, more than 100 billion messages are sent daily on the app, highlighting its integral role in daily communication (K. Kaushik and Y. Katara, 2022 [4]) This project aims on innovating an approach for analyzing WhatsApp chat data, providing users with a powerful tool to upload their chat histories in text format and receive detailed analytical reports through interactive visualizations. The application serves as an engine for comprehensive data analysis, examining critical aspects of WhatsApp conversations. Users can access insights related to the total number of messages exchanged, the overall word count, and the number of media files and links shared within the chats. Moreover, the tool features timeline analyses that display chat activity on a daily, weekly, and monthly basis, allowing contributors to understand patterns in their communications over time. It identifies the busiest days and months, helping users pinpoint when their conversations are most active, as well as highlighting the most engaged users within their groups. Additional features include an exploration of common and top words used in conversations and detailed emoji analysis, which offers a fun yet insightful view of emotional expressions in chats. The project is to be built using Python for robust data analysis and MERN stack or creating a user-friendly, interactive web application. This platform is particularly advantageous for institutions and businesses that leverage WhatsApp as a communication channel, as it offers a clear and concise view of customer interactions, marketing efforts, and group dynamics. By providing valuable insights into chat patterns and user engagement, the project empowers users to make informed, data-driven decisions, for improving communication effectiveness with scope for enhancing their overall business strategies. Ultimately, this project not only addresses the need for analytical tools in WhatsApp but also opens up avenues for optimizing user interactions and driving better outcomes in a fast-paced digital environment across multiple conversational avenues.

Keywords:
Time - series data, Conversational Analysis, Text Analysis.

INTRODUCTION

In today's digital age, social media and messaging platforms such as Facebook Messenger, WhatsApp, and iMessages have become primary tools for communication, enabling individuals to share their evaluations, opinions, questions, and critiques in both public and private settings. Naturally, it's a valuable source of data that is not only rich in linguistic form, but also provides a plethora of time-related metadata. This includes 'time sent,' 'read receipts,' and 'last seen,' which impact communication quality and correlate to healthy use of said technology (N. Blabst and S. Diefenbach, 2017 [1]). Contrary to the increasing reliance and growth in text-based communication, the area of time-driven statistical analysis of such corpora remains largely unexplored. Through analytical techniques, one can predict the user's sentiment or mood or understand customer sentiment towards certain products or services. At the periphery, such an analysis can provide an understanding of social behavior and engagement patterns for professional online chats (P. Anand, R. Gupta, et al., 2022 [2]). Previous studies have highlighted the importance of robust text data pre-processing, especially for time-series modeling, which can improve analysis accuracy in such datasets (J. Pomenkova, P. Koráb, and D. Štrba, 2023 [5]). Additionally, sentiment analysis approaches have been enhanced for text containing mixed-case language, which brings out the linguistic nuances present in digital communication (P. Awatramani et al., 2021 [8]).
Conversation Analysis (CA) is a methodological approach used to study social interactions, focusing on the structure and organization of conversations. The proposed research seeks to explore how patterns within these digital interactions can be linked to broader organizational issues and communication processes. CA examines

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

how individuals interact and respond to one another in a conversation, particularly how they structure their turns and manage exchanges during interactions. Rather than solely focusing on linguistic structure, CA is concerned with how language functions in everyday social exchanges are by-products of the time related metadata that the applications provide. By analyzing conversations on text-based platforms, CA provides valuable insights into how people perceive and react to one another, offering a deeper understanding of the dynamics of virtual communication. This method allows researchers to examine a wide range of communication styles and behaviors within different contexts, providing insights into both personal and professional insights. Such a tool can be of utility in various genres that have a role in explaining why people act the way they do during a conversation.

METHODOLOGY

The methodology for the development of CTDA involves a structured approach encompassing stages such as data collection, preprocessing, feature extraction, machine learning model selection, and the design of an interactive web interface. At this stage, the project has been carefully planned, and the development is to be done in phases.

Data Collection

The data for the CTDA project will be sourced from WhatsApp chat exports. Users will upload their WhatsApp chat history in the form of a plain text file, which contains the conversation history, including message content, timestamps, and user identifiers.

1. **WhatsApp Chat Export**: Users will download their chat history via WhatsApp's export feature and upload the raw text file for analysis. This data includes messages, timestamps, and participant information, which will be the foundation for further processing.
2. **Data Preprocessing**: Before any analysis is performed, the chat data will undergo several preprocessing steps to ensure it is structured and standardized. This includes:
    a. **Timestamp Parsing**: Associating messages with the correct timestamps to analyze communication patterns.
    b. **User Identification**: Identifying users to analyze individual communication behaviors.
    c. **Cleaning the Data**: Removing irrelevant system messages and formatting issues to ensure consistency.

Feature Extraction

Following preprocessing, meaningful features will be extracted from the chat data to provide insights into the conversations. These features will support various types of analysis, including message content analysis and temporal patterns.

1. **Message Analysis**: Each message will be analyzed for various features that reveal trends and behaviors within the conversation:
    a. **Word Count**: Calculating the number of words in each message.
    b. **Media and Link Count**: Counting the number of media files and links shared.
    c. **Sentiment and Emoji Analysis**: Classifying each message as positive, negative, or neutral based on sentiment, and counting emojis to assess emotional expression.
2. **Temporal Analysis**: Time-based features will provide insight into communication trends:
    a. **Daily, Weekly, and Monthly Activity**: Tracking the volume of messages sent over different time periods.
    b. **Peak Activity Periods**: Identifying times when conversations peak, such as certain days or weeks.
    c. **Response Time**: Analyzing the average time taken for users to reply to each other's messages.

Machine Learning Models

Machine learning models will be integrated to extract deeper insights from the processed data, such as detecting sentiment, evaluating user engagement, and identifying communication trends.

1. **Sentiment Analysis**: Sentiment analysis will be applied to categorize each message based on its emotional tone (positive, negative, neutral). This will help in understanding the mood and attitude of participants during conversations.
2. **User Engagement Analysis**: Key metrics to be analyzed include:
    a. **Message Frequency**: The number of messages sent by each user, reflecting their participation.
    b. **Conversation Participation**: The level of engagement, including how often users initiate messages and their response behavior.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

3. **Time-Series Analysis:** Advanced pre-processing techniques for time-series modeling could be incorporated to enhance the system's ability to detect long-term trends and patterns in communication behavior (J. Pomenkova, P. Koráb, and D. Štrba, 2023 [5]).
4. These models will be refined over time as data is collected and analyzed.

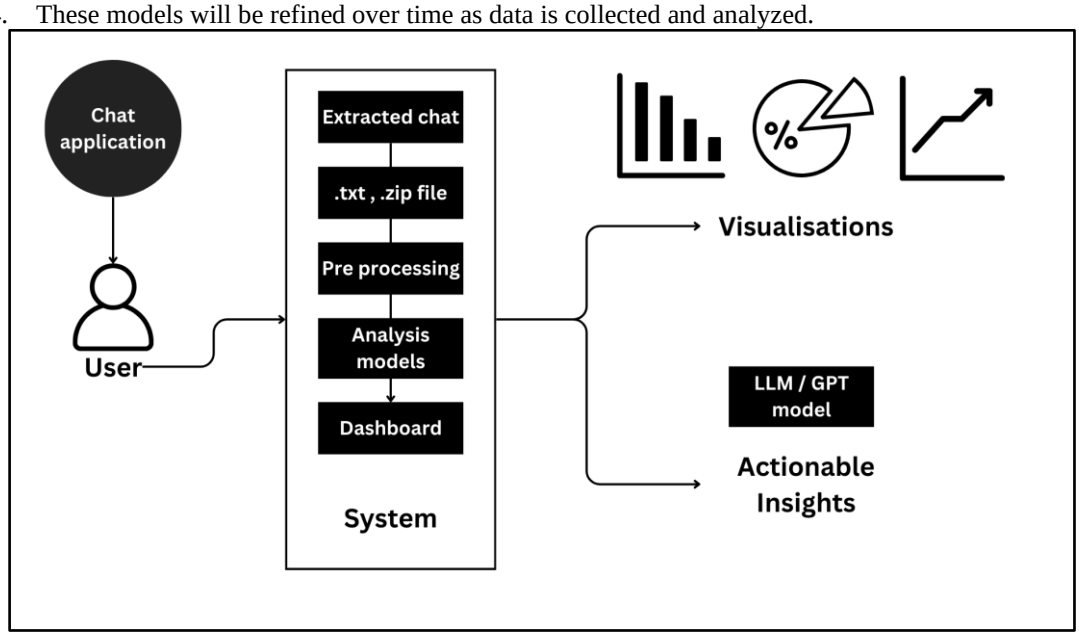

Figure 1. Proposed System

Development Stack and Implementation

The CTDA web application will be developed using the MERN stack (MongoDB, Express, React, Node.js), chosen for its robustness, scalability, and ease of integration.

1. **Frontend (ReactJS)**: The user interface will be designed using ReactJS to display interactive charts and graphs for visualizing the results of the chat analysis. This will help users easily interpret the insights generated.
2. **Backend (Node.js and Express)**: The backend will utilize Node.js with Express to handle API requests, manage data processing, and facilitate communication between the frontend and the database.
3. **Database (MongoDB)**: MongoDB, a NoSQL database, will be used to store both processed chat data and analytical results. Its flexible structure will allow for easy scalability as the project evolves.

RESULTS AND DISCUSSION

1. For the target of performing temporal and statistical chat based analysis, Whatsapp chats prove to be the ideal source for input corpus as it contains rich, unstructured data, including text, emojis, multimedia (such as images and videos), timestamps, and more. This offers a complex, multi-dimensional data source for analysis. It is also the most trouble-free to acquire with user's consent.
2. WhatsApp provides both personal (one-on-one) and group chat interactions, making it suitable for analyzing different dynamics in communication, such as social behavior in groups versus individual exchanges.
3. Use of Traditional Machine Learning approaches is best suited for text classification and statistical analysis, whereas Deep Learning approaches facilitate linguistic analysis.
4. The use of visualizations like word clouds, sentiment analysis charts, and message volume graphs allows for a clearer understanding of communication patterns, sentiment distribution, and user engagement within the WhatsApp group.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

ACKNOWLEDGEMENT

1. **Mrs. Parvati Bhadre, Associate Professor, Department of Information Technology, D. Y. Patil College of Engineering** - For her expert guidance, valuable feedback, and constant support throughout the course of this project.
2. **D. Y. Patil College of Engineering, Akurdi**- For providing the infrastructure and resources that facilitated the research, including the hardware, software, and internet facilities.
3. **Fellow Students and Research Peers** - For their collaboration, discussions, and feedback that helped refine various aspects of this work.
4. **WhatsApp Users** - For allowing their conversation data to be analyzed, which forms the basis of this research.

CONCLUSION

Effective feature extraction is one of the important steps to construct an effective conversation analysis. The corpus of input data and its source along with desired outcomes, determine the approach to be used (Traditional model learning, Deep learning and ensemble approach). For comparison, related works in the field have used lexical and contextual features for CA as they are proven to be effective (R. Sujatha and K. Nimala, 2022 [7]). The proposed system deviates in a direction of temporal analysis of conversations, and aims to build on top of proven models. The Conversational Time-Coded Data Analyzer (CTDA) is designed to provide institutions with a valuable tool for analyzing communication dynamics within professional settings. With the initial setup of the web application completed and the development of core components of data acquisition and data pipeline in place, CTDA shows significant promise in delivering temporal insights from WhatsApp chat data. The project aims to extract meaningful patterns from time-coded data, allowing institutions to gain a detailed understanding of communication flows, team interactions, and prevalent discussion topics, by leveraging advanced data processing and analytical techniques.

REFERENCES

[1] N. Blabst and S. Diefenbach, "WhatsApp and Wellbeing: A study on WhatsApp usage, communication quality and stress," Ludwig-Maximilians-University Munich, (**2017)**.
[2] P. Anand, R. Gupta, et al, "Determining the Polarity and Statistics of Chat based on Sentiment Analysis," Galgotias University, **(2022)**.
[3] R. K. Ravishankara, Dhanush, Vaisakh, and S. I. Srajan, "WhatsApp Chat Analyzer," Srinivas Institute of Technology Mangalore, **(2020)**.
[4] K. Kaushik and Y. Katara, "Forensic Analysis of WhatsApp chat data," University of Petroleum and Energy Studies, Dehradun, India, **(2022)**.
[5] J. Pomenkova, P. Koráb, and D. Štrba, "Text Data Pre-Processing for Time-series Modeling," Brno University of Technology, Brno, and Zeppelin University, Friedrichshafen, Lentiamo Prague, **(2023)**.
[6] P. Debnath, S. Haque, S. Bandyopadhyay, and S. Roy, "Post-disaster Situational Analysis from WhatsApp Group Chats of Emergency Response Providers," IIEST, Shibpur, Indian Institute of Management Calcutta and Heritage Institute of Technology, Kolkata, **(2016)**.
[7] R. Sujatha and K. Nimala, "Text-based Conversation Analysis Techniques on Social Media using Statistical Methods" SRM Institute of Science and Technology, Kattankulathur **(2022)**.
[8] P. Awatramani, R. Daware, H. Chouhan, A. Vaswani, and S. Khedkar, "Sentiment Analysis of Mixed-Case Language using Natural Language Processing" Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, Mumbai **(2021)**.