WhatsApp Chat Analysis Using Machine Learning

Sakshi Pituk[1], Janvi Shinde[2], Vanshita Jadhav[3] , Vaishnavi Bidoo[4], Balasaheb Balkhande[5], Parmeshwar Manegopale[6]

[1,2,3,4]CO student, Dept. of Computer Engineering from Vasantdada Patil Pratishthans College of Engineering Mumbai, Maharashtra, India.

[5.6]CO professor, Dept. of Computer Engineering from Vasantdada Patil Pratishthans College of Engineering Mumbai, Maharashtra, India.

---***---

Abstract - With each passing year, the number of data in WhatsApp is increasing. Visual representation of data gives a piece of effective information rather than manually. The "WhatsApp Chat Analysis using Machine Learning" project is a multifaceted exploration focused on deriving valuable insights from WhatsApp conversations. Through a methodical approach encompassing data cleaning, pre-processing, and statistical analysis, this study endeavors to offer a comprehensive understanding of communication patterns within the WhatsApp platform.

Employing advanced data processing techniques and compelling visualizations, our project seeks to illuminate the dynamics of user interactions. It delves into sentiment trends, identifies emerging topics, and uncovers peak activity periods within the digital messaging landscape. By leveraging machine learning algorithms, the analysis is designed to extract meaningful information, providing a nuanced perspective on how individuals engage and communicate in the digital realm. This project aims to contribute to a deeper understanding of the intricacies of WhatsApp conversations and offers a foundation for future studies in the evolving landscape of digital communication.

***Key Words*:** WhatsApp Chat, Analysis, Sentiment analysis, Spam detection, Translation.

1. INTRODUCTION

This project gives us a deep analysis of the data of WhatsApp Chats. The increasing significance of analyzing chat data in the digital age stems from the transformative impact of digital communication platforms on the way individuals, communities, and societies interact. As technology continues to evolve, the volume of data generated through online conversations has surged, presenting researchers with a valuable resource to gain insights into social behaviors.

The datasets have been read into pandas data frames and have been combined, data cleaning has been done on the combined dataset. The initial phase of implementing a machine learning algorithm involves defining the optimal learning experience from which the model initiates its improvement process. Data pre-processing is crucial for efficient machine learning. To optimize the model, ample high-quality data is essential. This project seeks to address this research gap by concentrating specifically on WhatsApp as a communication platform, showcase the transformative potential of machine learning in unraveling the complexities of digital communication.

2. LITERATURE REVIEW

WhatsApp has revolutionized modern communication, reshaping how individuals connect and share information in the digital era. Advanced machine learning techniques enhance this experience, allowing us to extract insights from raw WhatsApp chat data. By delving into communication patterns and trends, we create a dynamic portrayal of interactions. Notably, statistics reveal that Android users spend an average of 33.5 minutes daily on WhatsApp, with India boasting the highest number of monthly active users at 487.5 million.

The project, led by Dr. T. Siva Ratna Sai and his team, centers on leveraging Natural Language Processing (NLP) techniques for chat analysis on WhatsApp. Intending to enhance the accuracy of machine learning models, the project aims to gather essential data and conduct a comprehensive exploration of various WhatsApp chat types. A key focus lies in utilizing a variety of Python modules to ensure efficient code representation and facilitate user understanding. Through this endeavor, the team endeavors to unlock valuable insights from WhatsApp conversations and contribute to advancements in NLP-based chat analysis methodologies.[1].

In the study conducted by C.B.S. Reddy, S. Kowshik, K.M. Rakesh, O.N.K. Reddy, and G.

Gopichand, the focus is on analyzing and predicting emotions within WhatsApp chats using sentiment analysis. This involves the application of natural language processing techniques to decipher the emotional tone conveyed in chat messages. The primary objective of sentiment analysis is to identify and categorize emotions expressed in the chats, ranging from happiness and sadness to anger or neutrality. Through this research, the team aims to gain insights into the emotional dynamics of WhatsApp conversations and develop predictive models to anticipate emotional responses within the chat environment[2]. In "Detecting Spam in WhatsApp Group Chat Using Machine Learning Techniques," research focuses on leveraging machine learning techniques to address the prevalent issue of spam messages within WhatsApp group chats. Through the application of machine learning algorithms, the authors propose a methodology for effectively identifying and filtering out spam messages from group conversations. This study contributes to the field of spam detection by offering a practical approach tailored specifically to the WhatsApp platform.[3]

Sentiment analysis involves automatically identifying the subjectivity (whether a text is objective or subjective), polarity (positive or negative sentiment), and intensity (strength of sentiment) of a given text. [4]This area of research is rapidly expanding, particularly due to the valuable insights that can be derived from analyzing opinions and sentiments expressed online. Approaches to sentiment analysis have tackled the problem from two different angles: word-based or semantic approach, or a machine learning (ML) approach. The word-based approach involves utilizing dictionaries containing words tagged with their semantic orientation (SO) to determine sentiment. Sentiment is evaluated by summing the values associated with the words present in a given text or sentence.. A study conducted on students in Abu Dhabi, reveals that 85% of female students and 70% of male students use emojis to replace facial expressions on text. These students used WhatsApp every day for an average of 1 to 7 hours. Hence, emojis an important for communication over WhatsApp [5]. In another study where 3.88 million users from various countries were studied, it was revealed that 7.01% of their 6.06 billion messages contained at least one emoji. This study reveals how popular usage of emojis is among users [6].

3. SCOPE OF STUDY

Here, in this project, through machine learning techniques, users can upload chat files, preprocess data, and generate detailed analyses. The system will offer features such as exploratory data analysis, spam detection, and translation of the chats. Visualizations will enhance insights presentation, covering message frequency, user activity trends, and media sharing patterns. Compatibility with diverse platforms and scalability for large datasets will be ensured, providing users with a powerful tool for efficient and effective analysis of their WhatsApp chats.

4. PROPOSED FRAMEWORK

The "WhatsApp Chat Analysis" system provides users with a powerful platform to delve deep into their WhatsApp conversations. With an intuitive interface, users can seamlessly upload exported WhatsApp chat files in (.txt) format, initiating the analysis process with a simple click of the "Show Analysis" button.

Upon inputting the chat files, the system undergoes an initial phase of exploratory data analysis. The insights are then visually represented through a variety of charts and graphs, including pie charts, bar graphs, and timelines, offering users a comprehensive overview of the sentiment distribution within their chats.

Moreover, the system generates informative line graphs, showcasing metrics such as user and message counts for each date and participant. An ordered graph correlating dates with message counts offers insights into conversation trends over time. Additionally, the system highlights messages with unattributed authors, ensuring transparency in the analysis process. To further enhance the quality of analysis, the system integrates sophisticated spam detection mechanisms. These mechanisms work diligently to filter out irrelevant messages, ensuring that users receive accurate and actionable insights from their conversations.

One of the standout features of the system is its translation functionality. This feature enables users to translate messages into various languages, facilitating comprehensive multilingual analysis. By breaking language barriers, users can gain deeper insights into the content and sentiment of their conversations, regardless of the language used.

Overall, the "WhatsApp Chat Analysis" system offers users a comprehensive and visually engaging exploration of their WhatsApp conversations. With its powerful analytical capabilities and user-friendly interface, it empowers users to uncover valuable patterns and trends hidden within their chat data.

5. IMPLEMENTATION

This project consists of five primary modules, each comprising sub-modules for a comprehensive workflow:

1. Data Extraction: This module initiates data extraction and serves as the starting point for gathering WhatsApp chat data. It facilitates the extraction process by leveraging WhatsApp's export feature, allowing users to easily obtain their chat history. Once users initiate the export process through the designated button within the WhatsApp application, the chat data is saved into a text file format, preserving the raw information as it appears in the chat interface.

2. Data Collection: The Data Collection process allows users to obtain their WhatsApp conversation data effortlessly. Through the utilization of WhatsApp's "Export Chat" feature, users can seamlessly send their entire conversation to their designated email address in text format. The data will be collected by clicking on “Show Analysis”. This functionality streamlines the process of gathering chat data, enabling users to access and preserve their conversations for further analysis or reference. The preservation of chat conversations, empowering users to analyze or refer back to their conversations with convenience.

3. Data Preprocessing: The system undertakes essential steps to refine the raw chat data before analysis. This involves a series of transformations aimed at enhancing the quality and relevance of the dataset. Firstly, the module eliminates unnecessary elements such as stop words and punctuation marks, which do not contribute to the analysis and may introduce noise. Additionally, the text is tokenized, breaking down sentences into individual words or tokens to facilitate further processing.
The module is responsible for cleaning the text by removing any irregularities or inconsistencies present in the raw data. This ensures that the data is standardized and uniform, ready for meaningful analysis. By retaining only, the information required for analysis, the dataset is streamlined, reducing unnecessary clutter and improving the efficiency of subsequent analytical processes.

4. Data Analysis: Users have the option to choose between conducting a comprehensive group analysis or a more targeted examination of individual users within the system. Once users have made their selection, they can proceed by clicking the "Display Analysis" button to initiate the evaluation process of the imported WhatsApp file. Following this, the system generates and presents the analysis results derived from the WhatsApp text file in a user-friendly format, making it easy for users to review and interpret the outcomes of the analysis.

5. Statistical Representation: This module emphasis lies in presenting the preprocessed data through a range of graphical elements. These visual representations serve as effective tools for conveying the analyzed data in a clear and intuitive manner, facilitating the interpretation of insights derived from the analysis. By utilizing graphical elements, users are provided with visual cues that enhance their understanding of the data and enable them to discern patterns, trends, and relationships more easily. This approach fosters a more engaging and insightful exploration of the analyzed data, ultimately enriching the user experience and enabling informed decision-making based on the insights revealed through visualization.

6. ARCHITECTURE

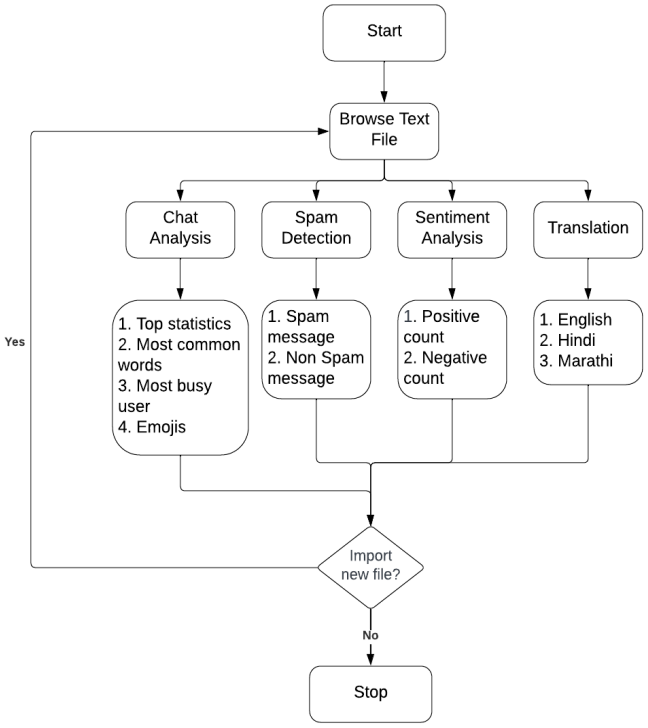

Figure 6.1: The System Architecture

The system architecture depicted in Figure 6.1 provides an overview of the website's functionality. Users initiate the process by exporting their WhatsApp chat file in text format and importing it into the WhatsApp chat analysis tool. Within the tool, users have the option to perform an overall analysis of the chat data or focus on a specific user's activity. The chat analysis feature offers insights such as identifying the most active users, common messages, word cloud representations, top statistics, and emoji usage patterns. Additionally, the sentiment analysis tool categorizes messages as positive or negative, while the spam detection feature identifies the number of spam messages present. Moreover, the translation functionality enables users to translate chat texts into English, Hindi, and

Marathi languages, enhancing accessibility and cross-cultural communication.

7. MODULE DESCRIPTION

The module descriptions provide an insight into the various components of our WhatsApp chat analysis system, each designed to fulfill specific tasks in the analysis process.

Chat Analysis: The chat analysis component of our project focuses on extracting valuable insights from WhatsApp conversations. It involves identifying key statistics such as the most active users, common messages, and overall activity trends. Additionally, it includes features such as word cloud generation to visually represent frequently used words, as well as emoji analysis to track emoji usage patterns within chats.

Spam Detection: Spam detection plays a crucial role in maintaining the integrity and quality of WhatsApp conversations. This component utilizes machine learning algorithms to identify and filter out spam messages effectively. By analyzing message content, frequency, and other relevant features, the system distinguishes between legitimate messages and spam, ensuring a clutter-free communication environment for users.

Sentiment Analysis: Sentiment analysis aims to understand the emotional tone and sentiment expressed in WhatsApp messages. By employing natural language processing techniques and machine learning algorithms, this component categorizes messages as positive, and negative. Users can gain insights into overall sentiment trends within their chats, allowing them to gauge the emotional dynamics of their conversations.

Translation: The translation component enhances cross-cultural communication by enabling users to translate WhatsApp messages into different languages. Supporting languages such as English, Hindi, and Marathi, this feature facilitates seamless communication among users from diverse linguistic backgrounds. Leveraging machine translation models, the system ensures accurate and efficient translation of chat texts, promoting inclusivity and accessibility within the platform.

8. SYSTEM DESIGN

In the software design process, several key principles have been adhered to for a structured and efficient system architecture:

1. Modularity and Partitioning: Organize the system into distinct modules based on functionality, including data preprocessing, analysis algorithms, visualization, user interface, and data storage. Define clear boundaries between modules to ensure each component has well-defined responsibilities and communicates with others through standardized interfaces.

2. Coupling: Strive for low coupling between modules by minimizing direct dependencies. Utilize interfaces and abstract classes to decouple components, facilitating easier maintenance and updates.

3. Cohesion: Emphasize high cohesion within modules by focusing each on a single, well-defined task. Encapsulate related functionalities within modules to achieve strong internal coherence.

4. Shared Use: Identify opportunities for code reuse across different modules by identifying common functionalities or algorithms. Develop shared libraries or utility classes to promote code reuse and maintain consistency. Implement mechanisms for sharing resources such as data structures or configuration files among modules to enhance efficiency and prevent duplication of efforts.

9. DESIGN AND DEVELOPMENT

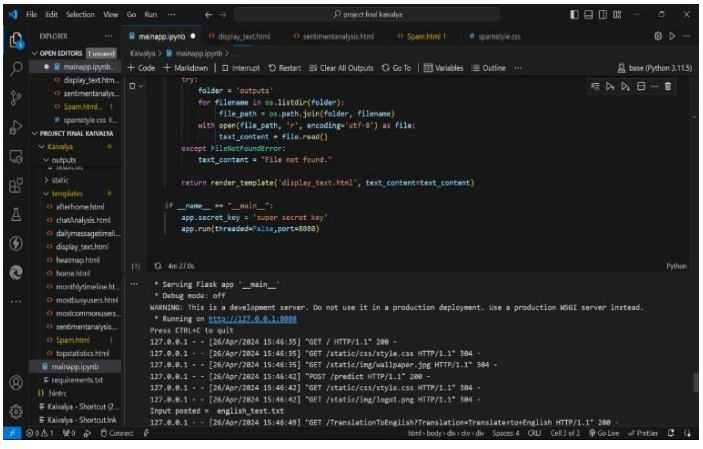

Figure 9.1: Local Host Connection

Figure 9.1 illustrates the connection to the local host webpage, facilitating access to the chat analysis functionalities for users.

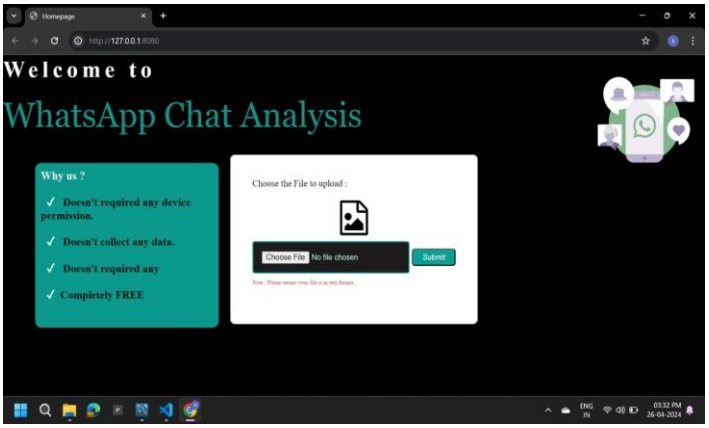

Figure 9.2: Home page

Figure 9.2 showcases the homepage interface, where users can upload their WhatsApp chat text file by clicking on the "Choose File" option.

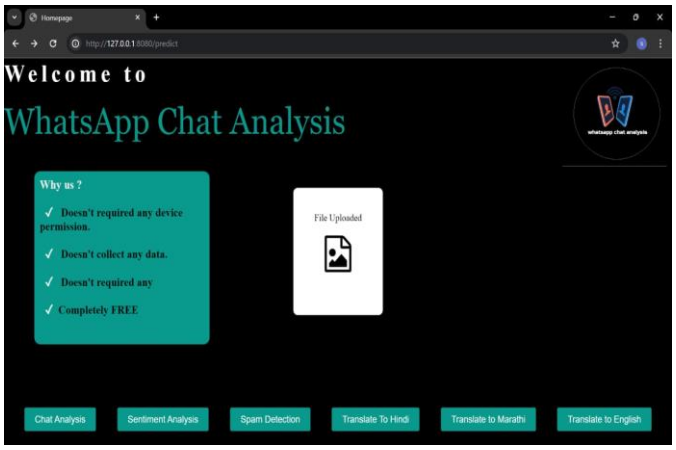

Figure 9.3: Home Page

Figure 9.3 illustrates the system's capability to import selected files for subsequent analysis and processing, streamlining user interaction and facilitating effortless data input.

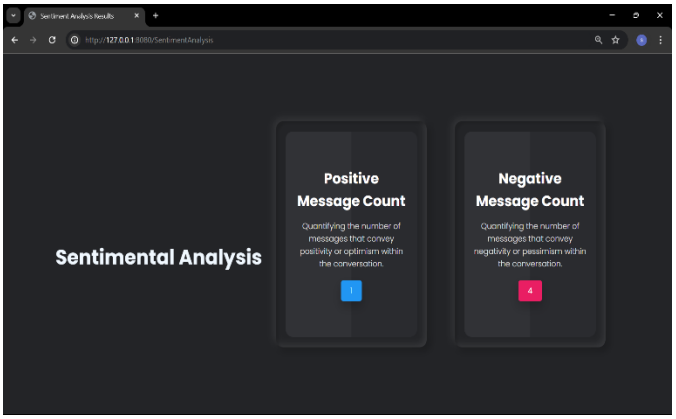

Figure 9.4: Sentiment analysis page

In Figure 9.4, we see the sentiment analysis page, where the system presents an overview of positive and negative sentiments extracted from the chat data.

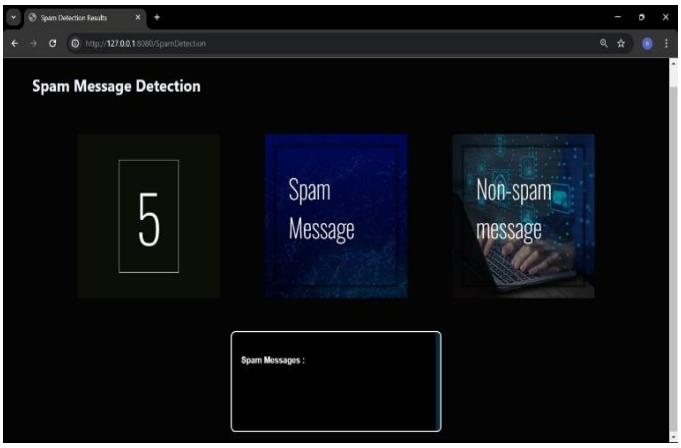

Figure 9.5: Spam detection page

Figure 9.5 illustrates the spam detection page of our system, where users can access insights regarding the total number of messages present in the chat file. By offering a summary of message count, users can understand their chat activity level and gauge the potential impact of spam messages on their communication.

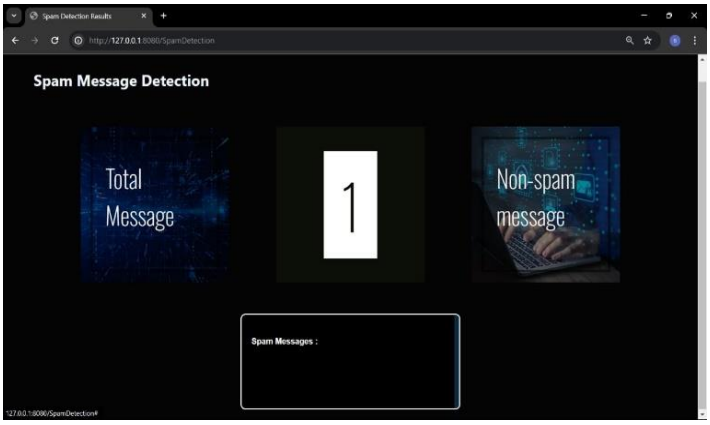

Figure 9.6: Spam detection page

Figure 9.6 illustrates the proportion of spam messages relative to the total number of messages present in the chat file.

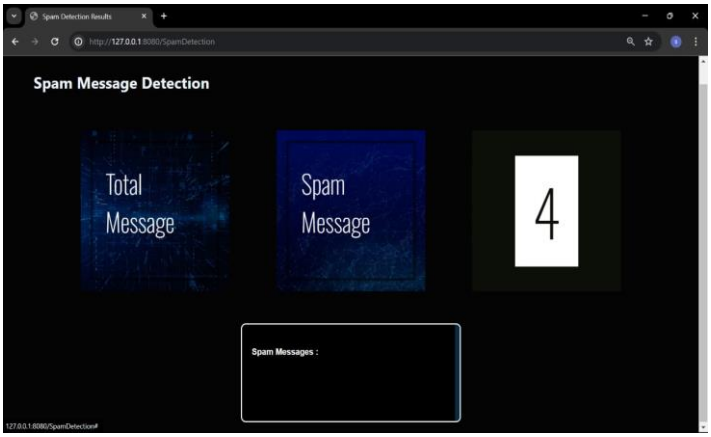

Figure 9.7: Spam detection page

In Figure 9.7, the graph illustrates the proportion of non-spam messages within the chat file relative to the total number of messages.

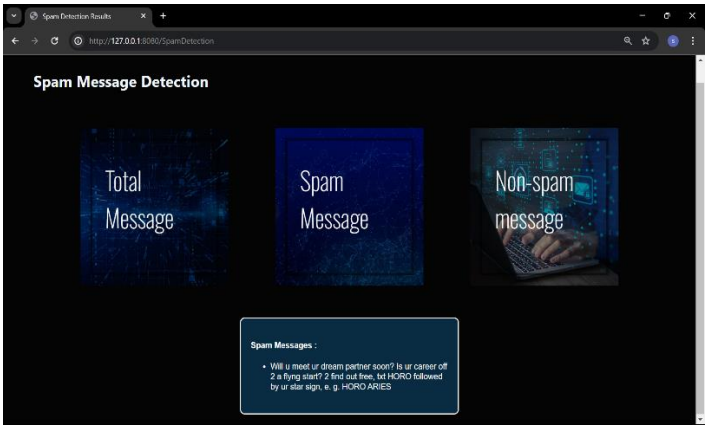

Figure 9.8: Spam detection page

In Figure 9.8, our system's spam detection page interface is highlighted, offering users an overview of essential metrics concerning spam messages. This involves showcasing identified spam messages to users, offering insights into the occurrence and characteristics of spam in their chat data.

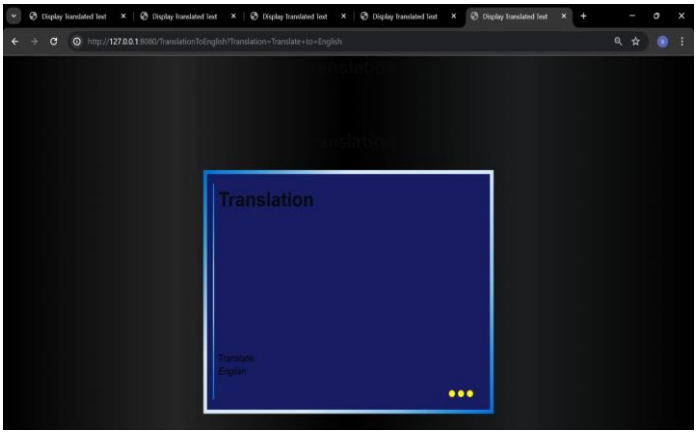

Figure 9.9: Translation page

Figure 9.9 illustrates the translation interface within our project's webpage.

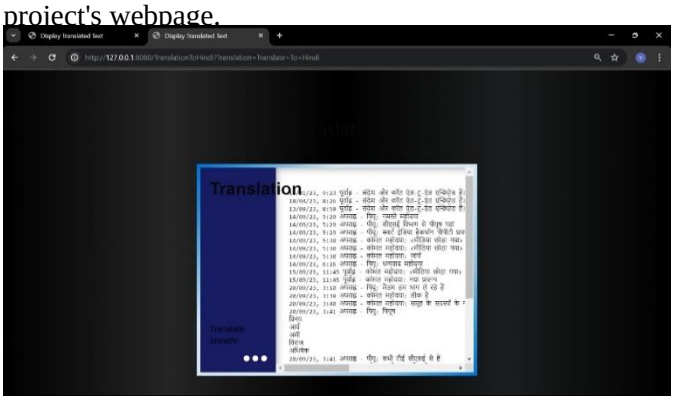

Figure 9.10: Translation page

In Figure 9.10, observe the translation webpage of our project, facilitating the translation of chats into Marathi.

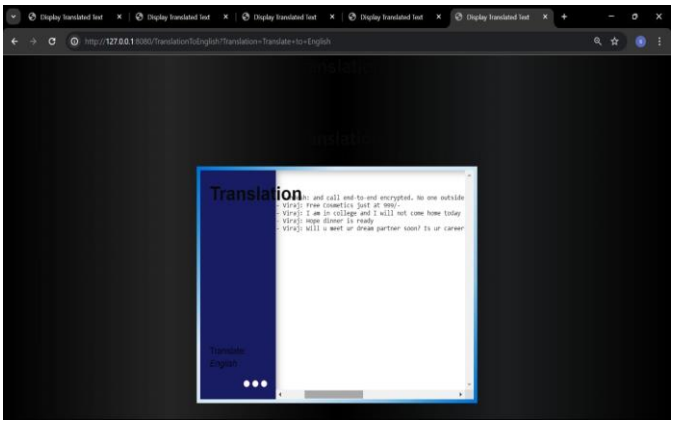

Figure 9.11: Translation page

In Figure 9.11, we present the translation webpage within our project, offering users the capability to translate chats into English.

10. LIMITATIONS

The “WhatsApp Chat Analysis” project endeavors to provide comprehensive WhatsApp chat analysis capabilities, but it's important to acknowledge certain limitations. Firstly, the accuracy of analysis results, particularly in sentiment analysis and spam detection, may be influenced by the complexity and nuances of natural language processing. Additionally, the effectiveness of translation functionality may vary depending on the language pair and the quality of machine translation models employed. Furthermore, the project's scalability may be constrained by hardware resources and processing capabilities, potentially limiting its ability to handle large volumes of chat data efficiently. Moreover, the project's usability may be impacted by user familiarity with machine learning concepts and the complexity of the analysis interface. Finally, the project's reliance on external APIs or libraries for certain functionalities may introduce dependencies and potential points of failure. Acknowledging these limitations is crucial for managing user expectations and guiding future development efforts to address these challenges effectively.

11. CONCLUSION

The WhatsApp Chat Analysis using Machine Learning project successfully implemented advanced data processing techniques, achieving a detailed understanding of communication patterns, sentiments, and emerging topics in WhatsApp conversations. The integration of machine learning models enhanced analysis accuracy, and graphical visualizations provided clear representations for pattern recognition. The project validates the efficacy of machine learning in extracting insights from messaging data, showcasing real-world applications in marketing, customer engagement, and user behavior analysis. The systematic approach and cutting-edge technologies pave the way for an informed and data-driven approach to WhatsApp chat analysis.

12. FUTURE RESEARCH

The "WhatsApp Chat Analysis using Machine Learning" can be further enhanced to provide greater flexibility and performance with certain modifications whenever necessary. Chat analysis could be used in educational settings to understand student engagement, collaboration, and learning outcomes. Machine learning models could predict trends, sentiment shifts, translation and emerging topics based on historical chat data. Real-time analysis of WhatsApp chats could become more prevalent, allowing businesses and individuals to monitor conversations as they unfold. This capability could facilitate tasks such as customer support, crisis management, and trend identification, enabling proactive engagement with ongoing conversations.

13. REFERENCES

1. Chat Analysis on WhatsApp using Machine Learning (2023) T.Naga Nandini, M.Harsha Vardhan, Journal of Engineering Sciences, Vol 14 Issue 04,2023

2. Analysing and Predicting the Emotion of WhatsApp Chats Using Sentiment Analysis, March - April 2020 ISSN: 0193-4120 Page No. 15454 – 15461

3. Whatsapp Chat Analyzer, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV9IS050676. Vol. 9 Issue 05, May-2020

4. Keagan Stokoe (2020) Text and Sentiment Analysis of WhatsApp Messages. Available: https://levelup.gitconnected.com/text-and sentiment-analysis-of-Whatsapp-messages-1eebc983

5. Sayali Meshram, Ms. Nisha Balani, Ms. Parul Jha (2019). Sentiment Analysis of Statement. IOSR Journal of Engineering (IOSRJEN). Vol. 09, Issue 5 (May. 2019), ||S (XI) || PP 46-49

6. Joshi, Sunil. (2019). Sentiment Analysis on WhatsApp Group Chat Using R. 10.1007/978-981-13-6347-4_5.

7. Yaakub, Mohd Ridzwan & Abu Latiffi, Muhammad Iqbal & Safra, Liyana. (2019). An Overview of Sentiment Analysis Methods and Their Practical Applications. IOP Conference Series: Materials Science and Engineering. 551. 012070. 10.1088/1757-899X/551/1/012070.

8. Kootbodien, Ammaarah & Prasad, Nunna & Ali, Muhamad. (2018). Trends and Influence of WhatsApp Usage among Students in Abu Dhabi. Media Watch. 9. 10.15655/mw/2018/v9i2/49380.

9. Z. Feng, Q. Chen, Z. Xu, and M. Yu, " "Analyzing Emotions in Conversations using Recurrent Neural Networks," presented at the Conference on Empirical Methods in Natural Language Processing, 2018.

10. 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" (2016).

11. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). "Google's multilingual neural machine translation system: Enabling zero-shot translation." Published in Transactions of the Association for Computational Linguistics, Volume 5, pages 339-351.

12. Vaswani et al.'s paper titled "Attention is all you need" was presented at the Advances in Neural Information Processing Systems conference in 2017, specifically appearing on pages 5998-6008.

13. Gyawali, Nepal, and Arora's paper, titled 'Spam Detection in Online Social Networks: A Survey,' was published in IEEE Access in 2019.

14. Dhanapal, V., & Saravanan, S. (2020). "WhatsApp Spam Detection: Machine Learning Approach." Published in International Journal of Computer Applications.

15. "Detecting Spam in WhatsApp Group Chat Using Machine Learning Techniques" by T. S. Shinde and D. S. Kapse. (Published in International Journal of Computer Applications, 2021)