WhatsApp Chat Sentimental Analysis

Shailesh Kumar Verma
Galgotias University
Greater Noida, India
shailesh29062001@gmail.com

Dr. Brijesh Kumar Singh
Galgotias University
Greater Noida, India
brijesh.singh@galgotiasuniversity.edu.in

Sudhanshu Tripathi
Galgotias University
Greater Noida, India
sudhanshutripathi8888@gmail.com

Abstract*— With the exponential growth of social media users, people are really curious about the type of conversation they have with other people and also how communication leads in group chats, with this Web App tool we focus on providing the user with insightful analysis of their chats with other people, some of the statistics include, total messages, total words, most used words, emoji analysis, and some more, off all the social media chatting apps we have chosen WhatsApp as it is used purely for conversations purposes so that it can provide us with the most helpful dataset, after performing these analysis, we have also performed machine learning analysis using supervised machine learning algorithm Random Forest Classifier in order to perform sentimental analysis, we have used Random forest classifier as it is a boosting algorithm and provides extremely good results for binary classification. We will export our WhatsApp chats and we can directly upload it to our Web App to get the sentimental analysis.

Keywords—WhatsApp Chat, nlp, classification, random forest, sentimental analysis, web app, machine learning.

I. INTRODUCTION

In an age dominated by instant communication, platforms like WhatsApp have become an integral part of our daily lives. WhatsApp, a globally recognized instant messaging application, has revolutionized the way people connect and communicate. Consequently, it has become a valuable source of information for examining the emotional dynamics of human interaction. Sentiment analysis, a subset of Natural Language Processing (NLP), plays an important role in understanding the emotions and ideas expressed in these digital conversations The aim of this paper is to explore the aspect of sentiment analysis in WhatsApp conversations The process of identifying and classifying emotional or emotional tone has been expressed in writing, as positive, negative, or neutral Has seen applications in many fields, such as customer feedback research, product reviews , social media management, and, now, interpersonal analysis in popular messaging services. Anecdotally, WhatsApp provides a fertile ground for studying how individuals express their emotions in digital conversations Analyzing WhatsApp conversation data not only provides an opportunity to explore the emotional nuances of these interactions but also if they have implications it is also useful for mental health support, customer service, and user engagement f Machine learning algorithms have proven to be effective in text classification tasks.

Its ability to handle large and complex data sets, while producing reliable results, makes it ideal for sensitivity analysis. This paper aims to demonstrate the capability and performance of random forest in WhatsApp chat sentiment analysis. We organized the rest of the paper as follows. Section II shows several related works. Section III describes the proposed approach. Section IV shows the evaluation of the proposed approach and the experimental results. Finally, Section V concludes the paper.

The following sections of this paper will delve into the methodology, data collection, and preprocessing steps, followed by a detailed explanation of how the Random Forest classifier is used to predict sentiment in WhatsApp chats. We will also present the results and insights gained from analyzing a real-world dataset, illustrating the practical implications of our research. Ultimately, by exploring WhatsApp chat sentiment analysis using the Random Forest classifier, this research paper contributes to the ever-evolving landscape of NLP and sentiment analysis, offering valuable insights into the emotions expressed in one of the most popular instant messaging platforms globally.

II. RELATED WORK

Many people have proposed models to perform sentimental analysis on social media applications specially on WhatsApp which provide amazing dataset for analysis, some of the papers, with their conclusion on the research are discussed in this section. The first one we are going to discuss is Church, K. **[1]** draw comparisons between SMS and MIM to illustrate WhatsApp's success. Additionally, they transmit that "WhatsApp messages are generally more conversational, social, and informal in nature, whereas SMS is perceived as more formal, privacy-preserving, and generally more reliable." Sana Shahid **[2]** set out to investigate if WhatsApp fosters interpersonal relationships and found that professionals use it more frequently to accomplish work-related objectives. He found and recorded the frequency, demographics, and usage patterns of WhatsApp users as well as the kinds of conversations they had.

Popescu **[3]** used n-gram generations for sentiment prediction and POS tagging analysis using NLP-based approaches. Alekh Agarwal and Pushpak Bhattacharyya **[4]** Focused on adjectival words, Agarwal et al. found that these words function as polarity indicators more so than non- adjectival words.Hai-Bing Ma **[5]** Since these are the most often used polarity terms, Hai-Bing Ma et al. advise starting by identifying the positive words—such as good, great, outstanding, awesome, and fine—and the negative words—such as bad and poor. Kaur, R. **[6]** In order to assess the orientation of the group, the attitudes expressed during visits were categorized into ten different groups in this article. Using devices, the material was divided up and then processed to provide the most appropriate terms. After that, the information was finally classified by comparing it to the lexicons in the installed libraries. Sarcastically worded sentences must be carefully extracted from irrelevant context and examined independently for each characteristic to be analyzed.

Winarko, E., &Cherid **[7]** Sarcastic sentence classification is a challenging undertaking as each feature requires physical processing. The outcome illustrates the significance of sarcasm identification, regardless of the degree of sarcasm and the intended context. We may infer the assessment of an announcement with mocking attitude from the findings. This feature may be used to extract satirical sentiment from a group of WhatsApp messages that contain material in Indonesian.
[8] The research, titled "Impact of WhatsApp on the Youth: A Sociological Investigation," involved the random selection of 100 WhatsApp users aged between 18 and 30. Findings revealed that 63% of users engage with the app around fifty times a day, 21% use it roughly twenty times daily, and 16% exceed a usage frequency of over a hundred times per day. This suggests a significant influence of WhatsApp on the communication patterns among young individuals. Kootbodien **[9]** In a study conducted among Abu Dhabi students, it was found that 85% of women and 70% of men use emojis to convey emotions in text, serving as a replacement for facial expressions. On average, their usage spans between 1 to 7 hours per day. Aravind K. Joshi **[10]** In an initial study conducted in 1982, Aravind K. Joshi defined code switching or code- mixing as the intentional transition between languages within a single utterance among speakers in certain bilingual communities. The creation and understanding of utterances involving intra-sentential code-switching are considered integral to the linguistic abilities of both the speakers and listeners within these specific communities.

III. METHODOLOGY

In this approach first, we need to install the required libraries for the model to train on the given dataset, after installing we need to import them, then we can start pre- processing the data for extraction, cleaning and classification, data pre-processing is a multi-step process:

- Column Modification: As this dataset is not perfect for our analysis, we need to modify it to a certain extent to make it acceptable for analysis, first we modified the “annotation” column in our data frame and drop the column called “extras”.
- Content Modification: We used the “nltk” library to remove stop words and punctuation from the dataset so that it can be ready for tokenisation.
- Tokenisation: Now we used PoterStemmer and tokenised the content and convert it into lower case.

Now for the next part, we move to the feature extractionstep.

In this step, we modify the textual data in a format that can be fed to the Machine Learning Algorithm. We will use TfidfVectorizer which stands for Term Frequency-Inverse Document Frequency, which tells us how important a word is in a series or corpus is to a text. The meaning can significantly increase if we increase the frequency of the words, we are using this so that we can give some type of weight or relevance metrics to the words. The next step is to perform the classification of the data, as in this paper we are using the Ensemble methods we are dividing the classification model.

As it was difficult to get relatively reliable labelled dataset for WhatsApp, we have performed the sentimental analysis on a reliable dataset found on Kaggle which extracts its data from twitter conversations.

The last step is to extract the data we have derived from our WhatsApp chat and pre-processed it for analysing in our Random Forest

We will judge our model based on the confusion matrix and 4 more criteria; the equation used to solve them are –

1. Precision
2. Accuracy
3. Recall

IV. GUI

Graphical User Interfaces (GUIs) are essential components of many software applications, providing users with intuitive ways to interact with and manipulate data. GUIs typically consist of visual elements such as windows, buttons, text fields, menus, and other graphical objects that users can interact with using input devices like a mouse or keyboard.

The GUI is created from using Python Library called Streamlit, that is used for creating graphical user interface for web apps, it is easy to use, reliable, scalable, and robust.

The purpose of this GUI is to provide simple analysis on the WhatsApp chats of users by just using exported chats from their WhatsApp with the hope of integrating sentimental analysis feature in the Web App.

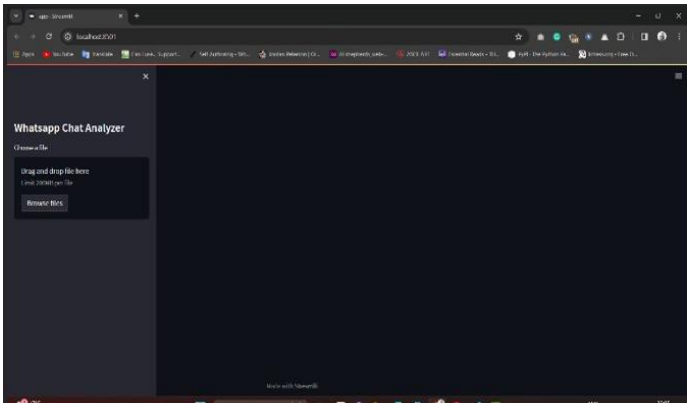

Fig. 1. Showcasing GUI made from Streamlit

The features of GUI include:

1. Upload exported chats from WhatsApp.
2. Check analysis for both individuals and overall.
3. Monthly timeline graph, showing monthly message frequency.

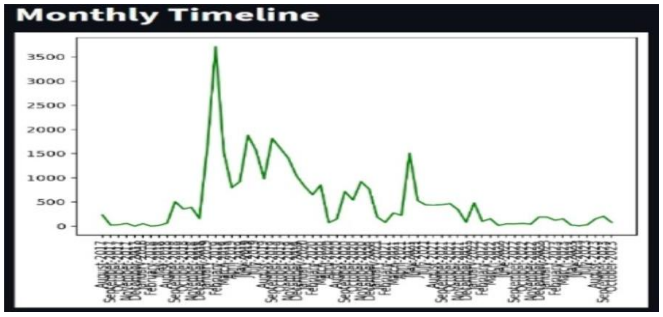

Fig. 2. Monthly Timeline Graph

1. Daily Timeline graph showing daily activity frequency.

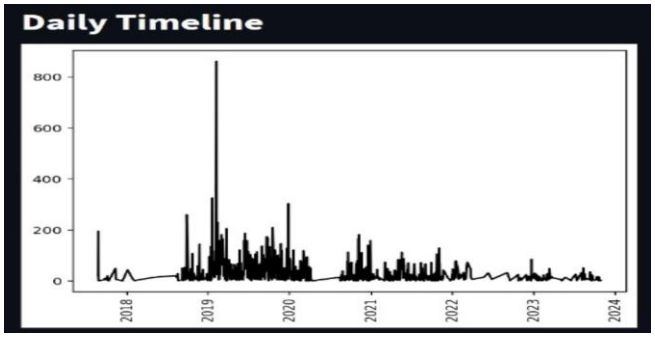

Fig. 3. Daily Timeline Graph

2. Activity Map showing graph of busy months and days.

Fig. 4. Activity Map

3. Weekly Activity heat map.

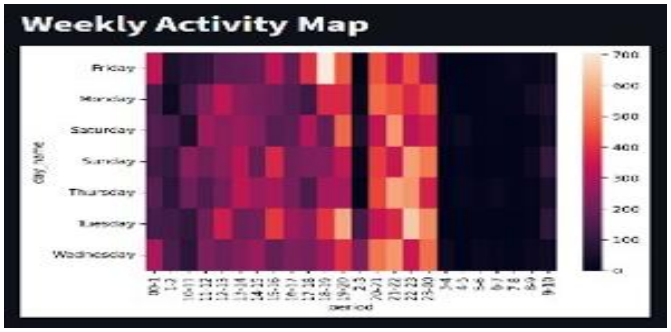

Fig. 5. Weekly Activity Map

4. Word Cloud.

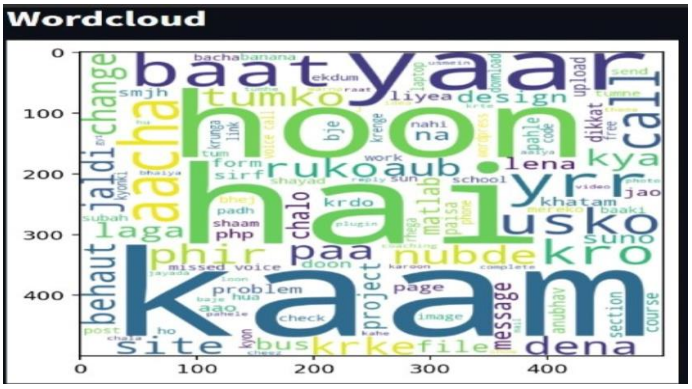

Fig. 6. Word Cloud

5. Most Common Words.

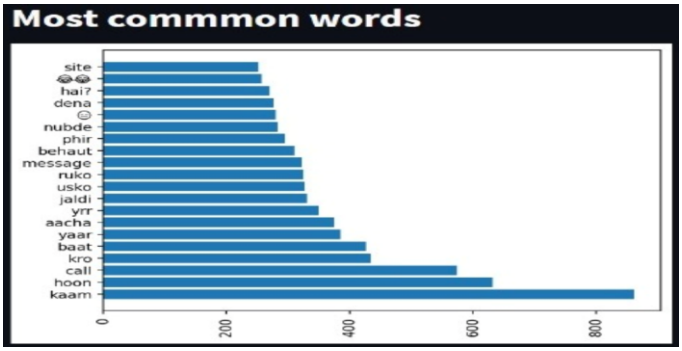

Fig. 7. Most common words

The preprocessing of data from exported WhatsApp chat is done differently from preprocessing of data taken from Kaggle dataset, the following steps were taken:

1.Convert textual data from txt file to a data frame.
2.Convert columns based on time format of date/month/year, hour: minute.
3.Applying regular expression to identify extraction of date time – regex used (([\ w \ W]+?):\s).
4.From the extracted date time extract date time individually.

V. EXPERIMENTAL RESULTS

In this section, description of experimental result of our proposed approach is given. The evaluation of the proposed approach on the dataset which was taken from Kaggle is done. In the following, we describe the data and the result.

A. Data Description

The dataset is taken from Kaggle which was partially manually labelled and is publicly obtainable. The dataset in general contains 20000 conversations taken from Twitter, the differentiation between offensive and non- offensive conversations is by the column annotation.

The Labels are divided into the categories: (Positive) 1 (Negative) 0. There are 12179 total Positive Cases, and 7822 to Negative cases.
This data is reliable as it is formally labelled to be used for supervised machine learning algorithm, like the one we have used in this project.

TABLE I. THE DATASET STATISTICS

Source of Data	Twitter
Source of Dataset	Kaggle
Total amount of Conversations	20000
Positive	12179
Negative	7821

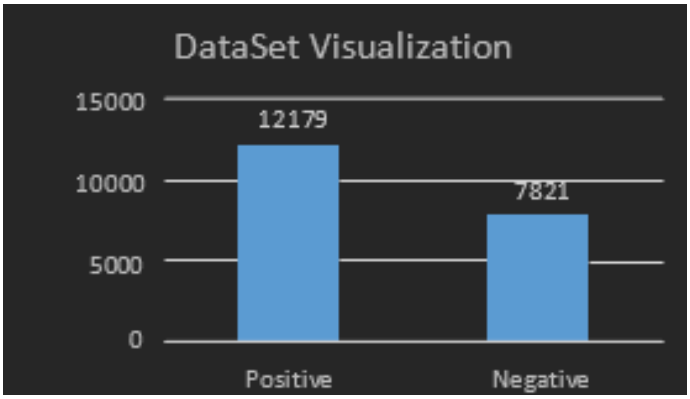

Fig. 8. Total Number of Positive and Negative Data

B. Results

After processing the dataset, we are following the steps mentioned in Section III of the research paper to extract the features. (0.9,0.1) is the ratio in which a dataset is split for training and testing. To properly measure the performance of our classifier we have used metrics like Precision, Accuracy, Recall and f-score.

As mentioned earlier, we have used 2 types of classification methods, normal classification, and ensemble classification, as an experiment to understand the difference between their classification strength and how better ensemble methods work. The ensemble method converts the main dataset into bootstrap samples without replacement to have multiple models do classification and then it will use a voting system to choose the best classification. Table II summarizes the accuracy of the method, namely, Random Forest Classifier.

TABLE II. ACCURACY OF THE MODEL

Classifier	Accuracy %
Random Forest	92.80%

In addition to accuracy, we have also shown in Table III and Table IV that three classifiers Precision and Recall for better comparison. With the tested data we have seen that Random Forest Classifier has given us the best result on all the grounds whether it be accuracy, precision, recall or f- score.

TABLE III. PRECISION OF THE MODEL

Classifier	Precision %
Random Forest	90.50%

TABLE IV. RECALL OF THE MODEL

Classifier	Recall %
Random Forest	97.43%

VI. CONCLUSION

In conclusion, the sensitivity analysis using random forest classification proved to be very effective, yielding an impressive accuracy of 92% This indicates that the model is robust and self-confirming at the accurate distribution of sensitivity in the analyzed data. The ability of random forest algorithms to deal with complex relationships and capture subtle patterns in data has contributed to the success of high-precision sensory differentiation.
Not only is the records knowledge accurate but actual usability, projects, social media, and other types of domes can also benefit from the customer's assessment. For applications of cleaning, customer and satisfaction surveys, etc. The model is well suited. However, it is important to realize that although accuracy is an important metric, it is not the only measure of model performance. To obtain a detailed evaluation of the efficacy of the model, other metrics should also be considered such as accuracy, recall, and F1 scores In conclusion, the success of sensitivity analysis with random forest classification is exactly 92% of uses open the way to practical use.
As machine learning continues to evolve, the integration of such models across industries can be increased, providing valuable insights and contributing to improvements in decision-making processes

VII. REFERENCES

[1] Church, K., de Oliveira, R.: 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, pp. 352– 361 (2013)

[2] Sana Shahid, "Content Analysis of Whatsapp Conversations: An Analytical Study to Evaluate the Effectiveness of WhatsApp Application in Karachi", International Journal of Media, Journalism and Mass Communications (IJMJMC) Volume 4, Issue 1, 2018, PP 14-26 ISSN 2454-9479 http://dx.doi.org/10.20431/2454-9479.0401002 www.arcjournals.org.

[3] Popescu, Ana-Maria and Oren, Etzioni, "Extracting product features and opinions from reviews," In Proceedings of EMNLP,

2005.

[4] Alekh Agarwal and Pushpak Bhattacharyya, Augmenting WordNet with Polarity Information on Adjectives, Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings,
pp. 3–8. c Masaryk University, 2005.

[5] Hai-Bing Ma, Yi-Bing Geng, Jun-Rui Qiu, Analysis Of Three Methods For Web-Based Opinion Mining, Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 10-13 July, 2011

[6] Kaur, R. (2019). Insight to Emotional tones in WhatsApp Through Sentiment Analysis. IJRAR.

[7] Winarko, E., &Cherid, A. (2017, November). Recognizing the sarcastic statement on WhatsApp Group with Indonesian language text. In 2017 International Conference on Broadband Communication, Wireless Sensors and Powering (BCWSP)

[8] Bhatt, Anshu & Arshad, Mohd. (2016). Impact of WhatsApp on youth: A Sociological Study. IRA-International Journal of Management & Social Sciences (ISSN 2455-2267). 4. 376. 10.21013/jmss.v4.n2.p7.

[9]Kootbodien, Ammaarah & Prasad, Nunna & Ali, Muhamad. (2018). Trends and Impact of WhatsApp as a Mode of Communication among Abu Dhabi Students. Media Watch. 9. 10.15655/mw/2018/v9i2/49380

[10] Aravind K. Joshi. (1982) Processing of sentences with intra-sentential code-switching. In Proceedings of the 9th conference on Computational linguistics - Volume 1 (COLING '82). Academia Praha, CZE, 145–150. DOI:https://doi.org/10.3115/991813.991836

[11] Marada Pallavi, Meesala Nirmala, Modugaparapu Sravani, Mohammad Shameem. WhatsApp Chat Analysis. International Research Journal of Modernization in Engineering Technology and Science. Volume: 04/Issue:05/May-2022

[12] Shaikh Mohd Saqib. Whatsapp Chat Analyzer. International Research Journal of Modernization in Engineering Technology and Science. Volume: 04/Issue:05/May-2022

[13] K, Ravishankara & Dhanush, & Vaisakh, & S, Srajan. (2020). Whatsapp Chat Analyzer. International Journal of Engineering Research and. V9.10.17577/IJERTV9IS050676.

[14] D.Radha, R. Jayaparvathy, D. Yamini, "Analysis on Social Media Addiction using Data Mining Technique", International Journal of Computer Applications (0975 – 8887).

[15] https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

[16] https://www.analyticsvidhya.com/blog/2021/06/build-web-app-instantly-for-machine-learning-using-streamlit/

[17] Meng Cai, "PubMed Central", PMCID: PMC7944036, PMID: 33732917

[18] E. Larson, \"[Research Paper] Automatic Checking of Regular Expressions,\" 2018 IEEE 18th International Working Conference on Source Code Analysis and Manipulation (SCAM), 2018, pp. 225-234, doi: 10.1109/SCAM.2018.00034.