***Abstract*— Today social media is the most used way to express emotions and share feelings, and WhatsApp is one of the most popular social media platforms. The model emerging from this research study aims to perform a data and sentiment analysis of WhatsApp chats of text messages of the user. The creation of a model that categorizes opinions as either positive, negative, or neutral is crucial. The mining of behavior is another name for sentiment analysis. Using Natural Language Processing (NLP) and information retrieval techniques, sentiment analysis comprises behavior, opinions, and sentiments of the text, chat, or online communication. The model proposed in the paper aims to examine the trends of WhatsApp usage in the youths belonging to the age group of 18-25 years and their sentimental behavior. The behavior patterns are eradicated during the project. The only group chats are focused for the proposed model and study. It will make it easier to determine whether the person enjoys using emojis and whether they do so as an adjunct to verbal communication or to express their emotions more fully. The benefit of the model is that it was developed using simple Python modules like pandas, matplotlib, seaborn, and NLTK which are used to create information frames and plot various types of graphs**.

Analyzing WhatsApp Chat Sentiments using Natural Language Processing (NLP) Algorithm

Nikhil Kulthe, Jatin Shevkari, Omkar Harkulkar, Abhijeet Patil

School of Computer Engineering and Technology

MIT Academy of Engineering

Alandi (Devachi), Pune, Maharashtra, India

[nikhil.kulthe@mitaoe.ac.in](mailto:nikhil.kulthe@mitaoe.ac.in), [jatin.shevkari@mitaoe.ac.in](mailto:jatin.shevkari@mitaoe.ac.in), [omkar.harkulkar@mitaoe.ac.in](mailto:omkar.harkulkar@mitaoe.ac.in), [Abhijeet.patil@mitaoe.ac.in](mailto:Abhijeet.patil@mitaoe.ac.in)

***Keywords*—WhatsApp; Sentiment Analysis; Natural Language Processing; Emotion Analysis; Emoji; Chats; KNN**

# I. INTRODUCTION

With more than 2.24 billion registered users across the world and one billion messages delivered daily in 60 different languages, WhatsApp is one of the most popular social media sites today [1]. WhatsApp plays a significant role in our lives. WhatsApp is used to communicate and express the most private feelings, which are frequently described as deep, meaningful conversations. Youth generation between the ages of 18 and 25 was found to spend 8 to 16 hours per day using WhatsApp, according to a survey conducted in India by Mood of the Nation (MOTN). It can be inferred from this that WhatsApp messages can reflect a person's mental process and express their feelings [2].

Sentiment analysis has two driving forces. Customers’ opinions about products and services are tremendously valuable to both consumers and manufacturers. Sentiment analysis has therefore observed correct smart effort from the industry. Sentiment analysis is to use this data to gather essential knowledge about beliefs to create better corporate decisions, political campaigns, and increased product consumption [3].

The pre-processing stage of text data mining is crucial. Any process involving the usage of data should start with data pre-processing. This is so that it produces data sets that are far more manageable, clearer, and coherent [4]. Depending on the type of data, there are several procedures involved in pre-processing it. In the proposed model in the paper, the dataset includes both textual and graphic information. So, it is essential to do segregation before using them for further analysis.

The model takes into account the importance of emojis while communicating via messaging and emotion categories for sentiment analysis in addition to data pre-processing and NLP used for sentiment analysis. The model put forward in the research tries to help people comprehend the different sorts of communication. This study shows that it is possible to supply ML models—which essentially chat data investigators—with superior input. The proposed model's accuracy is increased by the necessity for the appropriate learning situations. The suggested model ensures detailed exploratory data analysis of various WhatsApp chat kinds.

Section 2 of the paper introduces past research on texting psychology, WhatsApp chat analysis, and emotion classification. Section 3 contains the methodology for data gathering, data pre-processing, and key insights. Section 4 discusses the proposed system. Section 5 depicts the experimental setup of the system. Lastly, section 6 presents the findings, and section 7 presents the conclusion.

# II. LITERATURE REVIEW

WhatsApp got here into the marketplace rather than SMS. it is far used to make voice or video calls in addition to percentage media. Authors of “Survey Analysis on the Usage and Impact of WhatsApp Messenger” of their study of a collection of WhatsApp customers of ages 18-50 found that approximately seventy-9 percent of their topics use WhatsApp every day for a minimum of 15-60 minutes [4]. In the look “Impact of WhatsApp on Youth: A Sociological Study,” a hundred WhatsApp customers had been decided on randomly, lying among 18 and 30. They have a look at observed that the frequency of utilization for sixty-three percent of users is fifty times a day, twenty-one customers are twenty instances an afternoon and 16 percent is more significant than one hundred times an afternoon. it can be concluded that for some kids, WhatsApp greatly influences how they communicate [5].

Authors of “Facebook Sentiment: Reactions and Emojis,” argue that linguistic textual messages and emojis adjust one another’s that means. The interplay of linguistic textual content with emojis can range. Emojis can be used to express emotions more strongly, to toughen up language, to express a subject's mood or attitude on their own, to emphasize an already present emotion in communication, or just out of politeness. Emojis can be used to express tough emotions depending on the context [6].

Authors of “Text Classification based Behavioral Analysis of WhatsApp Chats” posted in 2019, observed a machine to take multiple chats for each situation, and with the aid of an impartial community, every sentence and emoji is scored. The composition of the feelings expressed by the subject is less [7]. In the paper “Pre-Processing and Emoji Classification of WhatsApp Chats for Sentiment Analysis” published in 2020, the authors discovered that WhatsApp is widely used internationally and the growth in usage of emojis is growing which is the most beneficial fact for sentiment analysis. we ought to carry out various statistics mining techniques, due to which our version will become heavy and slows down velocity [8]. From the paper “A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis” posted in April 2020 we get there may be a want for a method that will categorize the critiques into wonderful, negative, and impartial sentiments and we will do it via NLP set of rules and records retrieval strategies. The method that categorizes the opinion of sentiment is mechanically run, so it could have much less accuracy [9].

“Code Mixing: A Challenge for Language Identification in the Language of Social Media” describes that despite the majority of social media content is in the English language, this still accounts for the majority of content globally. this is due to the fact that most users switch between languages in the middle [10]. In a look at the finish with the aid of M Tafaquh Fiddin Al Islami, Ali Ridho Barakbah, and Tri Harsono within the paper “Interactive Applied Graph Chatbot with Semantic Recognition” posted in October 2020, authors have discovered that there is a want to increase the retention charge of the customer and we can do it by using chat-bot and another cost-effective way is the way of studying sentiment via chats and emoji’s [11]

In a have a look achieved by the scholars of Abu Dhabi, “Trends and Impact of WhatsApp as a Mode of Communication among Abu Dhabi Students,” 80-five percentage of girls and seventy percent of guys expressed that they used emojis rather than facial expressions in the textual content. Their average usage becomes 1 to 7 hours/in line with the day [12]. In an observation “Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users” on four million users from different countries, it was discovered that six million, or seven percent of all messages, featured at least one emoji. This demonstrates that consumers are aware of emojis [13]. In the paper “An Emotional Topic Recommendation Chat Assistant” published in might also 2021, the authors have found out that assistants looking to address this trouble with the aid of recommending fantastic and negative emotional topics to customers. there is an absence of tone and nonverbal cues whilst chatting on WhatsApp [14]

In the version prescribed within the paper “Social emotion mining techniques for Facebook posts reaction prediction,” preliminary pre-processing like converting to lowercase, casting off hashtags, and so on. has been carried out using the Stanford CoreNLP Parser that is observed the use of a tokenizer to split posts based on spaces & after filtering the stop phrases [15]. The paper “Are Emoticons Good Enough to Train Emotion Classifiers of Arabic Tweets? “, showcases how emojis may be used to carry out the automatic labeling of tweets. automated labeling of tweets outperforms guide labeling of tweets [16].

The software of sentiment evaluation has been prolonged in latest years. for example, the sentiment evaluation techniques used for processing reviews given about clinical papers. "Sentiment analysis and opinion mining applied to scientific paper reviews" [17]. The sentiment analysis implemented Twitter textual content in "Twitter text mining for sentiment analysis on people’s feedback about Oman tourism” to investigate the comments of vacationers [18]. In another painting, the sentiment analysis software on product evaluations is presented in "Sentiment analysis on product reviews using machine learning techniques" [19]. In this paper, the class strategies which include NB have been used. For a complete survey on sentiment evaluation, readers are referred to "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers" [20].

# III. METHODOLOGY

The four stages of the method for sentiment analysis that is suggested in the study are:

*A. Data Collection*

*B. Data Pre-processing*

*C. Feature Extraction*

*D. NLP Algorithm working*

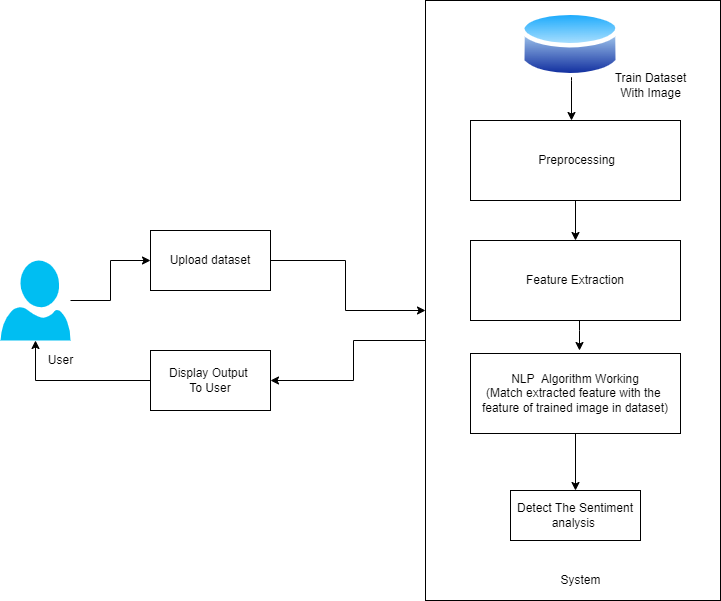
**

Fig. 1. Proposed Block Diagram

## *A. Data Collection*

The gathering of data is the first step in the analysis. One of the most important components of the sentiment analysis application is data collection. On a task that is particular to a domain, greater performance is not automatically guaranteed by massive datasets. The quality of the dataset affects the model's performance.

There are two possible ways to gather the data. The first is to collect preorganized data. Preorganized data is available on different sites, on these sites data is uploaded by developers for research purposes. The second way is the collect data manually. The model requires textual data so, chat data should be extracted into text files. This can be done by following steps:

* Creation of a WhatsApp account.
* Open WhatsApp. Select the chat (person/group) who wants to check his sentiment.
* Click on the tab, click on more, and select ‘export chat.’
* Select export chat ‘without media.’
* Copy the chats into the ‘.txt’ file.

During the proposed experimentation, we used the Python ‘textblob’ library to annotate the polarity of chats in pre-data.

## *B. Data Pre-processing*

Data pre-processing includes the steps we must take to change or encrypt data so that a machine can quickly and easily understand it. For the model to generate precise and accurate predictions, the algorithm must be able to quickly analyze the properties of the data. Due to their varied origins, the bulk of real-world machine-learning datasets is particularly prone to missing, inconsistent, and noisy data. Hence, pre-processing data is an essential step in sentiment analysis. Here are some pre-processing tasks:

* *Tokenization:* reduces the text to single clauses or semantic components
* Making up words like nouns, verbs, adjectives, adverbs, etc. is known as *part-of-speech tagging.*
* Standardizing words by reducing them to their base forms is known as *stemming and lemmatization.*
* *Stop word removal:* removing typical words like prepositions and articles that do not contribute much or any distinctive information.

Several opinions are conveyed in chats in various methods by various people. The WhatsApp dataset utilized in the proposed experimentation model work has already been categorized into classes for the neutral, negative, and positive polarity, making it simple to conduct sentiment analysis on the data and examine the impact of various aspects. Raw data with polarity is very prone to redundancy and inconsistency. Pre-processing of chats includes the following points,

* Eliminate all URLs (such as [www.abcd.com](http://www.abcd.com)), targets (such as @username) and media files.
* Check the spelling; repeating characters should be handled.
* Substitute their sentiments for each emoticon.
* Eliminate all punctuations, symbols, and numbers.
* Eliminate Stop Words.
* Expand Acronyms.
* Delete non-English chats.

## *C. Feature Extraction*

Feature extraction is the process of transforming unprocessed data into usable numerical features while preserving the original data set's content. The pre-processed dataset has many distinctive features. Using the feature extraction method, we extract the features from the processed dataset. Eventually, this characteristic is used to calculate a sentence's positive and negative polarity using models like unigram and bigram, which aid in gauging people's opinions.

Another open-source Python module that is available and that can be used to extract similar traits is called Natural Language Tool Kit (NLTK).

Machine learning techniques must be able to communicate their key points in order to process text or documents. These crucial qualities, which are used in the classification process and are known as feature vectors, are as follows:

*1. Words and Their Frequencies:* Unigrams, bigrams, and n-gram models with frequency counts are among the features. More research has been done on the use of word presence as opposed to word frequencies to better explain this quality.

*2.* Language Structure Adjectives, adverbs, and specific verb and noun groupings are effective indications of subjectivity and sentiment. Through parsing or dependency trees, we can produce syntactic dependence patterns.

*3. Opinion Terms and Phrases:* In addition to specific words, idioms, and phrases that express feelings can be employed as features.

*4. Words' Placement:* The positioning of a term within a text can affect how much the term alters the overall tone of the text.

*5. Negation:* Interpreting negation is a crucial but challenging aspect. The presence of a negative usually shifts the polarity of the opinion.

e.g., I am not feeling comfortable.

*6. Syntax:* Several researchers employ syntactic patterns, such as collocations, as features to learn subjective patterns.

## *D. NLP Algorithm Working*

The field of study known as natural language processing, or NLP for short, studies how language and computers interact with one another. It is at the intersection of computer science, artificial intelligence, and computational linguistics.

There are two parts to the algorithm first is the model selection and the second is model evaluation.

### *1. Model Selection:*

The pre-processed data must then be given to a classification model for additional processing. These models are constructed using various classification methods. In the study, the categorization was carried out using the k-nearest neighbor approach.

An algorithm for classifying a set of data into its assigned goal values—in this case, positive, neutral, or negative—is known as the KNN or k-Nearest Neighbor algorithm. Although KNN is frequently used for classification issues, it might also be utilized for regression issues.

Any classification model must now be trained on a target set before being used. Most of the references in the section on the literature review have manually set these target values to positive, negative, or neutral. To set the target for chats automatically, the Python Textblob package is used in the proposed system.

For this, a different Python module called sklearn, which includes a variety of classification models and encoders is utilized. This library is used for model selection, label encoding, and model evaluation, which are covered in the next section.

### *2. Model Evaluation:*

Logistic Regression or Confusion matrices are one of the most popular and useful techniques for classifier evaluation. In the table below, a generalized version of the confusion matrix is provided:

TABLE I. Generalized confusion matrix

|  |  |  |
| --- | --- | --- |
|  | Actually positive  (1) | Actually Negative  (0) |
| Predicted Positive (1) | True Positives  (TP) | False Positives  (FP) |
| Predicted Negative (0) | False Negatives  (FN) | True Negatives  (TN) |

The generalized evaluation parameters can be determined by using this method. These parameters consist of:

*1. Accuracy:* Accuracy is one metric for evaluating classification models. Informally, **accuracy** is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Accuracy can be calculated in terms of positive and negative as:

*2. Precision:* This is the portion of actually translated items relative to the total number of items to be translated, it can be calculated as:

*3. Recall:* This is the portion of successfully retrieved items to the total number of demanded items, and is calculated as:

*4. F-measure:* This is the harmonic mean of recall and precision, and calculated as:

The confusion matrix with advanced classification metrics is given below:

1. Confusion Matrix with advanced classification matrics

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive | Negative | ← Predicted Class |
| Positive | True Positive (TP) | False Negative (FN) | Recall |
| Negative | False Positive (FP) | True Negative (TN) | Specificity |
| Actual Class → | Precision | Negative Predictive value | Accuracy |

# IV. Proposed system

The general model for the sentiment analyzer, the architectural diagram,



Fig. 2. Architecture Diagram

The sentiment analysis architecture for phrases is shown in Figure 2 and is built on the model developed by Kharde and Sonawane [21]. To classify something, a training set of sentences must be generated that are positive, negative, and neutral. It is applied a text steaming and stop words are eliminated in the training set and the input text. After that, the input text's polarity is established.

## *A. Machine Learning Approach Used*

There are two approaches, namely *Unsupervised learning*, and *Supervised Learning* to classify text. In the proposed experimentation model, a supervised learning approach used for classification as the classifier makes simpler future predictions for unknowable data.

It is distinguished by the way it trains computers to accurately classify data or predict outcomes using labeled datasets. The model modifies its weights as input data is fed into it until the model has been properly fitted, which takes place as part of the cross-validation process. A training set is used in supervised learning to instruct models to produce the desired results. This training dataset has both the right inputs and outputs, enabling the model to develop over time. The loss function serves as a gauge for the algorithm's correctness, and iterations are made until the error is sufficiently reduced.

Three techniques are used for supervised learning in the proposed experimentation model are:

*1. TF-IDF Vectorizer*

*2. Support Vector Machines*

*3. Grid Search*

### *1. TF-IDF Vectorizer*

TF-IDF means the Term Frequency Inverse Document Frequency of Records. It is among the most well-liked and successful methods for natural language processing. With this method, you may determine how significant a term is about all other terms in a document.

The Term frequency (tf) with which a keyword appears in a specific document. Hence, it is unique to a document. It can be calculated by,

The Python sklearn library uses the following formula for calculating tf,

Inverse Document Frequency (idf) is a metric used to assess a term's frequency or rarity across the full corpus of documents. The important thing to remember is that it is shared by all of the documents. The idf value (normalized) will be close to 0 if the term is widespread and many documents contain it, or it will be close to 1 if the word is uncommon. It can be calculated by,

Where,

n = Total number of documents available

t = term for which the idf value has to be calculated

df(t) = Number of documents in which the term t appears

The Python sklearn library uses the following formula for calculating idf,

and

The product of a term's tf and idf determines its tf-idf value. The more relevant a term is in that document, the higher its value. TF-IDF Algorithm the following sequence:

1. Calculate each term's TF-values (word).
2. Take the IDF values associated with these phrases.
3. Calculate the TF-IDF values for each term by dividing the TF by the IDF.
4. We obtain a dictionary that includes TF-IDF calculations for each term.

## *2. Support Vector Machines (SVM)*

Classification or regression issues can be solved using the Support Vector Machine (SVM), a supervised learning machine learning technique. Most of the time, it is utilized in classification issues like text classification. Each data point is represented by a point in n-dimensional space (n being the number of features you have) in the SVM method, with the value of each feature being the value of a certain coordinate. The optimal hyper-plane that clearly separates the two classes is then found, and classification is thus achieved.

A hyper-plane is a specific SVM visualization, while Support Vectors are just individual observation coordinates. The SVM classifier is a frontier that best distinguishes between the two classes (hyper-plane/line).

## *3. Grid Search*

A tuning technique called grid search aims to identify the best hyperparameter values. On the specific parameter values of a model, a thorough search is conducted. The model is also known as an estimator.

A machine learning model with hyperparameters has several parameters that were not learned from the training set. The model's accuracy is controlled by these parameters. Before the model is trained, the caller of the model configures and provides the hyperparameters. Consider the Support Vector Classifier (SVC) model as well, which is employed to categorize data sets. The model needs a lot of different hyperparameters. The sk-learn library version of SVC can then be configured with a wide range of hyperparameters, some of which are widely used are:

* This regularisation parameter is *C.*
* The *kernel parameter* has several options, including linear, poly, rbf, sigmoid, precomputed, and callable.
* We can enter a custom *degree* to support the poly kernel parameter.
* *Gamma:* The rbf, poly, and sigmoid kernel parameters are all affected by this coefficient.
* The solver's maximum iteration count is indicated by the variable *max Iter.*

## *B. Data Flow*

The data flow of the proposed experimentation model is shown in Figure 3. WhatsApp chats are the key source of data. In a previous section of the approach, it is detailed how the data is gathered and processed. The dataset is uploaded by the user to the analyzer, where it is subsequently stored. The analyzer then begins processing the chats. The user receives an analysis report of the uploaded dataset after processing and applying the NLP algorithm, covered in an earlier portion of the technique.



Fig. 3. Data Flow Diagram

## *C. Flow of Model*

First, WhatsApp chats from various users are gathered. These conversations contain both sender and receiver texts. The chats of the intended subject are separated and stored separately in the second step because only the sender’s chats need to be analyzed.

Next, the dataset goes through pre-processing where from the chat non required things are omitted such that URLs, media files, and usernames. The spell gets corrected, non-English words are omitted and acronyms are expanded in this stage.

Next, we used a supervised learning approach for the classification of data, we used tf-idf vectorizer, SVM, and grid search, discussed in the previous section.

Following that, the chat file is run through a neutrality classifier. Here, the confusion matrix is calculated, we discussed the confusion matrix in an earlier section of the methodology. In this case, the sentence might be categorized as Positive, Negative, or Neutral. Calculate the Weightage Factor as 0.000-1.000 to lessen the impact of Neutral on our scores. The Neutrality Score in this case is 1.000. Hence, less neutral statements will have a greater impact on the score than less neutral sentences.

Now we get trained data with formatted order. The data with numerical calculations go through the testing phase where the sentiment of a sentence or whole chat is detected. Lastly, the detected sentiment is displayed to the user. The flow of the model is shown in the diagram below.



Fig. 4. Flow Chart

# V. EXPERIMENTAL SETUP

## *A. Hardware and Software Specifications*

Using jupyterbook and Spyder, the model has been fully created in Python 3.9.0 on the Anaconda interpreter. On a computer with the bare minimum requirements of Windows 10 and 4 GB of RAM that supports x86 64-bit architecture, experiments can be run. Python is used to create both the front end and the back end, along with a number of the supporting libraries that were described in earlier sections. Python Tkinter is used to build the front end. The Python binding for the Tk GUI toolkit is Tkinter. It serves as the de facto default GUI for Python and is the official Python interface to the Tk GUI toolkit.

## *B. Data Gathering*

For research purposes, two types of data are used, raw data collected from users and preorganized data from the internet.

Particularly, group chat data rather than private conversation data is gathered for experimentation. Different types of ten groups are selected which are based on different subjects of engineering backgrounds between users of age groups of 18-25 years. The chats are extracted into the textual format without including the media files in the ‘.txt’ format using the ‘Export Chats’ feature of WhatsApp with the consent of concerned users. The group chat’s raw data is shown in Figure 5 after being extracted and saved in ‘.txt’ format.

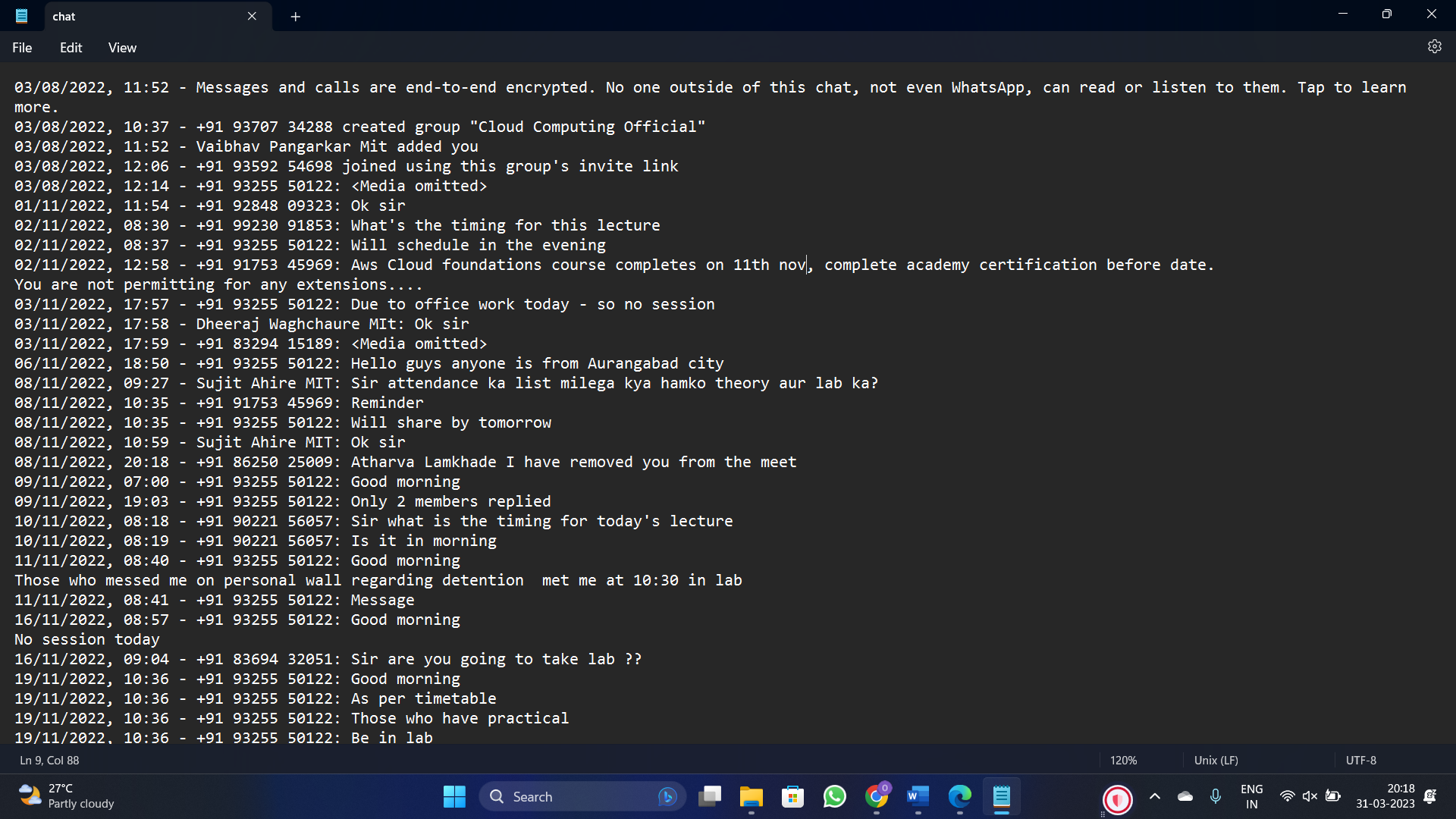


Fig. 5. WhatsApp Chat data in .txt format

The preorganized or open-source dataset from Kaggle [22] is utilized for experimenting. There are 12413 distinct chat statements in the sample. The Kaggle dataset can be shown in Figure 6.

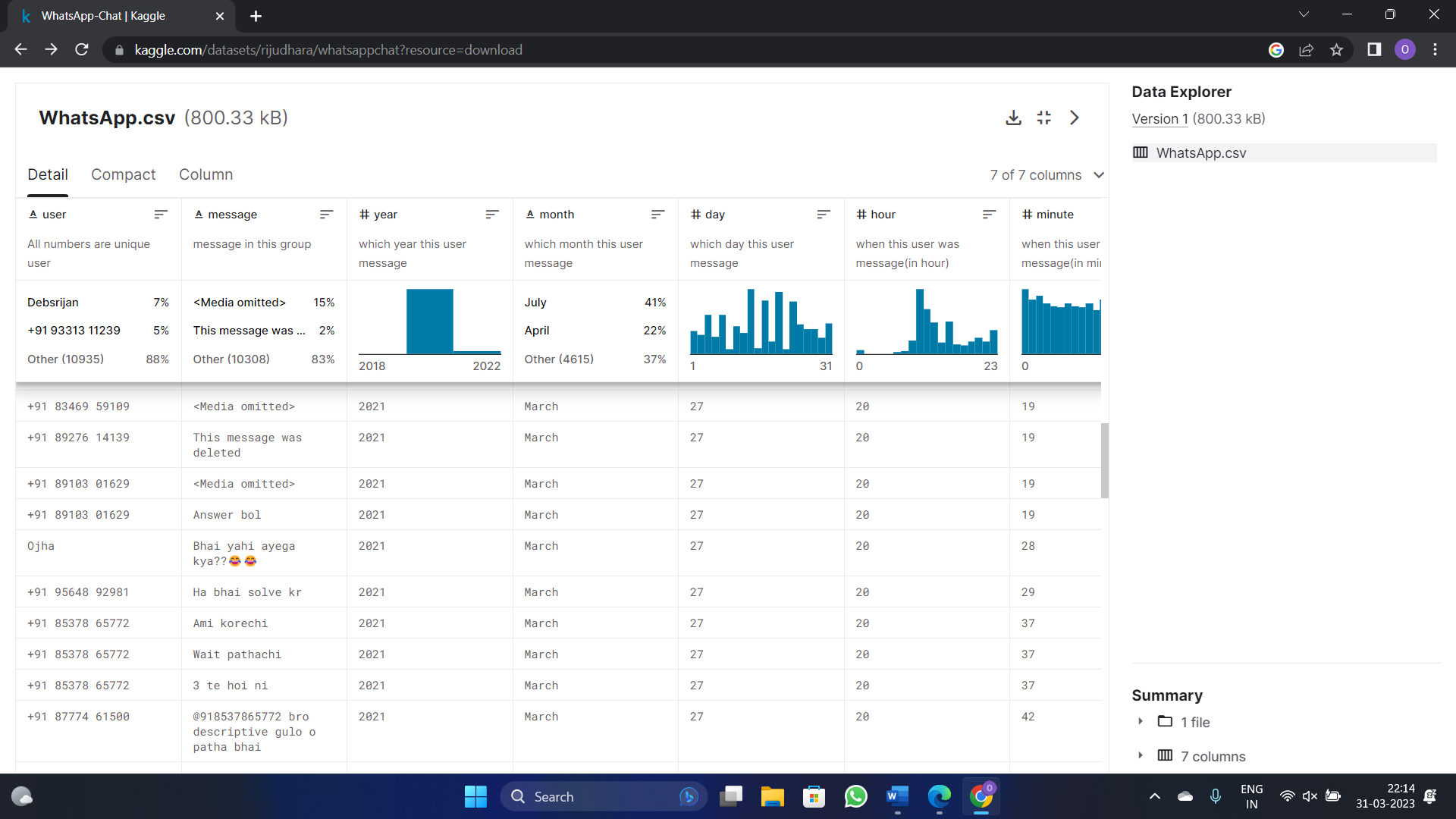


Fig. 6. Kaggle Dataset (preorganized dataset)

## *C. K-Nearest Neighbor*

A classification model must then be provided with the pre-processed data for further processing. These models are built using a variety of classification techniques. The k-nearest neighbor (KNN) method was used in the experimentation to carry out the categorization.

A non-parametric supervised classification approach called k-nearest neighbor (KNN) is straightforward but frequently efficient. Due to its effectiveness in producing efficient results and its simplicity, the KNN classifier is regarded as the most used classifier for pattern recognition. The fields of pattern recognition, machine learning, text classification, data mining, object recognition, and many more all make extensive use of them [23].

KNN was chosen since sentiment analysis is a binary classification and there are large datasets that can be processed.  In the proposed experimentation, the classifier is trained using a manually created training set. Within the training set, there is an X:Ys connection that shows the score of an opinion word as represented by x and the score of the word's positivity or negativity as represented by y [24]. KNN classifier receives as input a score of the opinion term associated with a feature in the chat.

## *D. Validity*

A dataset that had been pre-classified into the three emotions—positive, negative, and neutral—was needed to validate the experimentation model that had been suggested. After the collection of data, data is filtered into the required three emotions i.e., positive, negative, and neutral.

If the emotions corresponding to the top score matched the identified emotion in the data set, then that statement is labeled as successfully classified; if not, it is not. This is done to validate the model.

Several performance analysis criteria, such as precision, recall, and accuracy, which are covered in depth in the methodology section above, are taken into account to examine the performance of the proposed system.

The proposed system's performance is compared to the current system, which classifies chats as positive, negative, or neutral using an SVM classifier.

# VI. RESULT AND ANALYSIS

The testing trials revealed that model has accuracy of 0.6 or 60 percent for K-Nearest Neighbor (KNN) classifier.

Model implemented the logistic regression for training the dataset. The average F1 score obtained for the model trials is 0.0176.

A predictive probability in the logistic regression model is the likelihood that the dependent variable will take on a particular value when one independent variable's value changes while all other independent variables remain constant. The predictive probability obtained for the proposed model is:

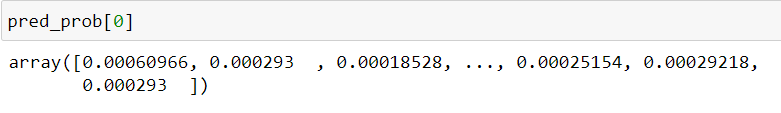


Fig. 7. Predictive Probability

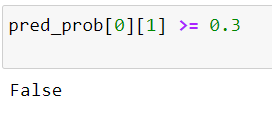


Fig. 8. Predictive Probability

Fig. 8 depicts that the overall model probability is false, i.e., the model incorrectly predicts the positive class.

**Sentiment Analysis:**

The dataset used for analysis is the WhatsApp group chat data publicly made available on the Kaggle web platform. The overall sentiment of the processed group chat was found as ‘Neutral.’

For the understanding of the pattern of positive, negative, and neutral chats, here some chats with the highest positive, highest negative, and neutral polarity chats are picked up:

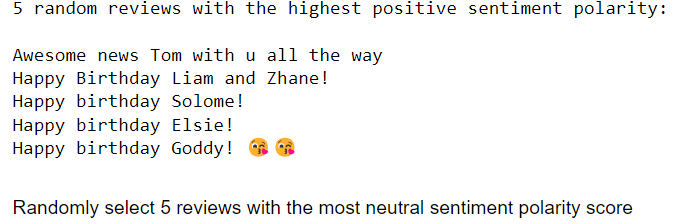


Fig. 9. Positive polarity chats

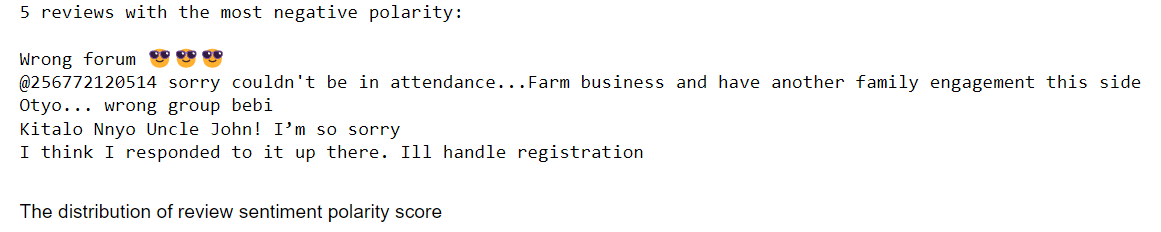


Fig. 10. Negative polarity chats

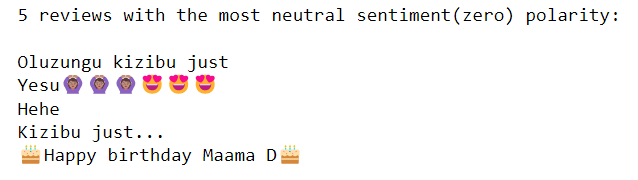


Fig. 11. Neutral polarity chats

Fig. 12 depicts the graphical representation of the polarity scores in the analysis of WhatsApp chats.

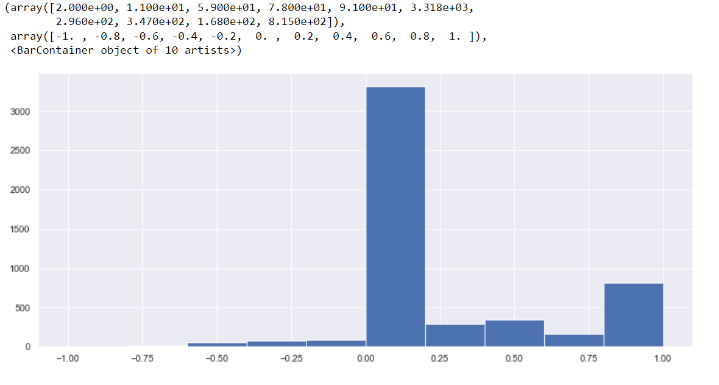


Fig. 12. Distribution of Review sentiment polarity score

**Data Cleaning:**

The practice of removing unnecessary and incorrect values from data that is intended for analysis is known as data cleaning in sentiment analysis. This phase in the sentiment analysis procedure is essential.

The unwanted items are removed from the dataset. The stopwords, URLs, usernames, media files, and emojis are omitted from the data for further analysis. Fig. 13 depicts the raw data after cleaning.

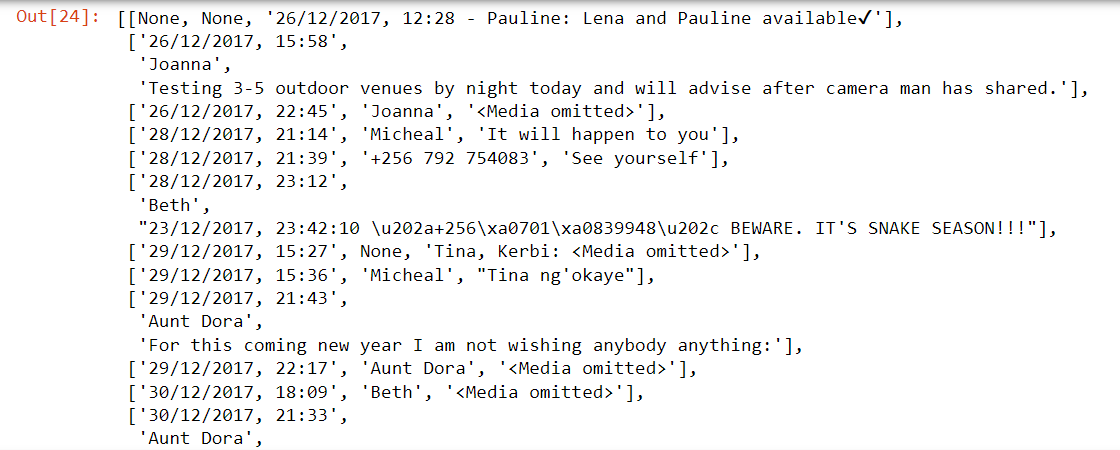


Fig. 13. Raw cleaned dataset

Further, the raw segregated data is converted into tabular form with the required fields. In the testing trials, the raw cleaned data is separated into date, time, author, and message fields of the tabular format.

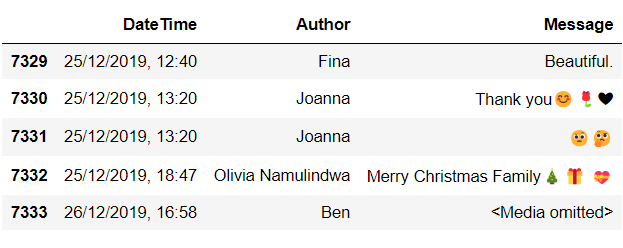


Fig. 14. The tabular data

**Analysis of WhatsApp chat data:**

Data Analysis of WhatsApp data that are going to use for sentiment analysis is important for understanding the behavioral patterns of users.

The count and the frequency of the chats in the data is depicted in Fig. 15.

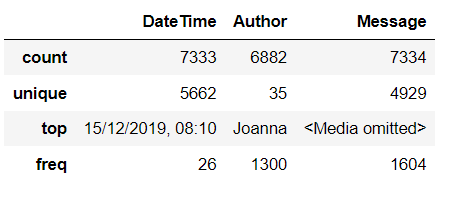


Fig. 15. Count and frequency of chats

The figure 16 presents the data of most media files sent by the user. It is in a graphical way which is easy to understand.

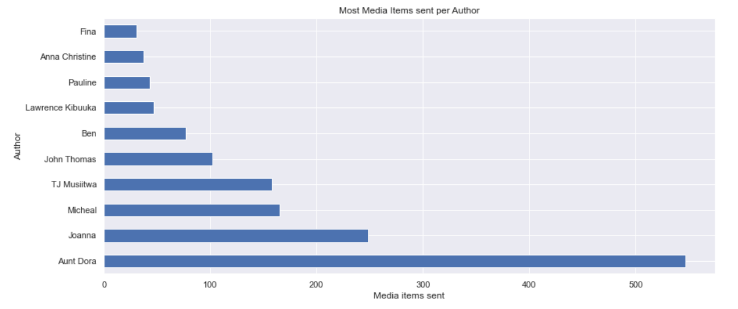


Fig. 16. Most media items sent per author

‘delete for everyone’ is one of the most popular feature of WhatsApp. By this feature we can permanently remove the message from the receiver’s inbox. Fig 17 shows the data about the permanently deleted messages per author.

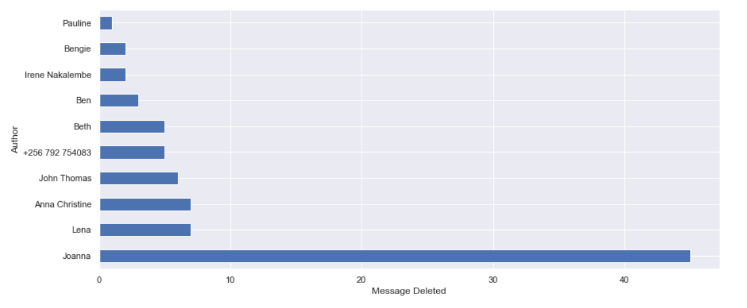


Fig. 17. Most deleted messages per author

Data exploration of the dataset is primarily depicts in the figure 18.

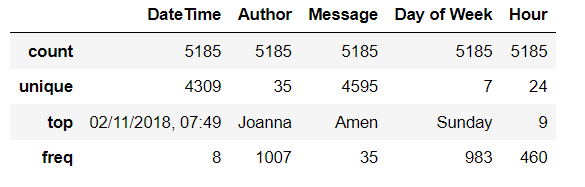


Fig. 18. Data Exploration

The following graphs give information about the word and letter counts in the dataset.

The total word count in the dataset which is the group chat data is 38057.

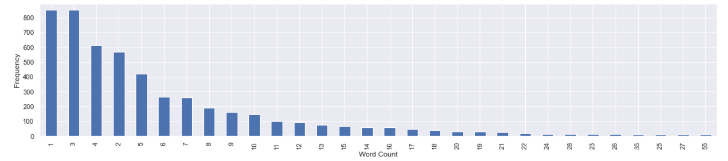


Fig. 19. Frequency of word count

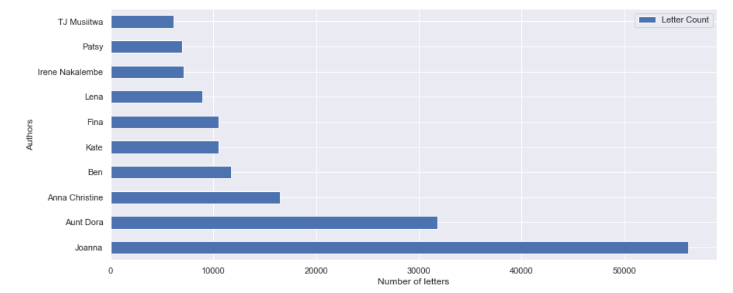


Fig. 20. Number of letter counts

Below three graphical figures depicts the information about the active time periods of users in the group in the WhatsApp chat. This time period is separated into months, days, and hours.

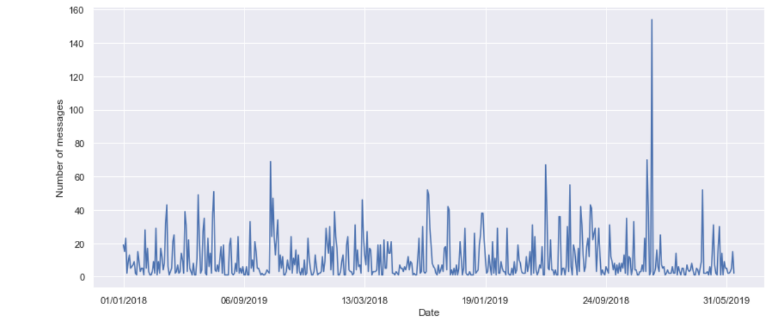


Fig. 21. Number of messages sent (month)

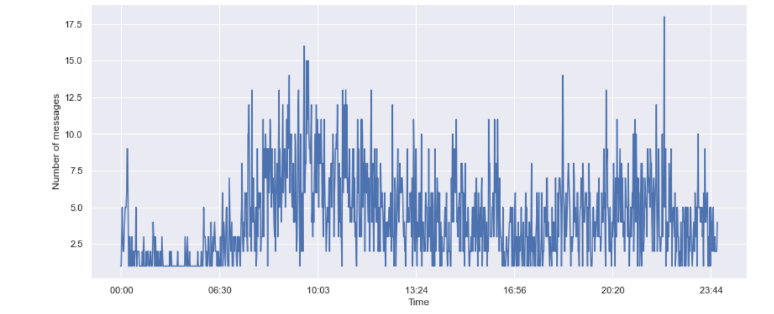


Fig. 22. Number of messages sent (day)

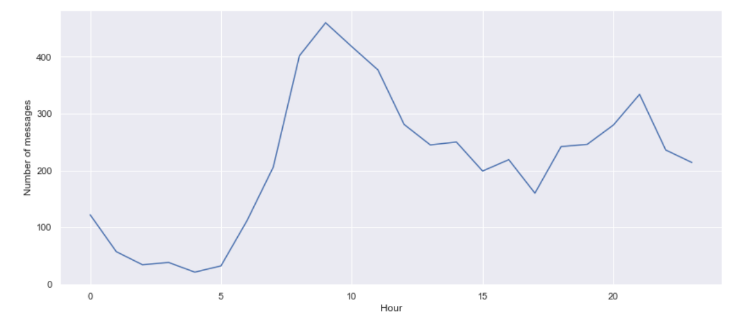


Fig. 23. Number of messages sent (hour)

From the above graphical information, it has been observed that the users are most active from 9 am to 10 am in the daytime and 10 pm to 11 pm in the night-time.

The most talkative users in the group are found in the dataset used for analysis, these are shown in Fig. 24.

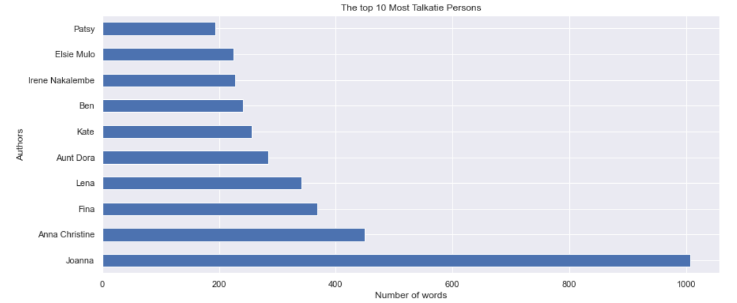


Fig. 24. Most talkative users

The figure 25 represents the most commonly used words in the form of word cloud.



Fig. 25. Most commonly used words

# VII. CONCLUSION

A crucial stage in any machine learning method is data pre-processing. Raw or unstructured data must be transformed into a structured representation prior to any further processing. Code-switching, which can be seen in a lot of the text data acquired from social media, complicates processing and must be removed. Emojis’ interactions with text messages and this code-switching are not taken into consideration by the present approaches. The algorithm described in this study successfully distinguishes the messages of the two participants from a variety of WhatsApp chats. Emojis and sentences are then separated, and the sentences are then translated into a specific language to make them general and united. Using the suggested formula, the chats are further categorized for sentiment analysis. The pre-processing data were then utilized to assess the respondents’ social conduct in context and determine whether the chats’ intentions are positive, negative, or neutral. The sentiment and data analysis methods mentioned in this paper can be applied to data pre-processing. Combining a scientific method for behavioral analysis with useful engineering objectives of assessing emotions in NLP texts can inspire a better and more intelligent strategy for developing systems that interact with people, such as chatbots. Additionally, psychologists can combine these strategies to improve their models for sociological research.

##### References

[1] “WhatsApp Statics 2023 – Usage, Users, Revenue and more”. Available: <https://meetanshi.com/blog/whatsapp-statistics/>

[2] Live Mint (July 2018) “How widespread is WhatsApp’s usage in India?”. Available: <https://www.livemint.com/Technology/>

[3] Bo Pang and Lillian Lee. “Opinion Mining and Sentiment Analysis.” Foundations and Trends in Information Retrieval Vol. 2, Nos. 1-2 (2008).

[4] Kumar, Naveen Sharma, Sudhahsh. (2017). Survey Analysis of “The Usage and Impact of WhatsApp Messenger.” Global Journal of Enterprise Information System.

[5] Bhatt, Anshu Arshad, Mohd. (2016). “Impact of WhatsApp on Youth: A Sociological Study.” IRA-International Journal of Management & Social Sciences (ISSN 2455-2267).

[6] Tian, Ye Galery, Thiago Dulcinati, Giulio Molimpakis, Emilia Sun, Chao. (2017). “Facebook sentiment: Reactions and Emojis.”

[7] Sonika Dahiya, Astha Mohta, Atishay Jain, “Text Classification based Behavioural Analysis of WhatsApp Chats”, Preceding of Fifth International Conference on Communication Electronics System (ICCES), 2019.

[8] Astha Mohta, Atishay Jain, Aditi Saluja, Sonika Dahiya, “Pre-Processing and Emoji Classification of WhatsApp Chats for Sentiment Analysis”, Proceedings of Fourth International Conference on I-SMAC, 2020.

[9] Seydeh Akram Saadat Neshan, Reza Akbari, “A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis”, 6th International Conference on Web Research (ICWR), April 2020.

[10] Barman, Utsab Das, Amitava Wagner, Joachim Foster, Jennifer. (2014). “Code Mixing: A Challenge for Language Identification in the Language of social media.”

[11]M Tafaquh Fiddin Al Islami, Ali Ridho Barakbah, Tri Harsono, “Interactive Applied Graph Chatbot with Semantic Recognition”, Published by IEEE, October 2020

[12] Kootbodien, Ammaarah Prasad, Nunna Ali, Muhamad. (2018). “Trends and Impact of WhatsApp as a Mode of Communication among Abu Dhabi Students.”

[13] Lu, Xuan Ai, Wei Liu, Xuanzhe Li, Qian & Wang, Ning Huang, Gang Mei, Qiaozhu. (2016). “Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users.”

[14] Xiao zhuhanling, “An Emotional Topic Recommendation Chat Assistant”, IEEE 15th International Symposium on Applied Computational Intelligence and Informatics, May 2021

[15] Krebs, Florian Lubascher, Bruno Moers, Tobias Schaap, Pieter Spanakis, Gerasimos. (2017). “Social Emotion Mining Techniques for Facebook Posts Reaction Prediction.”

[16] Hussien, Wegdan Tashtoush, Yahya Al-Ayyoub, Mahmoud Al Kabi, Mohammed. (2016). “Are Emoticons Good Enough to Train Emotion Classifiers of Arabic Tweets?”

[17] K. Norambuena, B. E. F. Lettura, and C. M. Villegas. "Sentiment analysis and opinion mining applied to scientific paper reviews." Intelligent data analysis 23, no. 1 (2019): 191-214.

[18] R. Vallikannu, and T. Meyyappan. "Twitter text mining for sentiment analysis on people’s feedback about Oman tourism." In 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), pp. 1- 5. IEEE, 2019.

[19] S. J. Rajkumar, V. S. Shirsat, and S. N. Deshmukh. "Sentiment analysis on product reviews using machine learning techniques." In Cognitive Informatics and Soft Computing, pp. 639-647. Springer, Singapore, 2019.

[20] M. Mika V., D. Graziotin, and M. Kuutila. "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers." Computer Science Review 27 (2018): 16-32

[21] V.A. Kharde and S.S. Sonawane. “Sentiment Analysis of Twitter Data: A Survey of Techniques.” International Journal of Computer Applications 139 (2016): 5-15

[22] “WhatsApp – Chat” (2022). Available: <https://www.kaggle.com/datasets/rijudhara/whatsappchat>

[23] Soudamini Hota and Sudhir Pathak, “KNN classifier-based approach for multi-class sentiment analysis of Twitter data.” International Journal of Engineering and Technology 7(3) (July 2018): 1372-1375.

[24] Akashdeep Dhiman and Dinesh Kumar, “Sentiment Analysis Approach based N-gram and KNN Classifier.” International Journal of Computer Applications (0975 – 8887), July 2018.