

## **Analysis of Online Social Behavior of Whatsapp Users**

### **Sonika Dahiya**

Department of Computer Science, Delhi Technological University, Delhi, India.

Email: Sonika.dahiya11@gmail.com

### **Aditi Saluja**

Department of Software Engineering, Delhi Technological University, Delhi, India.

Email: saluja.aditi5@gmail.com

### **Parteek Singhal**

Department of Software Engineering, Delhi Technological University, Delhi, India.

Email: psinghal767@gmail.com

### **Rahul Johari**

USICT, GGS Indraprastha University, Delhi, India.

Email: rahuljohari@hotmail.com

*Received August 05, 2021; Accepted November 20, 2021*

*ISSN: 1735-188X*

*DOI: 10.14704/WEB/V19I1/WEB19018*

---

### **Abstract**

Whatsapp is the most popular social media platform that allows users to interact with each other by sending texts, emoji, images, voice notes, videos etc. Whatsapp users spend hours chatting with each other exchanging information and also sharing intimate feelings that are often touted as deep meaningful conversations. This paper focuses on the study of online social behavior of Whatsapp users which are primarily students in the age group of 18-25 years. 300Whatsapp chats from 30 university students have been collected and scrutinized to determine their online behavior in terms of being expressive on Whatsapp or not and also identifying their general mood on Whatsapp based on their emotional manifestation in the chats. This analysis can be of compelling interest to various psychological researchers, social media influencers and counselors.

### **Keywords**

Whatsapp, Online Social behavior, Emoji, Emotion analysis, Clustering, Weighted K-means.

### **Introduction**

According to (Ilavarasan et al., 2020) in recent years, the online social networks have exponential growth and the field is booming as it aids in the promotion of brands, sharing

of information, staying connected with a circle of a network that consists of family and friends. Social networking has something or other to offer people of all age groups, and using them has almost become a part of our daily activities. This has attracted people around the world to show active participation in social media. Social Media offers a plethora of online networking sites allowing people to connect with each other and communicate their ideas and opinions to other people. According to a Pew Research Center report (Brenner, 2013), 73% of adults in America use social networking compared to just 15% seven years ago. The report also reveals that 40% of cell phone users have apps on their mobile devices that provide access to these social networking sites (Kio, 2015). Amongst the most popular networking sites such as Whatsapp, Facebook, and Twitter, it has been found that Whatsapp is preferred and used more often with a rank of 93.4%, which is followed by Twitter 76.4% and Facebook at 36.8%. In terms of time spent, 27.5% users spend more than three hours surfing on Whatsapp, 18.9% for Twitter and 5.5% for Facebook as stated in the results of the study conducted by (Alsanie, 2015). Hence, it can be stated that Whatsapp has major hold on online communication and has been successful in attracting over-enthusiastic user involvement which is growing enormously every day. Whatsapp supports free text messaging, various kinds of Emoji and emoticons, images, videos, stickers, GIFs (Graphic Interface Formats) and even voice notes. Thus, facilitates user to exchange information conveniently and express their emotions while chatting. Though there is no specific target audience for Whatsapp, it is a common observation that younger generation is infatuated and addicted to the application.

According to (Hussain et al., 2017) most university students use Whatsapp on a daily basis and check their account more than four times a day. (Minhas et al., 2016) stated that university students use Whatsapp for diverse reasons like exchanging academic details, chatting, communication with friends and family, sharing pictures and videos and even discuss current affairs. However, majority of respondents in his research used Whatsapp for general purpose chatting only to present their ideas and express their feelings. So, it can be clearly seen that younger generation uses Whatsapp as a medium to share their sentiments and thoughts. The online texting mediums have changed over years beginning from email then shifting to SMS and finally instant messaging services like Whatsapp. Along with the change in mode of communication, significant change can also be observed in chatting styles and patterns. Millennial use lots of Emoji or emoticons that reflects their mood in the chat along with the conventional text message. The wide variety of emoticons on Whatsapp conveys the whole gamut of human emotions enabling users to put across their feelings in fewer words. According to (Alsanie, 2015), the virtual personality of a person may differ from his regular personality. The virtual personality depicts the online social behavior of a

person and depicts his dealings with his peers in an online world. It is possible to deduce interesting behavioral characteristics of users from their message patterns and emoticons used in their chats.

Knowing the behavior of users when they operate in a social network has attracted and attracts the attention of the scientific community. Indeed, better understanding user behavior is a key issue in several contexts; users themselves to enhance awareness in this potentially insecure world; companies and government institutions to make better use of this huge network of people for their finalities; scientists to better understand individuals and communities as stated by (Borruto, 2015). Various kinds of studies and research have already been conducted to understand user's sentiments on social media and blogging sites. Understanding user behavior on these virtual platforms has been crucial for discovering user interests and enhancing marketing strategies taking into account user's opinion. (Cani et al, 2015) carried out a survey that investigates the use of social networking sites in relation to friendships in a large sample of Italian adolescents. (Young, 2013) explored the impact of social networking sites on making new friends online and maintaining existing offline friendships. (Zhao et al, 2016) built a social sentiment sensor for detecting trending topics on micro blogging sites and examining public opinion towards hot topics. (Sarlan et al, 2014) performed twitter sentiment analysis using machine learning techniques coupled with natural language processing methods to analyze customer's review of a product or a service as positive, negative or neutral. Many more improvements have come to sentiment analysis techniques using different kinds of twitter data encapsulating varied opinions of users on different subjects, for instance figurative sentiment analysis of Twitter data carried out by (Nguyen et al., 2017). Similar techniques have been applied on yet another popular networking site which we all love scrolling over during our free time i.e. Facebook. (Zamani et al., 2014) conducted a study to identify emotion patterns of users of Facebook using a lexicon based approach by analyzing comments on a randomly selected post on a particular Facebook page of a university. Based on the comments given by the users, their emotions are categorized as happy (positive), unhappy (negative) and emotionless. This in turn determined the type of response (positive, negative or neutral) yielded to a Facebook post by the users.

In the current research work, a model has been proposed to determine the online social behavior of Whatsapp users. University students in the age group 18-25 have been approached and convinced by sincere efforts to share their Whatsapp chats for the research. These chats are analyzed, using count of words and count of Emoji, and six features derived that are used to guide behavioral analysis. We used unsupervised learning to learn patterns from Whatsapp chats of different users and classify them as Extrovert or Introvert. Extrovert

as online social behavior represents being highly expressive on Whatsapp in terms of sending texts and Emoji while Introvert represents less expressive on Whatsapp and does not like to send a lot of messages. Further, emoji usage pattern is used to identify the general mood of users on Whatsapp as happy, playful, sad, angry or lovable.

### **Related Work**

Whatsapp is different with respect to Twitter and Facebook as it provides a more private form of communication. Generally, people use Twitter to voice their opinion on social issues or opinions on events, activities, and policies etc. affecting masses. Facebook is used more by people to show off fancy details of their life. For communicating with friends and relatives, most people use Whatsapp. People actually talk about their daily life events and happenings with loved ones on Whatsapp. Analyzing Whatsapp conversations of varied people can give us an insight into their actual lives and help us understand their behavior while chatting. (Kumar et al., 2016) did a ‘Survey Analysis on the usage and impact of Whatsapp Messenger’ to estimate the growing effect of Whatsapp on the people’s life, culture and society at large. The most striking results of the survey concluded that reason for using Whatsapp for a huge number of people was to communicate instantly anywhere anytime and express their emotions without seeing each other and 66% participants believed that Whatsapp enhanced their relationships with friends and family. While 64% people believed that Whatsapp was not harmful for their health, there were 19% people who stated otherwise and thought that Whatsapp was harmful for their health. It was rather interesting to find that despite realizing Whatsapp can be harmful for their health, people still use it as it gave them a feeling of belongingness and they felt valued when they got instant reply to their messages. This motivated us to deeply inspect user’s Whatsapp chats and find their online personality and emotional triggers (Johari et al., 2021).

(Premalatha et al., 2016) did an impressive research on ‘Emotion Analysis of Whatsapp Groups using Big Data Analytics’. Using text mining and analytical tools, they examined emotions like anger, fear, disgust, anticipation, joy, sadness, surprise and trust in the group chat collected over a period of time. This made us realize that there are innumerable human emotions and one cannot possibly detect all the variegated sentiments hidden in chats. Therefore, in the current research work, an attempt at putting users in certain broad categories based on most common emotions like happiness, playfulness/joy, sorrow, anger, love has been done and all the other emotions are not neglected but put in a miscellaneous category during chat analysis. (Bai et al, 2019) gave an interesting an review of application and usage of emoji in different contexts in A Systematic Review of Emoji: Current Research and Future Perspectives.

One of the important observations in The Study of Whatsapp Usage Patterns and Prediction Models without Message Content by (Rosenfeld et al., 2018) revealed that 99% of messages on Whatsapp are only text messages and only 1% include media and file sharing. Using this result, current research work focuses only on the text messages and the Emoji enclosed within the text messages to determine user behavior on Whatsapp. Another result highlights the differences in usage patterns across different age groups and gender. However, parameters like Social Level, Usage Level, Average messages in a day etc. for both male and female under the age of 25 years is nearly same and highly vary for the age group above 25 years. This formed another hypothesis in current work as our dataset comes from university students who are highly active on social media. Although, Gender is a significant attribute in analyzing online social behavior but both male and female in this age group are equally involved on social media and this feature is thus neglected during learning phase.

Various different techniques exist for applying social media analytics to mine special patterns (Kalra et al, 2021). (Singh et al, 2019) used big data analytics for their analysis of social media. (Sharma et al, 2018) gave a thorough analysis of sentiment analysis techniques that can be employed to mine meaningful patterns from social media data. In our analysis, we have employed a Machine learning method Weighted K means clustering to identify clusters of users possessing similar characteristics. K-means clustering algorithm has been applied to find and group people that exhibit similar online social behavior by analyzing their chatting patterns. K-means clustering algorithm is an effective and scalable clustering algorithm. It is a type of unsupervised learning which tries to discern groups which have not been explicitly labeled in the dataset, K being the number of groups and designates each data point to one of the these groups. According to (Sonagara et al. ,2014)the algorithm searches out the most effective division of n data points in k groups, in order to find the total distance of the group's members to its corresponding centroid, representative of the cluster, is reduced. Formally, the goal of the algorithm is to partition the n data points into k sets  $S_i$  where,  $i=1, 2 \dots k$  so that the within-cluster sum of squares (WCSS) is minimized, defined as

$$\sum_{j=1}^k \sum_{i=1}^n \| X_i^j - C_j \|^2 \quad (1)$$

Where, term  $\| X_i^j - C_j \|^2$  provides the distance between a data point and the cluster's centroid.

K-means is a widely accepted algorithm to dissect clustering problems due to its simple and easy implementation and its ability to perform well even with large datasets. One major problem in k-means algorithm is that it assigns equal significance to all variables in concluding the cluster memberships of objects. This feature of the algorithm makes it less

resourceful for many applications where data contains many divergent variables. To overcome this problem an extension of k-means is proposed which is known as Weighted K-means algorithm. This algorithm uses variables with weights designated to them according to their influence on clustering results. Since Weighted k-means is an extension to k-means it doesn't affect the scalability of k-means in clustering large datasets and gives more definite results due to the weights assigned to the variables. Weighted K-means shall be used in scenarios where some features have more importance over other feature.

Many different interesting applications of social media analysis have been discovered. (Aramburu et al, 2020) proposed a dynamic multidimensional model to identify different user profiles and deduce apposite events that are germane for Public Health Surveillance (PHS) in their paper 'Social Media Multidimensional Analysis for Intelligent Health Surveillance'. Another practical use-case for social media analysis has been expatiated by (Kankanamge et al, 2020) in their paper 'Determining disaster severity through social media analysis: Testing the methodology with South East Queensland Flood tweets'. Another cognitive implication of social media analysis in a different age group than just young people has been explored by (Khoo & Yang, 2020) in their paper 'Social media use improves executive functions in middle-aged and older adults: A structural equation modelling analysis'. Many pragmatic use case of the online behavioural analysis of users carried out in our paper can be considered. One such example can be curating a dating app based on the user's interests by analysing their online personality and matching them to the online personality of an apposite person.

## Proposed Approach

In this section, we are explaining the proposed model to determine the online social behavior of users by analyzing their Whatsapp conversations. The proposed model is diagrammatically represented in Fig 1.

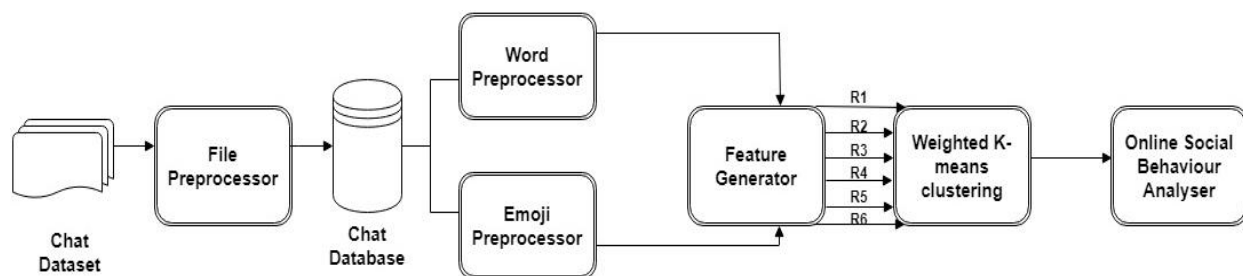


Figure 1 Proposed Architecture

## **Data Collection**

Whatsapp has Explore chat” feature that allows you to share your chat using applications such as Gmail, Bluetooth, Shareit etc. This feature was used as the tool for data collection. 300 Whatsapp chats of thirty participants were collected and stored in a database. For each participant, general demographic information including his name, age, gender and profession was also collected. University students in the age group of 18-25 are interminable users of social media and early adopters of advanced technological applications. Therefore, the targeted participants for our analysis are University students who are also our acquaintances.

Compared with regular text documents, Whatsapp messages can be informal containing slang and other millennial lingo. Any algorithm applied on such raw messages may result in very poor performance. Therefore, preprocessing techniques on raw messages are imperative for improving experimental results which are discussed in next sections.

## **File Pre-Processor**

The dataset collected using Whatsapp ‘Explore Chat’ feature contains user’s messages and related information in a text format. This chat file is fed as input to the File Preprocessor.

The file pre-processor performs the task of Data Cleaning and Data Integration i.e. organization of data in the form required for future work. It organizes user chats into meaningful information and stores them in a database. The dataset obtained contains the user’s name, the text messages sent by the user and the time at which messages were sent. The file preprocessor replaces the name of the user with pseudonyms to protect the user’s privacy and stores the corresponding messages sent by the user in the database.

## **Word Processor**

The text messages retrieved from the database formed after file preprocessing is used for further analysis. However, the chat database obtained is in an irregular format due to the presence of punctuation marks, Emoji and condensed form of words in text messages. Thus, preprocessing of this database is essential to remove the opacities in order to enhance the quality of the database. For this, the following steps are performed:-

1. Tokenization: The text messages are segmented into linguistic units such as words, punctuations, Emoji, numbers and alpha numeric symbols to segregate words and emoji from other tokens.

2. Stop word removal: Connecting words in a statement such as “a”, “an”, “the” etc. which are also known as non-information bearing words were eliminated. As enumeration of these words, doesn’t contribute any significance to determine the expressive nature of a person.
3. Enumeration: After stop word removal, the words remained in the text messages are known as informative words. Count of these words is taken as an indication of a person’s online behaviour in terms of expression. Thus, the informative words are counted and this count is termed as word count.

Word processing stage was exceedingly influential as any error committed at this stage could have caused a decline in efficiency and accuracy of results.







### **Emoji Processor**

Whatsapp is a host to a wide variety of Emoji that enhances user experience of online communication. Emoji depict user emotions in the most pertinent manner. It is symbolic representation of user’s emotions and demonstrates them in an exuberant demeanor.

Table 1. Shows the categorization of mostly used Whatsapp Emoji (mainly emoticons) in six groups as <Happy>, <Love>, <Sad>, <Playful>, <Angry> and <Miscellaneous>. The <Happy> and <Playful> categories comprises of a group of emoticons that project joy, excitement or exuberance. The <Love> category contains emoji that show love, affection or gratitude. The <Sad> category emoji represent disappointment, struggle or suffering. The <angry> category contains emoji that represent displeasure or annoyance. The <Miscellaneous> category incorporates other most popular and widely used Emoji/emoticons on Whatsapp. These categories are employed to interpret the mood of the conversation or to decipher the conversation polarity. Conversation polarity determines the state of mind of users or the pervading tone of conversation. Whatsapp conversation of different users is fed into the processor for examination of counts of Emoji belonging to varied Emoji set used by different users. The processor used a data structure called Hash map which stores <Key, Value> pairs. The key represents the category of Emoji and the value stores the combined count of all the Emoji present in the set of that category for a particular user. Therefore, the Emoji processor computes count of different categories of Emoji which is further used as a parameter for feature generation of users. The use of Emoji has transformed online texting entirely and made it easier for the user to show his feelings. The emoticons in the form of smileys and hearts convey a strong emotion and it is more pleasing than plain words. The Emoji count has been therefore exploited to determine user’s characteristics as measure of how expressive a person is and the count of different categories of frequently used Emoji to determine a person’s usual mood while chatting.



**Table 1 Categories of Commonly Used Whatsapp Emoji for Analysis**

Emoji Category	Emoji Set
Happy	
Playful	
Love	
Sad	
Angry	
Miscellaneous	

Considering the six category of Emoji discussed earlier in this section and informative words discussed in section 3.3, all possible labels are listed in Table 2.

L1 label is assigned to a user that has a leading role in a conversation and is very expressive, thus has a high input in terms of word and Emoji shared in communication. L2 is assigned to a user that is impassive and doesn't use too many graphics for expression. L3 is assigned to a user that extensively uses Emoji for communication. L4 label is assigned to a user who is quite expressive and is usually in a cheerful mood while chatting. L5 is assigned to a user who is also mostly in a joyous and happy mood while chatting but is comparatively less expressive. L6 is assigned to a user who is quite expressive and is usually pessimistic or upset while chatting. L7 is assigned to a user who is also usually sorrowful or upset but is comparatively less expressive. L8 is assigned to a user who is very affectionate and cordial and is quite expressive. L9 is assigned to a user who is also very affectionate but is comparatively less expressive. L10 is assigned to a user who is usually in an enraged mood and is very expressive about it. L11 is assigned to a user who is also in a furious mood while chatting but is comparatively less expressive. Mathematical Formulation of these labels has been discussed in next section.

**Table 2 List of user's Online Behavioral Characteristic Labels**

Label Name	Label Description
L1	Extrovert
L2	Introvert
L3	Emoji Maniac
L4	Happy/Playful + Expressive
L5	Mostly Happy/Playful Mood but Less expressive
L6	Mostly Sad Mood + Expressive
L7	Mostly Sad Mood but less expressive overall
L8	Mostly Romantic Mood + Expressive
L9	Mostly Romantic Mood but less expressive overall
L10	Mostly Angry Mood + Expressive
L11	Mostly Angry Mood but less expressive overall

## **Feature Generation**

After word processor and Emoji processor stages, the detailed enumeration of words and Emoji of respective categories is obtained. These values are employed to determine six features which individuate the participants. Six features are named as R1, R2, R3, R4, R5 and R6. The definitions, mathematical formulation and significance of the features are given as follow:-

Consider User1 is the person under consideration and User2 is the other person interacting with User1.

**Feature 1:** R1 is a ratio of summation of total number of words and Emoji (including all categories) sent by the User1 to total number of words and Emoji sent in the entire conversation, as shown in equation(1).

$$R1 = me\_Total/Total \quad (2)$$

Where me\_Total represents User1's word and emoji count in the chat and Total represents the summation of combined word count and Emoji count of both User1 and User2 in the chat.

**Feature 2:** R2 is a ratio of total number of Emoji sent by User1 to total number of words and Emoji sent by User1 in the conversation.

$$R2 = me\_TE/me\_Total \quad (3)$$

Where me\_TE represents User1's Emoji count in the chat while me\_Total represents summation of User1's word count and Emoji count in the chat.

**Feature 3:** R3 is a ratio of count of Emoji falling under <Happy> and <Playful> categories of User1 to the total number of Emoji used by User1 in chat.

$$R3 = me\_H\&P\_E/me\_TE \quad (4)$$

Where me\_H&P\_E represents User1's count of Emoji in chat belonging to happy and playful category and me\_TE represents User1's total Emoji count in the chat.

**Feature 4:** R4 is a ratio of count of Emoji falling under <Love> category of User1 to the total number of Emoji used by User1 in chat.

$$R4 = me\_L\_E/me\_TE \quad (5)$$

Where  $me\_L\_E$  represents User1's count of Emoji in chat showing love, romantic or flirtatious behavior while  $me\_TE$  represents User1's total Emoji count in the chat.

**Feature 5:** R5 is a ratio of count of Emoji falling under <Angry> category of User1 to the total number of Emoji used by User1 in chat.

$$R5 = me\_A\_E/me\_TE \quad (6)$$

Where  $me\_A\_E$  represents User1's count of Emoji in chat displaying anger and  $me\_TE$  represents User1's total Emoji count in the chat.

**Feature 6:** R6 is a ratio of count of Emoji falling under <Sad> category of User1 to the total number of Emoji used by User1 in chat.

$$R6 = me\_S\_E/me\_TE \quad (7)$$

Where  $me\_S\_E$  represents User1's count of Emoji in chat that depicts sorrow and despair whereas  $me\_TE$  represents User1's total Emoji count in the chat.

The online social behavior of a person varies with different people to a substantial extent. Thus, taking ratio features over a number of participating conversations is more helpful to get more factual and authentic values of the participant's behavior. For instance, if we collected one chat from each person and based our results on that one chat, the results obtained would be skew. This is due to the fact that the way we chat with anyone online is parallel to our relationship with that person in real life. The tone of conversation of a student with his professor would be formal and polite, while the tone with his friend would be friendly signifying a comfort level, and the tone of conversation with his lover would be emotional. Thus, count of words and Emoji used by a person in X number of chats factored by total count of words and Emoji in X chats with X number of people can be a significant attribute that guides his online social behavior.

The labels defined in Table 2 are assigned to a user by keeping in mind these six features R1, R2, R3, R4, R5 and R6. Analysis of these features help in assigning label to the person. If the ratio  $R1 \gg 0.5$  for a person, he can be classified as Extrovert (L1) as he plays the prominent role while conversing with other person on Whatsapp. If  $R1 \ll 0.5$  and all others ratios are approaching to zero, the person can be classified as an online Introvert (L2) i.e. he doesn't like to express through social media. If amongst all the ratios, R2 is most prominent and  $R2 \gg 0.5$ , the person is assigned as Emoji maniac (L3) i.e. he prefers to use graphics, smileys, symbols more often than words in a chat. If R1 and R3 are prominent for a person i.e.  $R1 \geq 0.5$  and  $R3 \geq 0.5$ , it implies that person is expressive and is generally in

a cheerful mood(L4). However, if  $R1 < 0.5$  and  $R3 \geq 0.5$ , it implies that person is not very expressive as such but since, he uses happy and playful emoji in excess, his tone of conversation is always happy and gleeful(L5). If  $R1 \geq 0.5$  and  $R6 \sim 0.5$ , it implies person is expressive of his thoughts and is mostly in a gloomy state of mind(L6). Likewise, if  $R1 < 0.5$  and  $R6 \sim 0.5$ , it implies person is less expressive and is usually sulking on social media (L7). If  $R1 \geq 0.5$  and  $R4 \sim 0.5$ , it shows person is expressive and his tone of conversation depicts romance and attachment (L8) while if  $R1 < 0.5$  and  $R4 \sim 0.5$ , it shows less expressive overall, but uses love, kiss and heart Emoji showing his general behavior is affectionate and full of warmth (L9). If  $R1 \geq 0.5$  and  $R5 \sim 0.5$ , it depicts that the person is expressive on Whatsapp and commonly in a grumpy mood (L10). If  $R1 < 0.5$  and  $R5 \sim 0.5$ , it depicts that the person is less vocal of his thoughts overall and uses emoji that show rage and anger suggesting that person's mood is mostly annoyed and displeased (L11).

## **Clustering**

Weighted k-means clustering algorithm is based on the virtue that not all features have equivalent contribution in predicting the clusters and hence different weights should be assigned to each feature depending on their prominence. In our algorithm we associated R1 with a weight of 0.3, R2 with a weight of 0.3 and R3, R4, R5, R6 collectively a weight of 0.4. i.e. individually with a weight of 0.1 each. R1 and R2 is assigned more weight i.e. 0.3 each because they account for a person's word count and emoji count contribution in a conversation and thus their values majorly define the behavioral characteristics of a person. However, R3, R4, R5 and R6 are assigned a weight of 0.1 each as these features indicate the person's general emotional mood while chatting. Experiments with different values of K are done to find an optimal value of K and get our desired clustering resolution. As discussed in section 3 of results, the resultant K-clusters will act as a trained model that will help in predicting online social behavior of new participants who is a student and lie in the age group 18-25. In the next section of proposed approach, algorithm formulated to conduct this analysis has been discussed in detail. The key notations and Whatsapp pre-processing of chats with respect to different users has been delineated alongside the feature scores and clustering algorithm to obtain the final results (Sonika et al., 2021).

## **Algorithm Formulated**

### **Notations:**

P: Person to be analyzed

WAP<sub>i</sub>: i<sup>th</sup> WhatsApp chat of P

n: number of WAP<sub>i</sub> of P

$em_{ij}$ :  $j^{\text{th}}$  emoji of  $WAP_i$   
 $w_{ij}$ :  $j^{\text{th}}$  word of  $WAP_i$

C1: set of Happy and Playful emoji  
C2: set of love emoji  
C3: set of Sad emoji  
C4: set of Angry emoji  
C5: set of all the remaining emoji

H&P\_E: frequency of  $em_{ij}$  belonging to C1  
 $me_{H\&P\_E}$ : frequency of  $em_{ij}$  belonging to C1 and sent by P

L\_E: frequency of  $em_{ij}$  belonging to C2  
 $me_{L\_E}$ : frequency of  $em_{ij}$  belonging to C2 and sent by P

S\_E: frequency of  $em_{ij}$  belonging to C3  
 $me_{S\_E}$ : frequency of  $em_{ij}$  belonging to C3 and sent by P

A\_E: frequency of  $em_{ij}$  belonging to C4  
 $me_{A\_E}$ : frequency of  $em_{ij}$  belonging to C4 and sent by P

M\_E: frequency of  $em_{ij}$  belonging to C5  
 $me_{M\_E}$ : frequency of  $em_{ij}$  belonging to C5 and sent by P

Total\_W: Total count of words in  $WAP_i \forall i$   
 $me_{TW}$ : Count of words sent by P in  $WAP_i \forall i$   
Total\_E: Total count of emoji in  $WAP_i \forall i$   
 $me_{TE}$ : Count of Emoji sent by P in  $WAP_i \forall i$

Step 1. SET H&P\_E=0, L\_E=0, S\_E=0, A\_E=0, M\_E=0,  $me_{H\&P\_E}$ =0,  $me_{L\_E}$ =0,  
 $me_{S\_E}$ =0,  $me_{A\_E}$ =0,  $me_{M\_E}$ =0  $\forall i$ .

Step 2. Collect in  $WAP_i$ .

Step 3. FOR  $i=1$  to  $n$

3.1. Start scanning from first letter of  $WAP_i$ .

3.2. For  $\forall w_{ij}$  in  $WAP_i$

IF( $w_{ij} \in \text{Stop\_Words}$ )

Continue;

ELSE

```
Total_W ← Total_W + 1  
END IF  
IF (send(P))  
    Me_TW ← Me_TW + 1  
ENDIF
```

3.3. For  $\forall em_{ij}$  in  $WAP_i$

```
IF( $em_{ij} \in C1$ )  
    H&P_E ← H&P_E + 1  
    IF(send(P))  
        me_H&P_E ← me_H&P_E + 1  
    ENDIF  
ELSE IF ( $em_{ij} \in C2$ )  
    L_E ← L_E + 1  
    IF(send(P))  
        me_L_E ← me_L_E + 1  
    ENDIF  
ELSE IF ( $em_{ij} \in C3$ )  
    S_E ← S_E + 1  
    IF(send(P))  
        me_S_E ← me_S_E + 1  
    ENDIF  
ELSE IF ( $em_{ij} \in C4$ )  
    A_E ← A_E + 1  
    IF(send(P))  
        me_A_E ← me_A_E + 1  
    ENDIF  
ELSE  
    M_E ← M_E + 1  
    IF(send(P))  
        me_M_E ← me_M_E + 1  
    ENDIF
```

Step 4. Set  $Total\_E \leftarrow H\&P\_E + L\_E + S\_E + A\_E + M\_E$

Step 5. Set  $me\_TE \leftarrow me\_H\&P\_E + me\_L\_E + me\_S\_E + me\_A\_E + me\_M\_E$

Step 6. Set  $me\_Total \leftarrow me\_TE + me\_TW$

Step 7. Set  $Total \leftarrow Total\_E + Total\_W$

Step 8. Set  $R1 \leftarrow me\_Total/Total$  and  $R2 \leftarrow me\_TE/me\_Total$

Step 9. Set  $R3 \leftarrow me\_H\&P\_E/me\_TE$ ,  $R4 \leftarrow me\_L\_E/me\_TE$ ,  $R5 \leftarrow me\_A\_E/me\_TE$ ,  
and  $R6 \leftarrow me\_S\_E/me\_TE$

Step 10. Using weighted K-means, for data set with six features as R1, R2, R3, R4, R5, and R6, Compute k clusters.

Step 11. Resultant k cluster which are represented by k centroids will act as our model for predictions online social behavior for any new person.

Step 12. Return Model

Refer Table 1 for set of emoji represented by C1, C2, C3, C4, and C5. In step 1, we initialize all the count variables. In step 2 to 3.2, we scan each Whatsapp post corresponding to a user and calculate count of informative words for the user under consideration and combined count of both users in the chat. Function send (P) depicts that message is sent by the user under consideration. In step 3.3, frequencies of Emoji in chat belonging to categories <Happy/Playful>, <Love>, <Sad>, <Angry> and <Miscellaneous> are recorded. In step 4 to 9, we calculate the values of several variables defined at the beginning of algorithm which lead us to the value of the feature ratios R1, R2, R3, R4, R5, R6 as described in section 5 of proposed approach. In step 10, we assign weight 0.3 to R1 and R2 and 0.1 each to rest four ratios and train the model using weighted K-means clustering. In final step, we return the model with k centroids representing k clusters.

## **Results**

The proposed model uses Whatsapp chats of university students belonging to the same age group and profession, and then extracts significant information about their online social behavior. Detailed description of the dataset used, feature generation process and clustering results is discussed in this section.

## **Dataset**

300Whatsapp chats belonging to 30 individuals are collected using the 'Explore chat' feature provided by Whatsapp. These chats are of university students in age group 18-25. Snapshot of one of 300 chats is shown below: -

User 1: Person 1(Person under consideration)

User 2: Some User X interacting with Person 1.

6/10/16, 6:48 PM - User 1: hii

6/17/16, 9:46 AM - User 2: Do not get disturbed by such reviews.after all you are going to work for HCL

6/17/16, 9:48 AM - User 2: It is not exactly BPO.why HCL will hire B Tech if they can get normal graduate cheaply

6/17/16, 9:50 AM - User 2: If you join HCL now say for 2 years then you have something to show in CV

6/17/16, 9:52 AM - User 2: Do not forget what happened with Nidhi. She was sitting home idle for 2 yrs

6/17/16, 9:54 AM - User 2: You are not losing anything by joining HCL.any gain will be plus for you.

6/17/16, 9:56 AM - User 2: Take my word those who went back are going to repent.

**Figure 2 Snapshot of chat exchange between Whatsapp users**

### Feature Matrix

As discussed in model, after data cleaning, data integration and data refining through word and emoji processor, chat data is converted is converted to a 6 featured dataset. A snapshot of obtained feature matrix is shown in Table 3 below.

**Table 3 Values Retrieved for Select Features consolidated together to form Feature Matrix**

S. No.	R1	R2	R3	R4	R5	R6
1	0.447353	0.098017	0.553512	0.051839	0.023411	0.008361
2	0.544365	0.138767	0.492063	0	0.142857	0.079365
3	0.524988	0.062847	0.75	0	0.058824	0.161765
4	0.257499	0.089005	0.691176	0.014706	0.044118	0.161765
5	0.499441	0.035821	0.75	0	0.0625	0.020833
6	0.514536	0.098156	0.80203	0.001015	0.018274	0.005076
7	0.459687	0.105152	0.788	0.006	0.01	0.006
8	0.454516	0.10291	0.82069	0.02069	0	0
9	0.502547	0.254378	0.231884	0	0.032609	0.039855
10	0.462649	0.141657	0.457627	0.050847	0.04661	0.038136

### Clustering Results

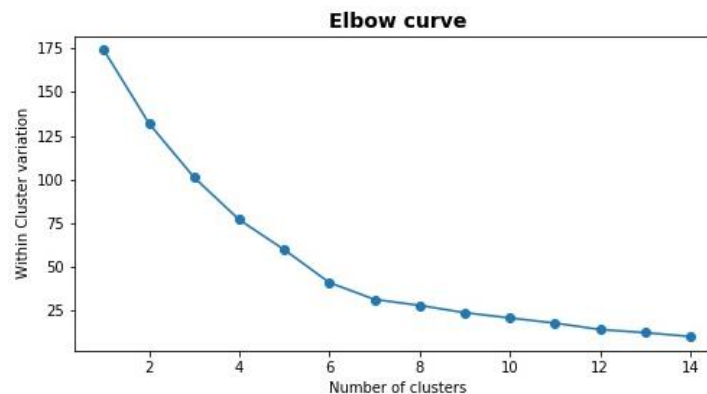
Weighted K-Means clustering is applied on feature matrix obtained to form K- clusters. Experimentation with different values of K is done to find the optimal value of K that groups our dataset in the best possible clustering configuration. For different values of K, results are shown below in Table 4 below.



**Table 4 Clustering Results for Different Values of K**

Cluster Number	Label	Number of persons in cluster at K=4	Number of persons in cluster at K=5	Number of persons in cluster at K=6
1	Happy/Playful & Expressive (L4)	6	6	6
2	Extrovert (L1)	12	11	9
3	Emoji Maniac (L3)	1	1	1
4	Mostly Happy/Playful Mood but Less expressive overall (L5)	11	11	11
5	Mostly Sad in mood but More expressive overall (L6)	NIL	1	1
6	In between L1 and L4, though more towards L1 i.e. is Extrovert	NIL	NIL	2

For validating the choice of K, we used the famous Elbow Method. The value of K may range from K=2 to K=15. The k-means algorithm tries to minimize the variation within the clusters by grouping similar objects in the same cluster. In Elbow method, we plot Within Cluster Variation against the Number of clusters. The location of bend (knee) in the plot is taken as the estimator of most suitable number of clusters.

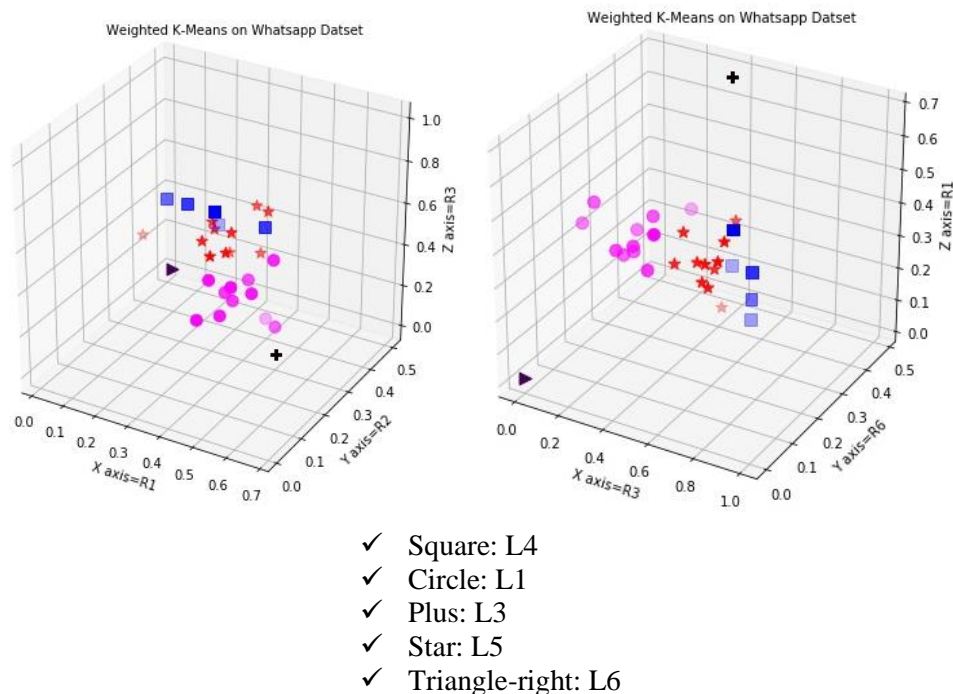


**Figure 3. Elbow Curve for selection of optimal value of K**

From Figure 3, we can observe a slight bend in graph between 4 and 6. For K=4, an interesting label depicting sadness as the general mood of the user is missed. From Table 4, we can see that for K=6, we come across a group having overlapping labels and it is difficult to assign one possible label. Thus, K=5 is chosen as the optimal number of clusters.

After applying weighted k-means on the dataset, K clusters are formed. The users belonging to the same cluster are grouped together and their feature ratios are examined. The dominant

ratios (ratios having significant values and not tending to zero) for each person in a cluster are determined. Based on the dominant feature values as explained in earlier sections, the cluster label is assigned. Thus, the cluster label is applicable to all users belonging to that cluster. For our dataset with  $K=5$ , the labels are Extrovert, Happy/Playful mood & Expressive, Happy/playful mood & Less expressive, Emoji maniac, Sad mood & Expressive. For graphical visualization of resultant clusters, 3-D graphs based on dominant features R1, R2, R3 and R6 are plotted. Two graphs one with R1, R2, R3 and other with R3, R6, R1 are plotted in three dimensional plane to pictorially represent the clusters, shown in Figure 4 below.



**Figure 4 Graphical Visualization of Clusters for  $K=5$**

Since, most of the participants fall under the same age group and profession, it was not expected that all possible behaviors listed in Table 2 could be detected. However, one interesting observation from results tells us that most people like to express a lot using Whatsapp i.e. they are social media extroverts. Also, the tone of conversation signals that the general mood is happy and playful for majority. This means most people are in jolly mood and have fun talking to their loved ones. It has been found that two people exhibit very distinct personality traits from remaining persons. One of these people prefers to express using Emoji a lot. So, it can be said that he is more expressive in terms of body language and facial expressions. There is another distinct trait showing that the general mood of the user is mostly sad on social media though he is very expressive.

In order to validate the labels assigned to each person, we enlisted the social media qualities of the various people used in our dataset as most of them are our friends or classmates and we are familiar with their texting patterns and online social behavior. The true label as per our knowledge of the person and the label assigned by clustering algorithm turned out to be same for almost all our participants thereby verifying our analysis.

## **Conclusion**

In this paper, an intensive analysis on Whatsapp chats of various university students has been performed for determining their online social behavior. It has been found from this analysis that students in this age group are mostly extroverts or extensively expressive and their general mood is happy and playful. In our small collection of data, we also found some outlier behavior like one student depicting sad mood in general and other student who is an emoji maniac. Such results encompassing outlier behavior can be beguiling for researchers working in psychological domain and provide counseling to help students exhibiting highly sad and depressed behavior to alleviate their stress. Thus, the results from this exploratory analysis provide new acumen about Whatsapp users and these results can be of interest to other groups such as social media influencers, applied behavior analysis, demographers etc.

This work is somewhat limited in containing only a small number of users and exclusively focusing on people between the ages of 18 and 25. Our dataset is biased in terms of age and profession and collecting data from all age groups and different professions could provide contrasting information. More accurate algorithms can be proposed through studying data from more users, with a wider range of ethnic backgrounds and professions. Enhanced form of clustering algorithms can be applied which may help to give faster results. Possible future work could also analyze people's social behavior on other social platforms. The time stamp in Whatsapp messages can be exploited to find out active durations of a user and it can be used for further online behavior analysis of a person. The above mentioned and all other issues shall be explored in greater detail in future work. Moreover, this analysis can be further extended to identify user interests and correlate user interests with user behaviors. Such analysis might be useful for generating a recommender system that can be employed to suggest suitable motivational videos, podcasts etc. catering to the specific interests of the user.

## **Acknowledgement**

The authors would like to thank the anonymous reviewers for their valuable and constructive comments and suggestions. They would also like to thank Ms. Kanak Sundriyal for her contribution throughout our journey to publish this work.

## References

- Alsanie, S.I. (2015). Social media (Facebook, Twitter, WhatsApp) used, and its relationship with the university students contact with their families in Saudi Arabia. *Universal Journal of Psychology*, 3(3), 69-72.
- Aramburu, M.J., Berlanga, R., & Lanza, I. (2020). Social Media Multidimensional Analysis for Intelligent Health Surveillance. *International Journal of Environmental Research and Public Health*, 17(7).
- Bai, Q., Dan, Q., Mu, Z., & Yang, M. (2019). A systematic review of emoji: Current research and future perspectives. *Frontiers in psychology*, 10.
- Biolcati, R., & Cani, D. (2015). Feeling alone among friends: Adolescence, social networks and loneliness. *Webology*, 12(2), 1-9.
- Borruto, G. (2015). Analysis of tweets in Twitter. *Webology*, 12(1).
- Dahiya, S., Gosain, A., & Mann, S. (2021). Experimental Analysis of Fuzzy Clustering Algorithms. *In Intelligent Data Engineering and Analytics*, 311-320.
- Hussain, Z., Mahesar, R., Shah, R., & Memon, N.A. (2017). WhatsApp Usage Frequency by University Students: A Case Study of Sindh University. *Engineering Science and Technology International Research Journal*, 1(4), 15-20.
- Johari, R., Kalra, S., Dahiya, S., & Gupta, K. (2021). S2NOW: Secure Social Network Ontology Using WhatsApp. *Security and Communication Networks*.
- Kalra, S., Johari, R., Dahiya, S., & Yadav, P. (2018). WAPiS: WhatsApp Pattern Identification Algorithm Indicating Social Connection. *In Advanced Computational and Communication Paradigms*, 595-603.
- Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, M. (2020). Determining disaster severity through social media analysis: Testing the methodology with South East Queensland Flood tweets. *International journal of disaster risk reduction*, 42, 101360.
- Khoo, S.S., & Yang, H. (2020). Social media use improves executive functions in middle-aged and older adults: A structural equation modeling analysis. *Computers in Human Behavior*.
- Kio, S.I. (2015). What students are saying on Facebook about their schools? *Webology*, 12(1).
- Krithiga, R., & Ilavarasan, E. (2020). A Novel Hybrid Algorithm to Classify Spam Profiles in Twitter. *Webology*, 17(1), 260-279.
- Kumar, N., & Sharma, S. (2016). Survey Analysis on the usage and Impact of Whatsapp Messenger. *Global Journal of Enterprise Information System*, 8(3), 52-57.
- Minhas, S., Ahmed, M., & Ullah, Q. (2016). Usage of Whatsapp: A Study of University of Peshawar, Pakistan. *International Journal of Humanities and Social Science Invention*, 5(7), 71-73.
- Nguyen, H.L., & Jung, J.E. (2017). Statistical approach for figurative sentiment analysis on social networking services: a case study on twitter. *Multimedia Tools and Applications*, 76(6), 8901-8914.
- Premalatha C, Arunachalam, Kumar (2017). Emotion Analysis of Whatsapp Group Chat Using Big Data Analytics. *International Journal of Innovations & Advancement in Computer Science*, 6(8).

- Rosenfeld, A., Sina, S., Sarne, D., Avidov, O., & Kraus, S. (2018). A study of WhatsApp usage patterns and prediction models without message content. *arXiv preprint arXiv:1802.03393*.
- Sarlan, A., Nadam, C., & Basri, S. (2014, November). Twitter sentiment analysis. *In Proceedings of the 6th International conference on Information Technology and Multimedia*, 212-216.
- Sharma, D., Sabharwal, M., Goyal, V., & Vij, M. (2020). Sentiment Analysis Techniques for Social Media Data: A Review. *In First International Conference on Sustainable Technologies for Computational Intelligence*, 75-90.
- Singh, S., Arya, P., Patel, A., & Tiwari, A.K. (2019). Social Media Analysis through Big Data Analytics: A Survey. *In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*.
- Sonagara, D., & Badheka, S. (2014). Comparison of basic clustering algorithms. *International Journal of Computer Science and Mobile Computing*, 3(10), 58-61.
- Young, K. (2013). Adult friendships in the Facebook era. *Webology*, 10(1), 1-18.
- Zamani, N.A.M., Abidin, S.Z., Omar, N.A.S.I.R.O.H., & Abiden, M.Z.Z. (2014). Sentiment analysis: determining people's emotions in Facebook. *In Proceedings of the 13<sup>th</sup> international conference on applied computer and applied computational science*, 111-116.
- Zhao, Y., Qin, B., Liu, T., & Tang, D. (2016). Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog. *Multimedia Tools and Applications*, 75(15), 8843-8860.