

## Grouping of Twitter users according to contents of their tweets

Farha Naznin, Anjana Kakoti Mahanta

Department of Computer Science, Gauhati University, Guwahati, India

---

### Article Info

#### Article history:

Received Oct 27, 2022

Revised Mar 30, 2023

Accepted Apr 2, 2023

---

#### Keywords:

Cluster characterization

Clustering

K-means algorithm

Twitter dataset

User analysis

---

### ABSTRACT

In today's world most of the people use social networking sites such as Twitter. They share their opinions and their views. through these media. Grouping these users will help us in different ways such as product recommendation, opinion mining, characterization of users based on their way of expressing their feelings. In this work, we present a technique to group the users based on the textual contents of the tweets. This technique is based on an unsupervised approach of machine learning that is clustering. A method is presented for representing the users using vector space model and TF-IDF weight scheme. K-means algorithm is employed for grouping the users using cosine distance as a distance measure. For the evaluation of this method, we construct a Twitter user dataset by using the Twitter application programming interface (API). A new technique is also proposed for characterization of the clusters formed. The experimental results are promising and from the study, it is found that the users in the clusters formed could be well defined by using the proposed cluster characterization technique.

*This is an open access article under the [CC BY-SA](#) license.*



---

### Corresponding Author:

Farha Naznin

Department of Computer Science, Gauhati University

Guwahati, India

Email: farha.gu@gmail.com

---

## 1. INTRODUCTION

Now-a-days most of the people use social networking sites to present their opinions and their views [1]. Through social media, people post their thoughts and share information with each other. They express their feelings through their comments and posts. Different people have different writing styles, use different words to express their feelings or opinions on different topics. Words used by the users to express their feelings give an indication on the inherent character of the users and their involvement on different topics [2]. Words are used by human beings to express their opinion, feelings, observations, emotions with different intensity values. People with the same kind of opinion or feelings also use different words to share their opinion or feelings. Some of our words convey meaning, some convey emotions, and some produce actions. Expressing emotions on a social platform is an even more difficult and sensitive task. The words used by the users give a picture on the behavior of a person when communicating in a social media platform. The personality of the users can be explored by examining linguistic style of the users [3]. Basically, using the words the users use in their posts, the users could be categorized in different classes. In short, through the contents and writing styles of comments and posts posted by users, characteristics of users are believed to be predicted. Grouping of users may help us in many ways such as recommending products [4], characterization of users based on their way of expressing their feelings etc. Analysis of users is an important area because it helps us to identify and characterize different types of users [5]. So, in research on social networks, user analysis has gained attention [6].

Twitter is one of the most popular social media platforms [7], [8]. It is widely used by different kinds of people with different backgrounds. Therefore, it is a good platform to be used to classify various people and categorize them. At the same time, it is challenging too. In Twitter, users can express their feelings with a

limitation of 280 characters [9]. So, it is actually a difficult task to find the characteristics of a user with such a limitation in words. Therefore, using tweets to classify users is a challenging task. The contents of the tweets are used in this work and analyzed to classify the users to find hidden pattern and knowledge.

User analysis in sentiment analysis area, mainly involves supervised techniques (such as classification and regression) and unsupervised techniques (clustering) of machine learning [10]. The accuracy of the supervised techniques is relatively high but the main problem with these techniques is that they are costly and need human interaction. Also labeled data is needed to perform these techniques. In this era where a very large amount and complex nature of data are produced every now and then from various platforms, these data often are not labeled, and to label the data a significant effort is required to be made by individuals with domain knowledge. So, it is an important task to group the data into meaningful clusters without having any prior label information. So, the unsupervised techniques i.e., clustering techniques become very useful in user analysis where there is no labeled data available. Although the accuracy of clustering is not as high as supervised methods, using the clustering techniques, the knowledge which is hidden could be found. Moreover, in some works [11], [12], it has been proved that accuracy of unsupervised methods could be enhanced to the level as that of supervised techniques in sentiment analysis area.

In this work, we attempt to categorize the users of Twitter with the help of the contents of their tweets. Moreover, here inherent information from the users is attempted to be extracted i.e., a categorization is attempted to be achieved where groups that are not predefined are obtained. Several works have been done to categorize the users but a single and specific viewpoint is analyzed in most of them, such as organization detection [6], [10]–[12], bot detection [13]–[17], political orientation detection [18], and age prediction [19]. In this work, a single perspective is not considered; rather the users are clustered and analyzed in general but considering the sentiment of the users. Users are classified based on the sentiment related words they frequently use in their tweets to express their feelings or opinions towards various topics or matters. Therefore, a vector space model is used to represent the users where only sentiment related words used by the users are considered as features. To classify the users into different clusters, the unsupervised approach of machine learning, which is clustering technique, is used employing the K-means algorithm. Here cosine distance is used in K-means as a distance measure. Also, a new characterization technique for the clusters obtained is proposed in this work for analyzing the users in the clusters based on the words used by the users.

The works mentioned above that are done on user analysis are all based on supervised techniques. Only a few works on categorizing or grouping Twitter users using unsupervised techniques of machine learning are found in the literature. In the study done in paper [20], the authors collected the data from a particular company, Nike. Once the features are extracted from the data, they use principal component analysis (PCA) for pruning the features. Then the K-means clustering algorithm is used on the data to find out the clusters using those features. They used the silhouette measure to obtain the optimal number of clusters and developed a metric to measure the quality of the clustering. Finally, they performed manual labeling on the clusters obtained by their proposed method. In the work mentioned in the paper [21], a study is done to analyze the behavior of the users along with the activeness of the users with respect to the time. Moreover, the authors examined to see how the quality of the clusters formed at different time intervals could be enhanced for top-k trending topics using these criteria. To categorize the social media users effectively for top-k trending topics, an activeness score function is defined in this research. In another study done in paper [22], the optimal cluster set is attempted to be obtained by the authors. To do this, hierarchical clustering is applied to the users to produce overlapping clusters at different levels of hierarchy. Then, they use mean and standard deviation to obtain the optimal cluster sets from the different sets of clusters. However, our work is different from all the above work as the focus of the proposed work is on how the users are clustered or grouped based on their sentiment related views which are shared through their tweets.

The contributions of this work are: i) for the evaluation process, a new Twitter user dataset is generated by extracting the tweets from the Twitter using Twitter API; ii) a representation method is presented to represent the users using the sentiment related words; iii) presenting a clustering-based method for categorization of users of Twitter; and iv) a new characterization technique is proposed to characterize the clusters of users. The rest of the paper is organized as follows. Section 2 contains the method adopted in this research followed by section 3 in which the experimental results are discussed in detail. The paper is concluded in section 4 with some outlines of the future work.

## 2. METHOD

The proposed method for grouping the Twitter users can be divided into three major steps—first one is the preprocessing of the dataset and to construct the vectors for representing the users of the tweets, the second one is clustering the users. Again, the third step is the characterization of clusters formed in the second step. Figure 1 shows the working of the proposed method.

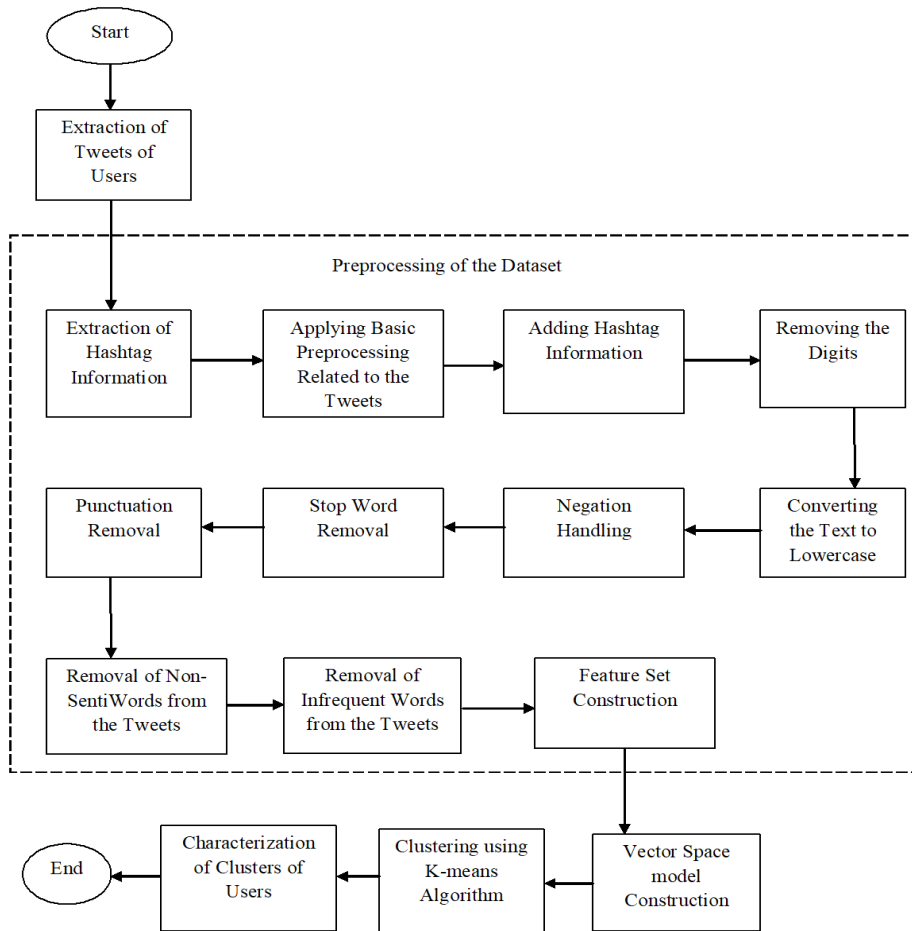


Figure 1. The proposed method

### 2.1. Extraction of the tweets of users

In this step, tweets are collected for each user from Twitter using the Twitter API. For each user, a separate file is maintained containing the tweets posted by that user. Then the collected data are subjected to pre-processing.

### 2.2. Pre-processing of the dataset

The input data for this phase is the separate text files, which are maintained for each user to store the tweets of that user. In this phase, the following processes are carried on the input data for preparing to be used in the next phase. The processes are described in detail.

### 2.3. Extraction of “hashtag” information

In this step, the “hashtag” information presented in the tweets is extracted from the tweets and stored in a variable. Since the “hashtag” may carry important information so we keep this information. In this step, “hashtag” information is not removed from the tweets but extracted to be stored. Here the “re” module of python which is used to work with regular expressions is used to accomplish it.

### 2.4. Applying basic pre-processing related to the tweets

In this step, the basic cleaning related to the tweets is done. This includes removing of emojis, hashtags, mentions, smileys, URLs, and reserved words (RT and FAV). For accomplishing it, the “preprocessor” package from python is used.

### 2.5. Adding “hashtag” information

It is already mentioned that “hashtag” information carries useful information since when “hashtag” is used in a tweet, all other tweets that includes the same “hashtag” are linked to it. So people can follow the

theme in which they are interested by following the “hashtag”. After basic cleaning, the “hashtag” information previously stored is added in the cleaned contents of the tweets obtained from the previous step.

### 2.6. Removing the digits

Digits are considered to be non-informative data in our analysis. Therefore, it is essential to remove this type of data from the set. After adding the “hashtag” information, the digits are removed from the tweets. The same python package “re” is used here for this purpose.

### 2.7. Converting the text to lowercase

As the data are English texts, we need to convert the text into a normalized case text. Usually, text data in the dataset may consist of several occurrences of upper-case characters. In this step, all the letters are converted to lowercase.

### 2.8. Negation handling

Negations can change the polarity of the other words in a sentence. For example, ‘not good’ will mean ‘bad’. In this work, negation handling is important, because here the words play the main role for categorizing the users. So, if negations are not handled properly, the meaning of the words in a sentence where a negation word is present will mean differently. For handling negations, we use the antonym relationships from WordNet, where WordNet is a large lexical database of English [23]. We keep a list where, the negation terms such as no, not, never etc. are stored. Whenever a negation term is found in a sentence, for each sentiment related word present in the sentence, we extract a list of the antonyms of that word from the WordNet. Here, adjectives and adverbs are considered as sentiment related words, since adjectives and adverbs are good representatives of the sentiment related words [24]. After this, the word along with the negation term is replaced by the first word in the antonym list.

### 2.9. Stop word removal

The sentiment perspective of users is considered in this work. Therefore, the focus of this work is on the sentiment related words only. So, stop words are not needed to be kept in this work and hence they are removed from the text.

### 2.10. Punctuation removal

Punctuations are removed from the data set prepared for processing, as they normally do not contribute to decision that we intend to extract. They are not needed in the text to be processed in the next steps. In this step, all such punctuations are removed.

### 2.11. Removal of non SentiWords from the tweets

Since we are going to classify the users according to their sentiment perspective, so only sentiment related words in these files are kept. As mentioned earlier, adjectives and adverbs are good representative of the sentiment related words. That is why, from the tweets of the users, the words other than the adjectives or adverbs are removed and only SentiWords (adjectives and adverbs) are kept in the files. To determine the adjectives and adverbs from text, the *nlk* pos-tagger of python is used here.

### 2.12. Removal of infrequent words from the tweets

In this step, for each user, a list of distinct words is extracted from all the tweets of the user. The frequencies of words in the tweets for that user are also stored. Words with lower frequencies are considered to be of lesser informative. For each user, words with the frequency are then removed.

### 2.13. Feature set construction

In this step, a master word list of distinct words is constructed from all the lists of the distinct words created for the users. The master word list represents the feature set and the words in this list represent the features in the feature set. These are considered as significant informative tokens and shall contribute to the decisions during analysis.

### 2.14. Vector space model construction

From the master word list created for the users, a matrix is formed where each row represents a user, and each column represents a feature in the feature set. Term frequency-inverse document frequency (TF-IDF) weight scheme is used to weigh the words (features) of the users in the matrix. As discussed earlier, each user is represented by the list of frequent sentiment words that are used by the user and so one such list will correspond to one document in the definition of TF-IDF. In this scheme, term frequency (TF) measures the

importance of a particular word in the document and the importance of the word in the corpus is measured by using inverse document frequency (IDF). The mathematical expression of TF is:

$$tf_{i,j} = \frac{t_i}{l_j} \quad (1)$$

where  $t_i$  is the frequency of term  $i$ ,  $l_j$  is the length of document  $j$  where term  $i$  has occurred. Again, the mathematical expression of IDF is:

$$idf_i = \log\left(\frac{D}{df_i}\right) \quad (2)$$

where  $D$  is the number of all the documents under consideration and  $df_i$  is the number of documents where term  $i$  occurs. Now for term  $i$ , the mathematical expression of TF-IDF is (3).

$$tfidf = tf_{i,j} \times idf_i \quad (3)$$

### 2.15. Clustering on users using K-means algorithm

In this work, on the matrix that we have created, K-means clustering algorithm is applied to categorize the users using the cosine distance measure. In general, in K-means the centroids are selected randomly [25]. If the selected centroids are of the same type or in same cluster, then this leads to poor clustering result. That is why we have used K-means++ to find out the initial centroids. K-means++ is a smart technique for centroid initialization [26]. Then the centroids obtained by using K-means++ are used as the initial centroids for K-means algorithm. Cosine distance: using this measure, the similarity of the two vectors is computed by measuring the cosine of the angle between them. The cosine distance between two vectors of attributes,  $A$  and  $B$ , is represented by:

$$\cos \theta = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (4)$$

### 2.16. Characterization of clusters of users

Next, we present a new characterization technique to characterize the clusters that were formed after applying K-means algorithm on the vectors representing the users. Our aim is to find out a set of good representative words for each cluster. A word will be a good representative for a cluster if it is more frequent in the cluster but not frequent or less frequent in other clusters. To do this, each cluster is represented by the words that are used by the users in the cluster. then a score is given to each word in the clusters using the concept of word frequency-inverse cluster frequency (WF-ICF). The score WF-ICF is the multiplication of word frequency (WF) and inverse cluster frequency (ICF) of a word. WF is the number of times the word appears in a particular cluster and ICF is the inverse of the average of the occurrences of that word in other clusters. Basically, WF is the frequency of the word in the cluster and ICF is the inverse of the average of CF values of the word in other clusters. So, if the WF-ICF value of a word  $w$  in a cluster is large then it means that the word  $w$  is frequent in the cluster but infrequent or less frequent in other clusters. So, we take the words with largest value of WF-ICF of a cluster as representative.

Here one problem may arise that if a word is frequent in cluster  $C$  but its appearance in other clusters is 0, then the ICF cannot be computed because it leads to 'division by zero' error while calculating the ICF value. That is why whenever we see that a word is frequent in one cluster but its appearance in the other clusters (CF value) is 0, the value of ICF is taken as 1. The reason behind doing this is that the ICF value increases as the CF value decreases. The theory behind this is that if a word is frequent in one cluster and the appearance of the word in other cluster is 0, according to our logic, its chance of becoming a representative for the cluster where it is frequent will be the highest. Definition of WF-ICF: Let  $C = (c_1, c_2, c_3, \dots, c_n)$  be a set of clusters. We are to find the characterization of the clusters in the set  $C$ . Let  $c$  be the cluster under consideration. The cluster  $c$  can be expressed as the set of words  $\{w_1, w_2, w_3, \dots, w_m\}$ , where each  $w_i$ , ( $i=1$  to  $m$ ) is a word used by a user in the cluster  $c$ . So, for a word  $w$  in a cluster  $c$ ,  $WF$  is the number of occurrences of  $w$  in cluster  $c$ .

$ICF$  is the inverse of the average of frequency values of the word  $w$  in clusters other than  $c$ . Let,  $CF$  be the total number of occurrences of  $w$  in other clusters and  $N$  be the number of such clusters containing  $w$ , then  $ICF$  can be expressed mathematically as (5).

$$ICF = \begin{cases} 1, & \text{if } CF = 0 \text{ which means } N = 0 \\ 1/((CF + 1)/N), & \text{otherwise} \end{cases} \quad (5)$$

From the definition mentioned above, it is clear that *ICF* values are strictly less than 1 in all cases except the first case in which its value is 1. Now word frequency-inverse cluster frequency, *WFICF* of a word *w* can be expressed mathematically as (6).

$$WFICF = WF \times ICF \quad (6)$$

In this scheme, for the same *CF* value, the *ICF* value will be less if the number of other clusters containing the term i. e., *N* is smaller. For example, suppose in case 1, *N*=1, *CF*=10 and in case 2, *N*=2, *CF*=10, then in case 1, *ICF* will be 1/11 and in case 2, it will be 2/11. So, the value of *ICF* in case 2 is greater than the value of *ICF* in case 1. It means if value of *CF* of a word is same, the appearance of the word in more clusters is better than its appearance in a smaller number of clusters. It is because, if a word appears in large number of clusters, then the frequency of that word in individual outside cluster will be less and this is needed for a word to become a representative of a cluster.

### 3. RESULTS AND DISCUSSION

For evaluating the proposed method, the silhouette measure is used. This is a metric using which the goodness of a clustering technique is computed [27]. We experimented with our created dataset, and experimental results are presented. The process which is used to create the dataset is also described here. Quantified information of the dataset is also included in this section.

#### 3.1. Experimental data collection

For the evaluation process of the proposed method, we have constructed a Twitter user dataset by using Twitter API. This dataset contains 150 tweets in average from 1,557 users. We created a list of names of Twitter users. For creating this list, we first search tweets using geo code for some locations and extracted the list of tweets. We use “39.8, -95.583068847656, 2,500 km” search string to get tweets from the lower 48 states of the USA with radius of 2,500 km, “53.81982, -2.406348, 500 km” search string to get tweets within 500 km of the whalley arms public house in Whalley UK. Also, we use the search string “40.714353, -74.0059, 5,000 km” to get tweets around the area within 5,000 km from New York City. From the information related to the tweets, for each location, names of the users of the tweets are extracted in separate lists. From these lists distinct users’ names are extracted. Then the lists are merged. Moreover, we add names of some random celebrities from India to the list of users’ names. Again, some users from the list of users’ names are not accessible. So, we eliminated those users from our final list of users’ names. So, the final list contains 1,557 users. In this way we have constructed the Twitter user dataset to be used for the experiment of the proposed method.

#### 3.2. Results and analysis

After basic pre-processing related to the tweets is done, we applied negation handling, stop word removal and punctuation removal to each of the tweets. Then for each user, a list is created which contains the distinct words that are extracted from the tweets of the user. In other words, these lists contain the words that are used by the user. After this step, we keep only SentiWords that is the adjective and adverbs from the list and other words are removed from the list. Now the words with low frequency are removed from the list using a considered threshold. In our experiment the words with frequency 2% of the total words for a particular user and above are kept. After doing these two steps, some files had no contents in it. So those files are removed. After getting the lists for each user, a master word list is created which contains all the distinct SentiWords from the lists of words. The master word list contains 1,003 words. Then the vector space model is created using TF-IDF weight scheme from the lists of words for each user and the K-means algorithm is applied on it. Here, we must select the optimal number of clusters. To get this, we ignore illogical solution of  $k=1$ , and consider  $k=15$  as upper-bound value in this work as in our work, it seems reasonable. Then K-means algorithm is run iteratively for the values from  $k=2$  to  $k=15$  and for each result silhouette coefficient is calculated. So experimentally we get optimal value of  $k=15$  with Silhouette measure=0.0353. From the 15 clusters two clusters are removed as these two contains a smaller number of users. That is why these two are considered as noise. So, we now have 13 clusters. After obtaining the clusters in the second phase, in the third phase, a computation is performed in which, a characterization of each cluster is done by extracting words that have high WF-ICF values. The first ten words with highest WF-ICF values are used as representative words for each cluster. In Table 1, the representative words of the clusters with their WF-ICF values are shown.

Using this scheme, some of the clusters obtained are very well definable. Cluster 1 is a group of users who mainly use words having negative emotion such as ‘violent’, ‘hateful’, ‘obsessed’, ‘sad’ etc. Cluster 2 is a group of users who use the words ‘kindly’, ‘dear’, ‘gracious’ frequently and so this group of people are sober, gentle, and soft spoken. Cluster 3 is a group of people who have a nature of appreciating other people or objects.

From the words used by cluster 5, a positive attitude towards society and the future is being seen. Cluster 6 is the only cluster where the terms ‘democratic’ and ‘needy’ are used and this group may represent people actively engaged or are interested in politics. Cluster 7 is a group of people who have a nature of being grateful and having a positive attitude. The word ‘global’ is present in cluster 8 with high frequency, which is not frequent in any other cluster. Also, the words ‘rank’, ‘high’, ‘heavy’, ‘difficult’ and ‘firm’ are frequent in this cluster. This may indicate that people in this cluster are interested in global problems. Cluster 9 is a group of users who dominantly talk about fatigue related words like ‘wan’, ‘tired’. Cluster 10 has the words ‘daily’ and ‘everyday’ with high frequency values. Also, the word ‘halal’ is present only in this cluster. Users in cluster 11 use words like ‘crying’, ‘cut’, ‘upset’, ‘poor’ which are related to sad emotion and so they are mostly unhappy. Users in cluster 12 use words like ‘game’, ‘round’. mostly and so have an interest in games. Similarly, the users in cluster 13 use a special word ‘pandemic’. Further investigations are required to ascertain the patterns that have been discussed here. From the results obtained from the experiments performed using the proposed method we are able to categorize the users in different groups, where the groups are not specified earlier. That means we are able to extract or find out some hidden information from the text used in the tweets by the users without giving any pre-specified labels. The groups are created based on the sentiment perspective of the users. Moreover, the users in the groups or the clusters are well defined using the proposed scheme.

Table 1. The representative words of the clusters with their WF-ICF values

No.	Representative words with their WF-ICF values
1	(True, 21.176470588235293), (violent, 2.4), (regular, 2.333333333333333), (finished, 2.0), (Sweet, 1.8285714285714285), (hateful, 1.7999999999999998), (similar, 1.6666666666666667), (vocal, 1.6), (obsessed, 1.5555555555555554), (sad, 1.5211267605633803)
2	(Kindly, 7.0), (dear, 3.6), (gracious, 2.0), (otherwise, 1.6), (course, 1.0810810810810811), (longest, 1.0), (Poor, 0.9090909090909092), (forced, 0.8999999999999999), (repellent, 0.8275862068965517), (Zero, 0.7777777777777777)
3	(Beautiful, 14.486486486486488), (cute, 7.459459459459459), (lovely, 7.125), (blessed, 5.0), (Lucky, 4.666666666666666), (gorgeous, 4.375), (fabulous, 4.0), (healthy, 3.3333333333333335), (Perfect, 3.142857142857143), (disgusted, 3.0)
4	(Main, 6.0), (innocent, 5.625), (mere, 5.0), (free, 4.551724137931035), (near, 3.230769230769231), (grand, 3.0), (tight, 2.4000000000000004), (calm, 2.25), (existing, 2.0), (mass, 1.9090909090909092)
5	(Small, 10.461538461538462), (fair, 10.4), (sure, 10.12820512820513), (open, 9.117647058823529), (virtual, 8.0), (latest, 7.0), (nice, 6.782608695652174), (social, 6.107142857142857), (honest, 6.0), (future, 5.862068965517242)
6	(Needy, 3.0), (original, 2.7777777777777777), (cursed, 2.0), (wild, 1.7999999999999998), (Democratic, 1.3333333333333333), (implicit, 1.25), (pacific, 1.0), (closer, 0.8333333333333334), (extended, 0.8), (Electric, 0.75)
7	(Happy, 8.901098901098901), (understanding, 5.0), (thankful, 4.8), (amazing, 4.240963855421686), (Excited, 4.186046511627907), (dramatic, 4.0), (fucking, 3.6458333333333335), (proud, 3.0357142857142856), (rough, 2.5), (truly, 2.3333333333333333)
8	(Global, 25.0), (dangerous, 6.461538461538462), (high, 5.296296296296296), (correct, 4.19047619047619), (Rank, 3.0), (immediately, 2.0), (heavy, 1.5), ('firm', 1.3333333333333333), ('difficult', 1.090909090909091), (Involved, 1.0769230769230769)
9	(Wan, 27.0), (damn, 15.279069767441861), (tired, 5.037037037037036), (fine, 4.324324324324324), (Proud, 3.0357142857142856), (handsome, 3.0), (fucking, 2.9702970297029703), (gone, 2.9189189189189193), (Quick, 2.6666666666666665), (alright, 2.571428571428571)
10	(Daily, 6.125), (everyday, 4.9090909090909091), (dry, 2.0), (halal, 1.0), (alone, 0.9473684210526315), (Forced, 0.8999999999999999), (finished, 0.875), (expensive, 0.7), (alright, 0.6666666666666666), (nasty, 0.625)
11	(Exactly, 22.8), (crying, 8.666666666666666), (actually, 8.316831683168317), (cut, 8.285714285714285), (Used, 7.791666666666667), (yet 7.479999999999999), (upset, 7.0), (always, 6.905027932960894), (Even, 6.477777777777778), (poor, 6.363636363636363)
12	(Game, 19.333333333333332), (full, 5.348837209302325), (round, 3.4285714285714284), (pat, 3.0), (Better, 2.9747899159663866), (annoying, 2.769230769230769), (wild, 2.739130434782609), (Best, 2.5044642857142856), (fast, 2.4705882352941178), (fake, 2.4347826086956523)
13	(Black, 12.222222222222221), (white, 5.0285714285714285), (supreme, 5.0), (variant, 4.0), (Corrupt, 3.333333333333333), (spiritual, 3.0), (pandemic, 2.8421052631578947), (fresh, 2.6666666666666665), (narrative, 2.4000000000000004), (new, 2.31404958677686)

#### 4. CONCLUSION

In this work, the users of Twitter are categorized according to the contents of their tweets. Here a technique is proposed to represent the users using the SentiWords i.e., sentiment related words that are used by them in their tweets. Then the users are grouped by applying clustering technique using the K-means algorithm. Again, a characterization technique is proposed to characterize the clusters formed. With the characterization technique, the representative words are extracted from the words that are used by the users in the cluster. From the study done here, it is seen that these words give an indication on the inherent characteristics of the users and their views on different topics. As the future work, an experiment also could be

done to see the effect of the clustering result with other words also along with sentiment related word. Again, the experiment could also be done on users from other platforms such as Facebook and WhatsApp. Emotion mining could also be done on the Twitter users.

## ACKNOWLEDGEMENTS




This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## REFERENCES




- [1] N. L. Devi, "Analyzing Twitter data using unsupervised learning techniques," *Journal of Network Communications and Emerging Technologies (JNCET)*, vol. 7, no. 11, pp. 28–32, 2017.
- [2] Â. Jusupova, F. Batista, and R. Ribeiro, "Characterizing the personality of twitter users based on their timeline information," in *Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao*, Sep. 2016, vol. 16, pp. 292–299, doi: 10.18803/capsi.v16.292-299.
- [3] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference," *Journal of Personality and Social Psychology*, vol. 77, no. 6, pp. 1296–1312, 1999, doi: 10.1037/0022-3514.77.6.1296.
- [4] E. Khazaei and A. Alimohammadi, "An automatic user grouping model for a group recommender system in location-based social networks," *ISPRS International Journal of Geo-Information*, vol. 7, no. 2, p. 67, Feb. 2018, doi: 10.3390/ijgi7020067.
- [5] F. K. De Oliveira, M. B. D. Oliveira, A. S. Gomes, and L. M. Queiros, "Statistical grouping methods for identifying user profiles," *International Journal of Technology and Human Interaction*, vol. 15, no. 2, pp. 41–52, Apr. 2019, doi: 10.4018/IJTHI.2019040104.
- [6] J. McCorriston, D. Jurgens, and D. Ruths, "Organizations are users Too: Characterizing and detecting the presence of organizations on twitter," in *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, 2015, pp. 650–653, doi: 10.1609/icwsm.v9i1.14672.
- [7] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning," *Information (Switzerland)*, vol. 9, no. 5, p. 127, May 2018, doi: 10.3390/info9050127.
- [8] A. Shelar and C. Y. Huang, "Sentiment analysis of twitter data," in *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCSI 2018*, Dec. 2018, pp. 1301–1302, doi: 10.1109/CSCSI46756.2018.00252.
- [9] K. Gligorić, A. Anderson, and R. West, "How constraints affect content: the case of Twitter's switch from 140 to 280 characters," in *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 2018, pp. 596–599, doi: 10.1609/icwsm.v12i1.15079.
- [10] R. J. Oentaryo, J. W. Low, and E. P. Lim, "Chalk and cheese in Twitter: Discriminating personal and organization accounts," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9022, 2015, pp. 465–476.
- [11] L. De-Silva and E. Riloff, "User type classification of tweets with implications for event recognition," in *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 2015, pp. 98–108, doi: 10.3115/v1/w14-2714.
- [12] K. E. Daouadi, R. Z. Rebaï, and I. Amous, "Organization vs individual: Twitter user classification," in *CEUR Workshop Proceedings*, 2018, vol. 2279, pp. 1–8.
- [13] M. Singh, D. Bansal, and S. Sofat, "Who is who on Twitter—spammer, fake or compromised account? A tool to reveal true identity in real-time," *Cybernetics and Systems*, vol. 49, no. 1, pp. 1–25, Jan. 2018, doi: 10.1080/01969722.2017.1412866.
- [14] D. Stukal, S. Sanovich, R. Bonneau, and J. A. Tucker, "Detecting bots on Russian political Twitter," *Big Data*, vol. 5, no. 4, pp. 310–324, Dec. 2017, doi: 10.1089/big.2017.0038.
- [15] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pp. 280–289, 2017.
- [16] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 French presidential election," *First Monday*, vol. 22, no. 8, Jul. 2017, doi: 10.5210/fm.v22i8.8005.
- [17] O. Loyola-Gonzalez, R. Monroy, J. Rodriguez, A. Lopez-Cuevas, and J. I. Mata-Sanchez, "Contrast pattern-based classification for bot detection on Twitter," *IEEE Access*, vol. 7, pp. 45800–45817, 2019, doi: 10.1109/ACCESS.2019.2904220.
- [18] D. Proetiuc-Pietro, D. J. Hopkins, Y. Liu, and L. Ungar, "Beyond binary labels: Political ideology prediction of twitter users," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, vol. 1, pp. 729–740, doi: 10.18653/v1/P17-1068.
- [19] R. G. Guimarães, R. L. Rosa, D. De Gaetano, D. Z. Rodríguez, and G. Bressan, "Age groups classification in social network using deep learning," *IEEE Access*, vol. 5, pp. 10805–10816, 2017, doi: 10.1109/ACCESS.2017.2706674.
- [20] V. Friedemann, "Clustering a customer base using Twitter data," *Cs*, vol. 229, no. 1, pp. 1–5, 2015.
- [21] T. T. Aurpa and T. C. Detection, "Clustering Active Users in Twitter Based on Top-k Trending Topics Clustering Active Users in Twitter Based on Top- k Trending Topics," no. October, 2020, doi: 10.13140/RG.2.2.19936.30726.
- [22] A. Paul, A. Dutta, and F. Coenen, "Cluster of tweet users based on optimal set," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, Nov. 2017, pp. 286–290, doi: 10.1109/TENCON.2016.7848008.
- [23] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990, doi: 10.1093/ijl/3.4.235.
- [24] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. S. Subrahmanian, "Sentiment analysis: Adjectives and adverbs are better than adjectives alone," in *ICWSM 2007 - International Conference on Weblogs and Social Media*, 2007, pp. 1–4.
- [25] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.
- [26] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, vol. 07-09-January-2007, pp. 1027–1035, 2007.
- [27] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. C, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.



**BIOGRAPHIES OF AUTHORS**

**Farha Naznin**    is a research scholar in the department of Computer Science in Gauhati University, Assam, India. She obtained B.Sc. degree in Computer Science from Gauhati University in 2009 and M.Sc. degree in Computer Science from Gauhati University in 2011. Her research interests include data mining, natural language processing and pattern recognition. She can be contacted at email: farha.gu@gmail.com.



**Anjana Kakoti Mahanta**    is working as a professor in the department of Computer Science in Gauhati University, Assam, India. She obtained B.Sc. degree in Mathematics from Gauhati University in 1981 and M. Sc. degree in Mathematics from Gauhati University in 1983. She obtained a PGDCSA degree from Gauhati University in 1986. She obtained a Ph.D. degree in Computer Science from Gauhati University in 1990. Her research interests include data mining, design of algorithms, and pattern recognition. She can be contacted at email: anjana@gauhati.ac.in.