Emotion Prediction using Machine Learning Techniques

Areeba Shamsi[1], Sabika Nasir[2], Mishaal Amin Hajiani[3], Afshan Ejaz[4], Dr Syed Asim Ali[5]

[1,2,3,4]Department of Computer Science Institute of Business Administration Karachi,Pakistan
[5]Department of Computer Science/ UBIT University of Karachi, Pakistan

Abstract
Sentiment analysis is the field of study that analyzes people's opinions, sentiments, evaluations, and emotions from written language or their voice. Nowadays, an increasing number of online users has led to the growing influence of human emotions on the online community. Understanding the opinions behind user-generated content automatically is of great help for commercial and political use, among others. This assignment can be conducted on different levels by classifying the polarity of words, sentences or entire documents. Various emotions are conveyed on social media, it helps to identify the mood of a user with which the review was written.
This project focuses on the implementation of unsupervised learning by applying different types of clustering techniques such as k-means and fuzzy c-means on a data describing human emotions. The emotions in the content are clustered with basic emotions such as fear, sad, happy etc. Emotional analysis can be used for efficient recommendation process.

Key words:
Recommendation process, sentimental analysis, unsupervised learning, user generated content, K-means, fuzzy c-means.

1. Introduction

Sentiment analysis refers to the task of natural language processing to determine whether a piece of text contains some subjective information or not. What sort of subjective information it expresses, i.e., whether the attitude behind this text is positive, negative or neutral. As human beings, we go through numerous emotions which largely influence our day-to-day tasks like, our reviews about a particular product, opinions and comments on social media. Emotions are considered to have a large impact on mood, attitude, character, and temperament. There are various ways that we can handle the emotions people use in writing reviews. In social media, information is present in a large amount. Extracting information from social media gives us several usages in various fields
Emotion analysis is an interesting blend of psychology and Technology and has become a popular task which tries to predict sentiments from texts and voice speech. Emotion classification is to classify the words in user reviews, blogs, and twitter posts into several clusters of emotions based on similarities of these words with each emotion to provide pattern and understanding of given user content. Happy, sad, fear, disgust, surprise, neutral, and anger are some of the basic emotions. Several types of research are being conducted in areas like Artificial Intelligence, human-computer interaction (HCI) along with the evolution of affective/social computing. and not as an independent document. Please do not revise any of the current designations.
Emotion detection is basically a way to determine how the users are responding to the website, blogs and social media posts. This results in an improvement in the areas of marketing and advertising by detecting human emotions and then acclimating user experiences to these emotions in real time. It can be used in conversational voice recognition systems and voice response systems like google voice, Apple's Siri or Cortana. It can also provide assistance to people with speech impediments. This technology aids to design robots which are not only human friendly but are also capable to recognize user's emotions and needs.

2. Literature Review

With the emergence of sentimental computing, a lot of research is being done on sentiment analysis. Some of the works are summarized below:
Sivaraman sriram, xiaobu yuan, an enhanced methodology for classifying Emotions using decision tree algorithm. As there are various ways to identify emotions for instance, textual conversation, facial recognition, and dynamic gesture recognition. Here artificial neural network model is also used for emotion detection, here they found out mean and root mean square for all values in the data. Techniques like data mining or gene prediction system can also be used, but this paper implements this above methodology in applications like classification of videos with respect to the emotion
Oscal T.-C. Chen et al, this paper describes the proposed age and gender recognition systems based on arousal intensities of speaker's emotions. Additionally, k nearest neighbors is used as a classifier. Four sentiments angry, calm, happy, and sad are evaluated first. It was known that angry and happy mostly have higher arousal as compared to calm and sad sentiments. But the results showed that the gender recognition prefers calm whereas the age recognition prefers angry and happy. Thus, this recognition

Manuscript received June 5, 2019
Manuscript revised June 20, 2019

systems can be widely used for multiple voice interface applications

Bhowmik et al, learns a multilevel classification model to classify review content into various emotion categories. This work considers the words present in the sentences in the oppositions of the subject, object and verbs which were used as features.

Ghazi et al arranges the emotions in terms of neutrality, polarity and hierarchically to improve the performance of emotion categorization.

Moreo et al proposes a lexicon-based sentiment analyzer which allows users to express their views in non-standard language and detects the target of users' opinions in a multi-domain scenario.

Yuxiamg et al, introduced reader sentiment classification model that let users to extract documents that contain related content and simultaneously, produce proper emotions

Shenghua Bao, Shengliang Xu, introduced social emotion mining for document classification so that the document can be selected on the basis of the emotional preferences of online users. It links the online document and user generated social emotion. Affective words are extracted using text mining and are connected witsh related emotions. This model can help in exposing the hidden topics that portrays strong emotions. This methodology can be applied in songs and emotion aware advertisement recommendation systems.

Similarly, the purpose of this project is to collect words from various reviews, blogs and posts and categorized these words into different categories helping to define user's mood. The techniques used in our model are based on unsupervised learning whereas, most of the project mentioned above have used the combination of supervised and unsupervised learning such as artificial neural networks and k nearest neighbours (KNN) for emotion detection.

3. Data Information

The dataset has been picked from Kaggle's website. This dataset contains 1104 rows defined against 8 columns.

A. Data Collection

Words from Blogs, Twitter, and social media were collected. Next, these words were categorized into 7 basic emotions like disgust, Surprise, Neutral, Anger, Sad, Happy and Fear. Then probabilities of existence of these words in Disgust, Surprise, Neutral, Anger, Sad, Happy and Fear Sentences were calculated and assembled in a CSV file. The Naive Bayes Algorithm was used to calculate the "probabilities of existence" of these Words

B. Data Understanding

Attribute information:

1. Words
2. Disgust
3. Surprise
4. Neutral
5. Anger
6. Sad
7. Happy
8. Fear

C. Software Used

- KNIME Analytic Platform
- R

D. Data Pre-Processing

The missing value node was used to cater any missing values in the dataset, a tuple containing any null cell was removed from the dataset. Since this dataset contains probabilities which range from [0-1], there was no need to normalize the data.

E. Data Sampling

The dataset is divided into 3:7- test: train ratio. 30% of data is kept for testing while 70% is used to train the model.

F. Limitations

This dataset is only restricted to 7 emotions, the dataset could have been classified to more emotions like anxiety or excitement. More attributes like user's gender and age group could have been brought into consideration at the time of data collection which may help to provide better results.

4. Proposed Model

To accomplish the objectives, the following process model is used:

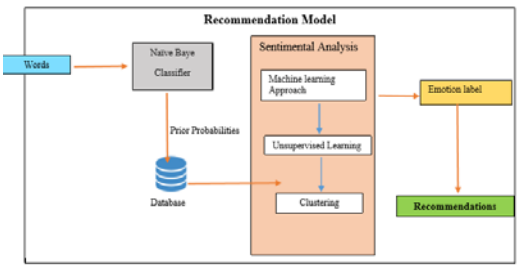


Fig. 1 Recommendation Process Model

The incoming word is first passed through the Naïve Baye classifier, then the word and its associated probability calculated from this classifier are collected and stored in the database. This data is then trained into a sentimental model. To classify sentiments two models are prepared which is discussed in the next section. Results from this model benefit the recommendation process.

5. Data Modelling

MODEL A: K-MEANS

A. K-means Algorithm

K-Means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because different location causes a different result. So, the better choice is to place them far away from each other as much as possible. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is unresolved, the first step is completed. At this point, it is necessary to re-calculate k new centroids as bar centers of the clusters resulting from the previous step. After getting these k new centroids, a new binding has to be done between the same data points and the nearest new centroid. A loop has
been generated. As a result of this loop, one may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

$$J = \sum_{j=1}^{k} (\sum_{i=1}^{n} (|x_i^j - c_i|^2))$$

Formula (1)

Where $|x_i^j - c_i|^2$is the chosen distance measured between data points and cluster centerc_i, is an indicator of the n data points from their respective cluster centers.

Steps:

1. Clusters the data into K groups where K is predefined.
2. Select *k* points at random as cluster centers.
3. Assign objects to their closest cluster center according to the *Euclidean distance* function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

B. Knime workflow for K-means

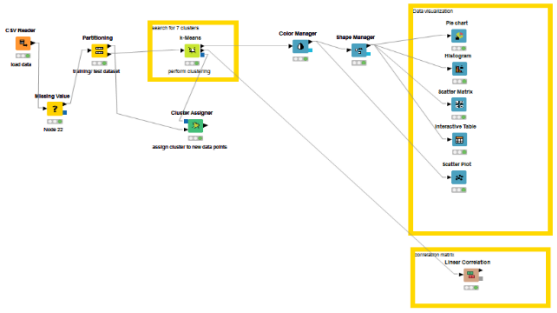


Fig. 2 K-means workflow with Pre-processing and sampling.

The data is first pre-processed by removing any record consisting a missing value it is then splitted into train and test data-70% 30% respectively. K-means algorithm is runned on the training set. When the dataset is trained with the k-means algorithm, it is assigned to the Cluster Assigner which caters the new data points and is used to test the model assigned to the assigner node.

C. Data Represention for K-means

Table 1

Key Chart for Knime visualization	
Cluster 0	Disgust
Cluster 1	Neutral
Cluster 2	Anger
Cluster 3	Surprise
Cluster 4	Sad
Cluster 5	Happy
Cluster 6	Fear

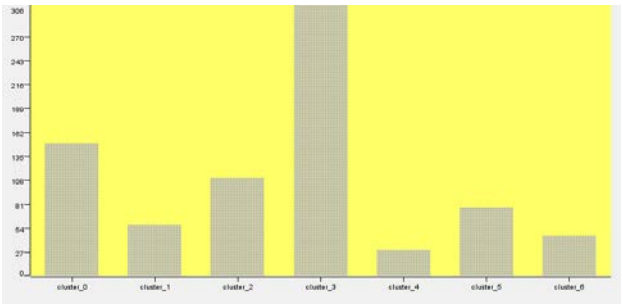
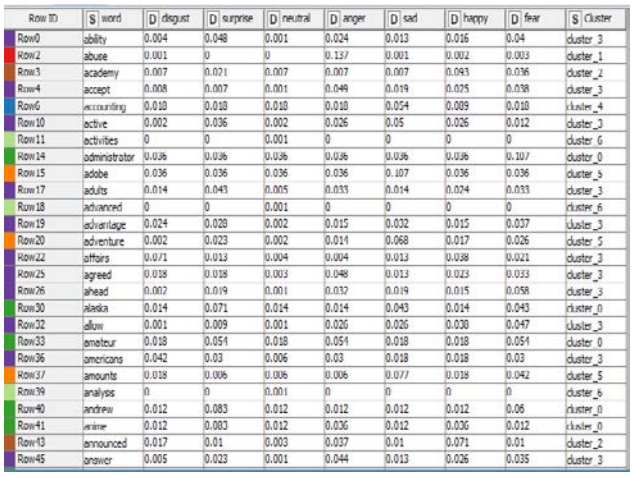


Fig. 3 Histogram: Categorical Grouping of Clusters

shows that highest number of words are clustered in cluster_3 representing “Surprised” while least numbers of words are clustered as “Sad” (cluster_4)



Row ID	word	disgust	surprise	neutral	anger	sad	happy	fear	Cluster
Row0	ability	0.004	0.048	0.001	0.024	0.013	0.016	0.04	cluster_3
Row2	abuse	0.001	0	0	0.137	0.001	0.002	0.003	cluster_1
Row3	academy	0.007	0.021	0.007	0.007	0.007	0.093	0.036	cluster_2
Row4	accept	0.008	0.007	0.001	0.049	0.019	0.025	0.038	cluster_3
Row6	accounting	0.010	0.018	0.010	0.018	0.054	0.009	0.010	cluster_4
Row10	active	0.002	0.036	0.002	0.026	0.05	0.026	0.012	cluster_3
Row11	activities	0	0	0.001	0	0	0	0	cluster_6
Row14	administrator	0.036	0.036	0.036	0.036	0.036	0.036	0.107	cluster_0
Row15	adobe	0.036	0.036	0.036	0.036	0.107	0.036	0.036	cluster_5
Row17	adults	0.014	0.043	0.005	0.033	0.014	0.024	0.033	cluster_3
Row18	advanced	0	0	0.001	0	0	0	0	cluster_6
Row19	advantage	0.024	0.028	0.002	0.015	0.032	0.015	0.037	cluster_3
Row20	adventure	0.002	0.023	0.002	0.014	0.068	0.017	0.026	cluster_5
Row22	affairs	0.071	0.013	0.004	0.004	0.013	0.038	0.021	cluster_3
Row25	agreed	0.018	0.018	0.003	0.048	0.013	0.023	0.033	cluster_3
Row26	ahead	0.002	0.019	0.001	0.032	0.019	0.015	0.058	cluster_3
Row30	alaska	0.014	0.071	0.014	0.014	0.043	0.014	0.043	cluster_0
Row32	allow	0.001	0.009	0.001	0.026	0.026	0.030	0.047	cluster_3
Row33	amateur	0.018	0.054	0.018	0.054	0.018	0.018	0.054	cluster_0
Row36	americans	0.042	0.03	0.006	0.03	0.018	0.018	0.03	cluster_3
Row37	amounts	0.018	0.006	0.006	0.006	0.077	0.018	0.042	cluster_5
Row39	analysis	0	0	0.001	0	0	0	0	cluster_6
Row40	andrew	0.012	0.083	0.012	0.012	0.012	0.012	0.06	cluster_0
Row41	anime	0.012	0.083	0.012	0.036	0.012	0.036	0.012	cluster_0
Row43	announced	0.017	0.01	0.003	0.037	0.01	0.071	0.01	cluster_2
Row45	answer	0.005	0.023	0.001	0.044	0.013	0.026	0.035	cluster_3

Fig. 4 Interactivity Chart

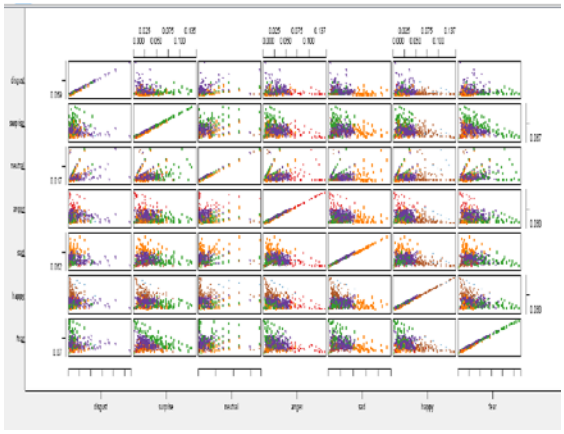


Fig. 5 Scatter Matrix

Above interactive table shows that words are categorized into seven different colours hence seven different clusters, each cluster defining an emotion.
The Scatter matrix in the Fig. 5 shows a correlation between two different emotions.

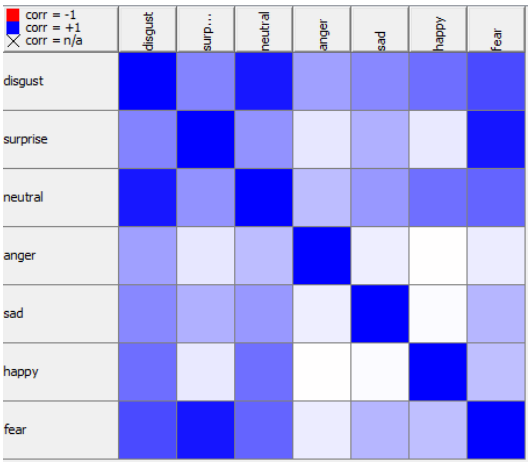


Fig. 6 Correlation Matrix for K-means

Fig. 6 shows another interactivity to show a correlation between different sentiments. The blue colour indicates that they are positively correlated.

Table 2 is the output of K-means showing several bits of information, to determine how well a k-means clustering is using SSE (sum of Squared Error).

K-means Statistics using R	
The total sum of squares***(totss)***	2.627687
Vector of within-cluster sum of squares, one component per cluster***(withinss)***	0.15837968 0.11831304 0.08208716 0.12798260 0.04518248 0.32856826 0.26968245
Total within-cluster sum of squares, i.e. sum(withinss)- ***(tot.withinss)***	1.130196
The between-cluster sum of squares, i.e. $totss-tot.withinss$.-***(betweenss)***	1.497491
(betweenss / totss)	57.0%

SSE is defined as:

$$SSE = \sum_{1} n(x - \bar{x})$$

Formula (2)

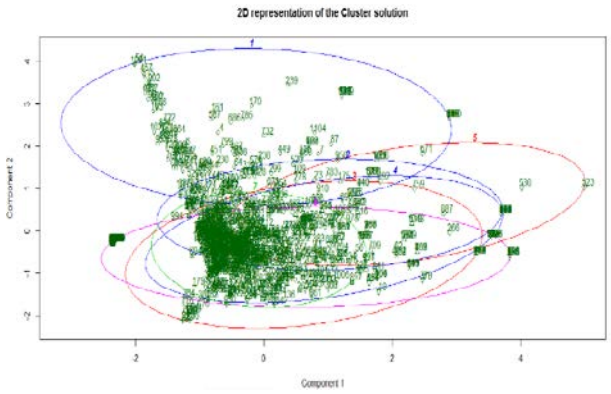


Fig. 7 clusters using R

The two components in Fig. 7 explains 45.27% of point variability.

MODEL B: FUZZY C-MEANS

A. Fuzzy C-means Algorithm

Traditional clustering approaches generate partitions; in a partition, each pattern belongs to one and only one cluster. Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function. The most widely used clustering algorithm implementing the fuzzy philosophy is FCM, initially developed by Dunn and later generalized by Bezdek. Despite this algorithm proved to be less accurate than others, its fuzzy nature and the ease of implementation made it very attractive for a lot of researchers that proposed various improvements and applications. The basic structure of the FCM algorithm is discussed below. The Algorithm FCM is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. It is based on the minimization of the following objective function:

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. *If* $|| U^{(k+1)} - U^{(k)} || < \varepsilon$ *then STOP; otherwise return to step 2.*

The algorithm comprises of following steps:

B. Knime Workflow for Fuzzy C-means

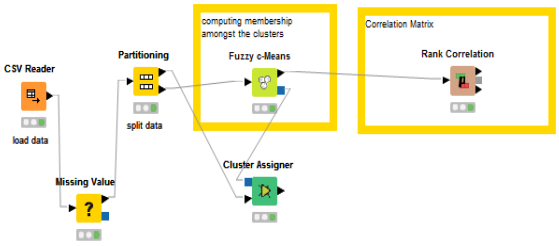


Fig. 8 C-means workflow with pre-processing and sampling

The data is first pre-processed by removing any record consisting a missing value it is then splitted into train and test data-70% 30% respectively. C-means algorithm is runned on the training set. When the dataset is trained with the C-means algorithm, it is assigned to the Cluster Assigner which caters the new data points and is used to test the model assigned to the assigner node.

C. Data Represention for C-means

Table 3

Key Chart for Knime visualization	
Cluster 0	Disgust
Cluster 1	Surprise
Cluster 2	Neutral
Cluster 3	Anger
Cluster 4	Sad
Cluster 5	Happy
Cluster 6	Fear

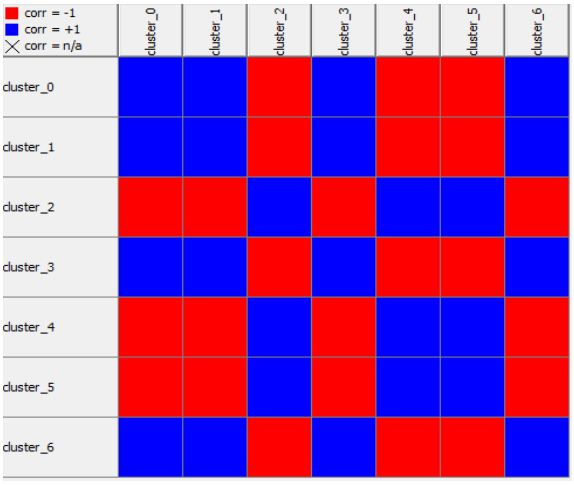


Fig. 9 Rank Matrix for C-means

The Rank Matrix in the Figure 11 shows the correlation between two clusters. Red color defines negatively strong correlation while Blue colour indicates the positively strong correlation.

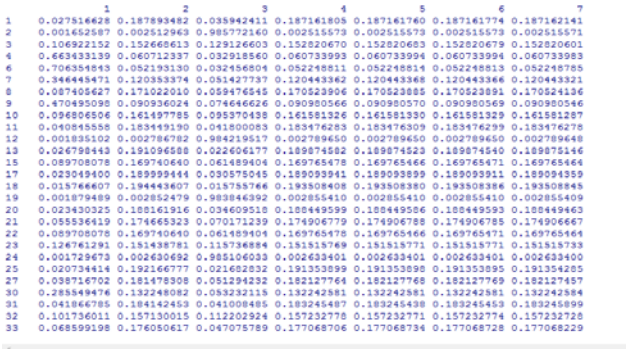


Fig. 10 Snapshot of R workflow showing membership of few rows against each cluster.Membership of cluster 6&7 are almost same; can be merged.

Table 4 shows the output of C-means clustering is using SSE (sum of Squared Error)

C-means Statistics using R	
The value of the objective function.- **(withinerror)**	0.0004685947

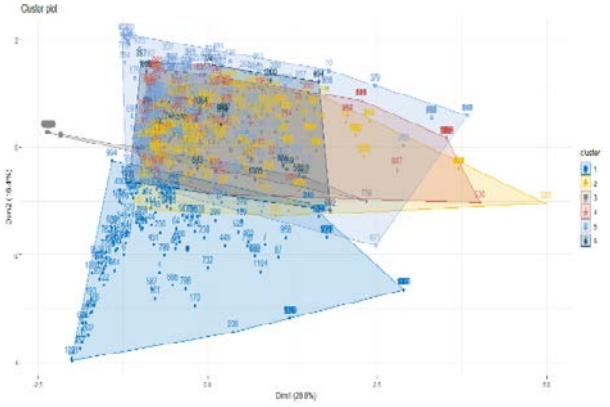


Fig. 11 Cluster Plot for C-means using R

Since the membership percentage of cluster 6 and 7 is almost same according to the membership table in Fig. 10, they merged in one cluster as cluster 6.

6. Analysis

Classification of emotions from reviews or blogs is a challenging task and this project analyzes the words with respect to categories of emotion. The models assign the words to emotions and help the websites or product owners to improve both sales and marketing strategies as the model helps them to predict the emotion of the user when he/she is expressing their reviews/feedback.

Approaches used by this project are independent of any datasets, instead, are more generalized and can be applied on different but normalized datasets.

This means the models that have been created in for this project can work for any other dataset related to sentiment analysis. The models of this project are based on k means and fuzzy c-means techniques. K means was found to be extremely fast than fuzzy cmeans in datasets that contain the clusters scattering in different patterns. Fuzzy c- means is a technique based on iterative fuzzy computations, so its performance was found relatively higher than expectations.However, the clustering failure of FCM and KM was found nearly equal.

7. Conclusion

This project presents an effective approach to an online reviews classification system based on Twitter, blogs and social media content. The number of reviews on the web are available from various sources. Instead of taking the entire review as the input, the system takes only a few words that reflect the emotion of the user. After the words have been extracted from the reviews, they are assigned to the basic emotions. Fuzzy C-Means (FCM) also known as soft clustering is better than K-Means (hard clustering) in term of the accuracy of clusters for emotion analysis. As a final conclusion, there is no particular algorithm which is the best for all cases. The datasets should be carefully examined for shapes and scatter of clusters in order to decide a suitable algorithm.

In future, this project can be extended to emotion classification with respect to user's personal information .

References

[1] Minho Kim, Hyuk-Chul Kwon(2011). Lyrics-based Emotion Classification using Feature Selection by Partial Syntactic Analysis, 2011 IEEE

[2] Oscal T.-C. Chen, Jhen Jhan Gu, Ping-Tsung Lu and Jia-You Ke(2012). Emotion-Inspired Age and Gender Recognition Systems, 2012 IEEE

[3] Jaskaran kaur , Sheveta Vashisht (2012) Analysis and Indentifying Variation in Human Emotion Through Data Mining, 2012 IJCTA.

[4] Shenghua Bao, Shengliang Xu September (2012). Mining Social Emotions from Affective Text, 2012 IEEE

[5] Sivaraman sriram, xiaobu yuan (2012). An enhanced approach for classifying emotions using customized decision tree algorithm, 2012 IEEE

[6] Ali, MA, Karmakar, GC & Dooley, LS 2008 'Review on Fuzzy Clustering Algorithms'. IETECH Journal of Advanced Computations, vol. 2, no. 3, pp. 169 – 181M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7] Bora, DJ & Gupta, AK 2014 'A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm'. Int. J. of Computer Trends and Technology, vol. 10, no. 2, pp. 108-113.

[8] Bezdek, JC 1981, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York., doi: 10.1007/978-1-4757-0450-1

[9] Ghosh, S & Dubey, SK 2013 'Comparative Analysis of K-Means and Fuzzy CMeans Algorithms'. Int. J. Advanced Computer Science and Applications, vol. 4, no.4, pp. 35-39.

[10] Gionis, A, Mannila, H & Tsaparas, P 2007 ‘Clustering Aggregation’. ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no.1, pp. 1-30.

[11] Di Martino, F & Sessa, S 2009 ‘Implementation of the Extended Fuzzy C-Means Algorithm in Geographic Information Systems’. J. of Uncertain Systems, vol. 3, no. 4, pp. 298-306

[12] Dunn, JC 1973 ‘A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters’. J. of Cybernetics, vol.3, no.3, pp. 32-57., doi: 10.1080/01969727308546046

[13] Borgelt, C & Kruse, R 2005 ‘Fuzzy and Probabilistic Clustering with Shape and Size Constraints’. Proc. of the 11th Int. Fuzzy Systems Association World Congress (IFSA'05, Beijing, China), pp. 945-950.