IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 5, May 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

| DOI: 10.15680/IJIRCCE.2021.0905118 |

# Tracking the Users' interest on Social Media

**Ankita Mandore[1], K.P. Adhiya[2]**

PG Student, Department of Computer Science, SSBT's College of Engineering & Technology, Bambhori, Jalgaon, Maharashtra, India[1]

Associate Professor,Department of Computer Science, SSBT's College of Engineering & Technology, Bambhori, Jalgaon, Maharashtra, India[2]

**ABSTRACT:** Social media are collaborating computer-mediated skills that facilitate the creation&sharing of data, ideas, career interests&other sorts of expression via virtual communities&networks. Social media analytics is that the process of assembly&evaluating data from social networks like Facebook, Instagram,&Twitter. Tracking the interest of user on social media is that the approach of collecting data from social media sites&blogs&evaluating that data to group the users consistent with their interest. This process goes beyond the standard monitoring or a basic analysis of retweets or" likes" to develop an in-depth idea of the social user. Dynamically clustering users within the context of streams of short texts may be a challenging task. To effectively infer users' dynamic interests, we evaluate the performance of existing collaborative strategies for clustering in documents or short text streams. We not only intend to enhance the performance of clustering strategies on document or short text stream but also to perform clustering supported user's location information for getting more relevant results.

**KEYWORDS**:UCIT, Clustering,ShorttextStreams,collaborative,bisectk-Means

## I. INTRODUCTION

The advent of online social networks has been one among the foremost exciting events during this decade. Many popular online social networks like Twitter, LinkedIn,&Facebook became increasingly popular.Social media analytics is that the process of gathering&analyzing data from social networks like Facebook, Instagram,&Twitter. Viewing consumers social media activity because the voice of the buyer. There are three main steps in analyzing social media: data identification, data analysis,&knowledge interpretation. Social networking sites allow users to share ideas, digital photos&videos, posts,&to tell others about online or real-world activities&events with people in their network. Clustering quality is very counting on how clean the info is, which is employed for clustering or categorization. Generally, the words or terms just like the,&don't support the cluster quality anyway. in order that the stop words are being faraway from the corpus or document before clustering. [1] Also, additionally to the present, there are a couple of transformation methods used for clustering the text documents. Clustering/segmentation is one among the foremost important techniques utilized in Acquisition Analytic. once you want to research the Facebook/Twitter/You tube comments of a specific event, it might be impossible to manually check out each&each mention&see where the sentiment regarding a specific brand/event/person lies. Due to necessity of enthusiastically clustering users inside the situation of short text streams mentioned in the comments on social medias, Clustering based on the user behavior technique is used**.** User clustering technique is interesting whenthe documents to be studied are long. To pact with this delinquent, it is recommended to follow a user clustering topic model (UCT).

## II. LITERATURE SURVEY

ShangsongLiang et.al. [1] presented, investigational outcomes proved that the representations are able to cluster users in short texts streams.The searches done by a person on social media over a long time are followed for improving the output of UCT.

Akanksha Kapoor et al. [2], presented partition-predicated clustering techniques, such as K-Means, K Means++ & object predicated Fuzzy C-Means clustering algorithm.

Yan Wu et al. [3], The problem of low efficiency & low accuracy posed by the fact that the influential users' interests will change during the event evolution, to address this problem, they proposed a user-interest model-based event evolution model, named HEE model.

Viktor Hozhyiet al. [4], they proposed a method for getting better clustering results by application of sorted & unsorted data into the algorithms.

Yukun Zhao et al. [5], they monitored all the activities of users on the social media&grouped them according to the same actions.

Ahmed Alsayat et al. [6], this paper, proposed a novel method to analyze social media data. Our method used K-Means algorithm along with Genetic algorithm&Improved Cluster Distance method to cluster the social media community based on headship, follower&attitude scores.

Kuldeep Singh et al. [7], they efficientlyfoundthecommonsearches of users by suggesting2methods of Gibbs sampling.

W. Chen, et al. [8], in their article Clustering Text Data Streams extended the semantic smoothing model into text data streams context firstly.

P. Xie et al. [9], in this paper, the problem of user interests mining from own photos is studied&proposed a User Image Latent tents.

Dibya Jyoti Bora et al. [10], presented an efficient method of combining the restricted filtering algorithm & the greedy global algorithm & use it as a means of improving user interaction with search outputs in information retrieval systems.

Yi Jiang et al. [11], discussed measurable, analytical representations of a consecutive corpus, active theme models offer a qualitative space into the fillings of a big document group.

X. Yan et al. [12], presented knowledge derived during pre-processing of documents for Clustering. Input is document vectors.

R. W. White et al. [13], they observed queries, clicks, scrolling,&text importance for millions of queries on the Bing marketable search engine to better understand the impact of user, task,&user-task interactions on user performance on search result pages (SERPs).

Vivek Kumar Singh et al. [14], This paper proposed the experimental work on applying K-means, heuristic K-means & fuzzy C-means algorithms for clustering text documents.

K. A. Abdul Nazeer et al. [15], this paper presents a experimental optional of the k-means algorithm which combines a heuristic method for result the initial centroids&an effective method for assigning the data points to the clusters.

Beil et al. [16], proposed a clustering algorithm based on term frequency called Frequent term-based text clustering. The method identifies a set of terms from the corpus & for each term extracted from the document it computes the term frequency.

Pankaj Jajoo [17], in this thesis he examined many existing algorithms&proposed two new ones. They achieve that it is barely conceivable to get a general algorithm, which can work the best in clustering all types of datasets.

## III. METHODOLOGY

Many people have already worked on the issue of fragmenting documents into differentgroups to present a genuine data that can be useful. [12, 13]. The classification of documents into different groups is not an easy task, but a problematicone. We need to segregate the similar kinds of data within the full information. Similarity function may be used to compare the items & get the final output as segregated data that can be useful to us.[11, 23].

Social networking sites allow users to share ideas, digital photos & videos, posts, & to inform others about online or real-world activities & events with people in their network.

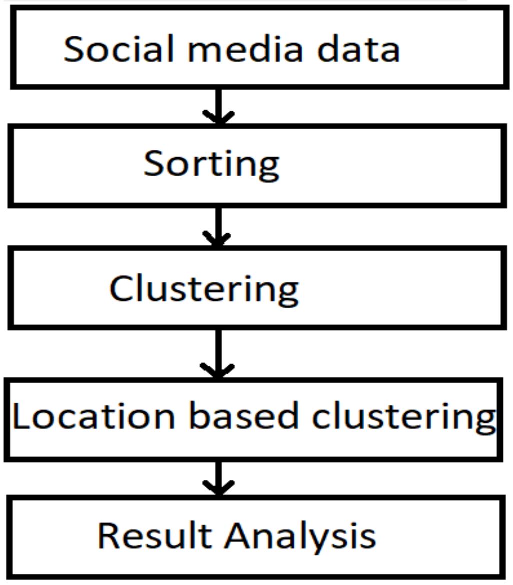

Figure. 1: System architecture of the proposed work

Figure 1 shows the system architecture for the proposed work in which we collect data from the social media platform, sort data by removing stop words, tokenization etc., perform clustering on the basis of users' choices and also find the location wise interest of the user. Finally, we analyze the result.

Using Social Media, connecting has become so easier that anybody can connect to anybody else in any part of the world at any time. Clustering quality is highly depending on how clean the data is, which is used for clustering or categorization. Generally, the words or terms like the, & do not support the cluster quality anyway. So that the stop words are being removed from the corpus or document before clustering. [1] Also, in addition to this, there are a few transformation methods used for clustering the text documents. To analyze social media comments of any events with a large userbase like Facebook, it is practically not possible to manually look at each & every mention.

Calculable data was the main part of work done in the conventional ways of clustering [14,15,16]. The data was in the form of numbers. But in recent times, Categorical data is taken into consideration. Its data is different from than that of numbers.

**Uses of Clustering Methodology:**

- **Managing Files & Web Searching:**

When a person searches for a document online, if it's in well-organized manner, the browsing becomes very easy & one can easily finds what needed to him. It organizes the documents into different types. Scatter method is one of the best tools for this.

- **Summary of Corpus:**

This method gives a short description about all the lengthy information within eitherword-clusters [17, 18] or cluster digests [8].This information may be utilized for giving a short description about the big information.Different forms of this technique are utilized for giving out a short description for a big file.

- **Categorization of files:**

Clustering technique needs no presence of any human interference. It doesnot require any supervision. But it's used to enhance the standard of the leads in case of under supervision.co-training [7] &word-clusters [17, 18] techniques are often utilized for enhancing the correctness of grouping.

Any text document is always indicated in the binary form in digital language. A binary array is made from the text in any file. Then categorical data clustering algorithms [1, 4] is applied onto it. A better indication of the groupings of different words within a file will then be given as output. It will also consist of repetency of different words in that file

& also in the entire worksheet, like Term Frequency &Inverse Document Frequency [8]. Calculative methodology [4, 7, 18] ismostly utilized here for improving the results. This is done so as to work out the foremost relevant groups of objects within the data.

### 1. Stop Words Removal:

The output of all the methodologies & technique in clustering is very hooked in to the effectiveness of attributes which will be utilized during the grouping method.Many words which are used almost everywhere won't be of any help for better results during clustering. For instance, words like "a", "an", "the". These words come in almost all the English statements. Hence, we have no use of such words for clustering. Hence, it becomes very important first step to choose attributes smartly. These repetitive words must be removed initially only & should not become the part of clustering. In Dimension Reduction Methodology, the inter-relation of different words is utilized. It helps us in better attributes to give main idea of the dataset.

### 2. Technique for Selecting Features:

For making category of texts, this technique is most widely used[99].This is a supervised technique. For deciding the attributes, a manual supervision is needed here. Though there already exist many options where no supervision is needed as such & they are simple to use as well, supervised ones will give better results. Few samples for this technique are as follows:

#### 2.1. Selecting a file on Frequency Basis:

This is the easiest technique available for our desirable output. Here, we utilize the frequency of different files & remove the content not relevant for us.Though IDF technique makes sure that those words in the file won't affect our results, having a support system is always a better option. As due to the repetitiveness of words to a larger extent, IDF single handedly may be inefficient to reduce the effect due to higher frequency of few words. It eliminates the most widely used common words of English like "for". As these words won't be useful for us from the point of grouping the file. Stop words is another name for these. Many techniques are already given for this purpose [6].Unformatted files taken directly from websites or social media reports will contain stop words more frequently than already well worked files. In case if a word is not able to distinguish among the clusters & it is still coming more frequently, then also it must be eliminated as it won't be of any use for us. The term frequency – inverse document frequency technique also eliminates few most repetitive words. Hence, it is always beneficial to have a good storage of these stop words as it saves us a lot of time & economic cost also. We need to include numbers to the texts for arranging them in the order of their relevance for us. Some of the techniques for high level of pruning are as follows:

##### 2.1.1. Term Strength:

This method comparatively gives the better results for pruning as compared to other methods [9]. Central theme of this methodology is applying those methods used under supervision to the non-supervision. This method calculates the usefulness of the given word to find out the relation between 2 or more different files.For an instance, assume 2related files asm&n respectively, then, the formula for calculating term strength s(t) of term t:

$$s(t) = P\ (t \in n | t \in m)$$

Here, P means probability.The big problem in this method is the logic behind the decision to say that file m & file n is related.It means we are usingthe supervised process for this. This method will be applicable when we already have a well mentioned category of dataset or files. The opposite side of this is when there will be a big set of data & files, we would be unable to decide which combinations are related & which are not. There is higher probability of making mistakes there if we do it on our own. That's why we need to automated system for this to make the work faster & reduce the errors. This method should also not be under supervision. We need a definition for creating a formula to say when a combination is related. This is given in [4].We can use the automatic well pre-defined formulas which itself will decide whether a combination is related or not. Here, s(t) is calculated by making unplanned samples of different combinations of the files from data set:

$$s(t) = \frac{No.of\ combinations\ in\ which\ t\ comes\ in\ both}{No.of\ combinations\ in\ which\ t\ comes\ in\ 1st\ combination}$$

In this case, we choose a primary pair in an unplanned manner. so as to reduce the attributes, the term strength could also be equated with that of our desired one,those were assigned in an unplanned manner in the training file but had almost equal frequency.

When term strengthisn't a minimum of 2sdmoreas compared to that for a word selected in an unplanned way, then it's faraway from the gathering.Here, we can start selecting feature in an unsupervised manner. This is the benefit we have in this method. While in other methods, we need to supervise in the early phase of selecting the feature, even when the actual working happens in unsupervised manner. We can obviously use this methodology for selecting a feature in categorization [1] as well as that in supervised clustering [4]. We can even us the same method when we have data of training with us.When we use this methodology for selecting an attribute, we notice that this methodology does not work well for clustering based on similarity. The reason behind this is that this method is based on thought of similarities within a file itself.

### 2.1.2. Entropy-based Ranking:

This method has been discussed in [2]. Here, when we eliminate a term, we calculate the amount of decrease in its entropy & accordingly decide the standard of it. The formula for the same has been given below:

$$E(t) = -\sum_{i=1}^{n}\sum_{j=1}^{n}(S_{ij} \cdot \log(S_{ij}) + (1 - S_{ij}) \cdot \log(1 - S_{ij}))$$

Where,
E(t) = Entropy of the term t during a collection of n documents.

Also,$S_{ij}$∈ (0, 1) gives us the similarity within the file no. i& jwhen we have eliminated the term t.
The calculation for $S_{ij}$is given by:

$$S_{ij} = \frac{2-dist(i,j)}{dist}$$

Where,
dist(i, j) = distance of 2 terms I & j when we eliminate the term t.
&dist= mean distance of files when we eliminate the term t.

It can be easily seen here that for calculating entropy for each term, we need O($n^2$) steps, which is almost not feasible when there is big data. In [2], it is already discussed aboutthe use of this approach along with the sampling techniques for giving us a better output.

### 2.1.3. Term Contribution:

This idea [6] is predicated on the very fact that the output during the process of text clustering are very much hooked to the similarity in file. Hence, we need to look at the benefaction of a term as the benefaction for similarity in the file. For instance, when we applysimilarity based on scalarproduct,it can be calculated as the scalar product of normalized frequencies. It gives us term’s contribution. Summation across pairs is done for calculating the team contribution. Similar to the earlier one case, this technique also needs O($n^2$) of period of time for each term.To fasten the process, sampling technique is also utilized. The disadvantage of using this technique is that it does not have a precise distinguishing ability. It may give an advantage to most repetitive words during grouping step.Though this technique has advantage of making it unsupervised, but here, there is a chance of biasness during the selection of the term. This might be due to our earlier assumption in the similarity function that this bias may come. This technique might give us different output if we use another similarity function. This means we need to be very careful while selecting a proper similarity function. As our final output is fully dependent on it.

### 2.2. Hierarchical Clustering:

To maintain records of various needs, we use this function. A summary of the normal agglomerative&hierarchical clustering algorithms has been given in [6,9]. The comparative analysis of few techniques has also mentioned in [11]. The advantage of this method is that it automatically makes a structure like a tree. Therefore, it benefits a lot of other searching techniques. This gives us a thumbs up over other during the search.This technique continuously joins the files on the basis of similarity. In the similar fashion, it also connects groups. All these different methods are differentiated on the basis of calculation of similarity. We may classify similarity as best, average or worst case. When we bundle the files continuously one over another creating levels, it forms a hierarchical structure. Here, leaves indicate separate files& nodes indicate merged ones. If we merge 2 groups, then it makes a new node.It’s 2 branches showthose 2 groups.

### 2.3. Clustering Algorithms:

#### 2.3.1. Hybrid Algorithm:

1. Calculate similarity among all clustercombinations.
2. Combine2 clusters which are most similar to each other.
3. Now, redo similarity matrix that will show combination-wise similarity among original & new cluster.
4. Keep on doing the above 2 steps still all but 1 cluster vanish.

#### 2.3.2. K-means Algorithm:

1. Choose K points indicating initial centroids.
2. Allot all the points to nearest centroid.
3. Calculate centroid for every cluster again.
4. Keep on doing the above 2 steps still centroids become constant.

## IV. RESULT & DISCUSSION

For Performance evaluation of the approach we measure it based on 2 parameters i.e precision & recall. Precision & Recall are well-defined in terms of a set of repossessed documents (e.g. the list of documents produced for a query) & a set of relevant documents (e.g. the list of all documents that are relevant for a certain topic). We have calculated the Recall & Precision using the below formulae:

Precision (i, j) = $\frac{nij}{\mathrm{nj}}$ & Recall (i, j) = $\frac{nij}{\mathrm{ni}}$

Where,
$n_{ij}$=No. of the members of cluster jin class i,
$n_j$= No. of the members of cluster j
$n_i$= No. of the members of class i.

F scoreis then calculated using the below given formula:

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}$$

| Methods | Precision | Recall | F-Measure |
|---|---|---|---|
| K-Means | 0.256 | 41352 | 120179 |
| LDA | 0.31 | 50174 | 111537 |
| UCIT Existing | 0.594 | 95973 | 65558 |
| Proposed | 0.627 | 101283 | 60248 |

Table 1: Comparison of Performance

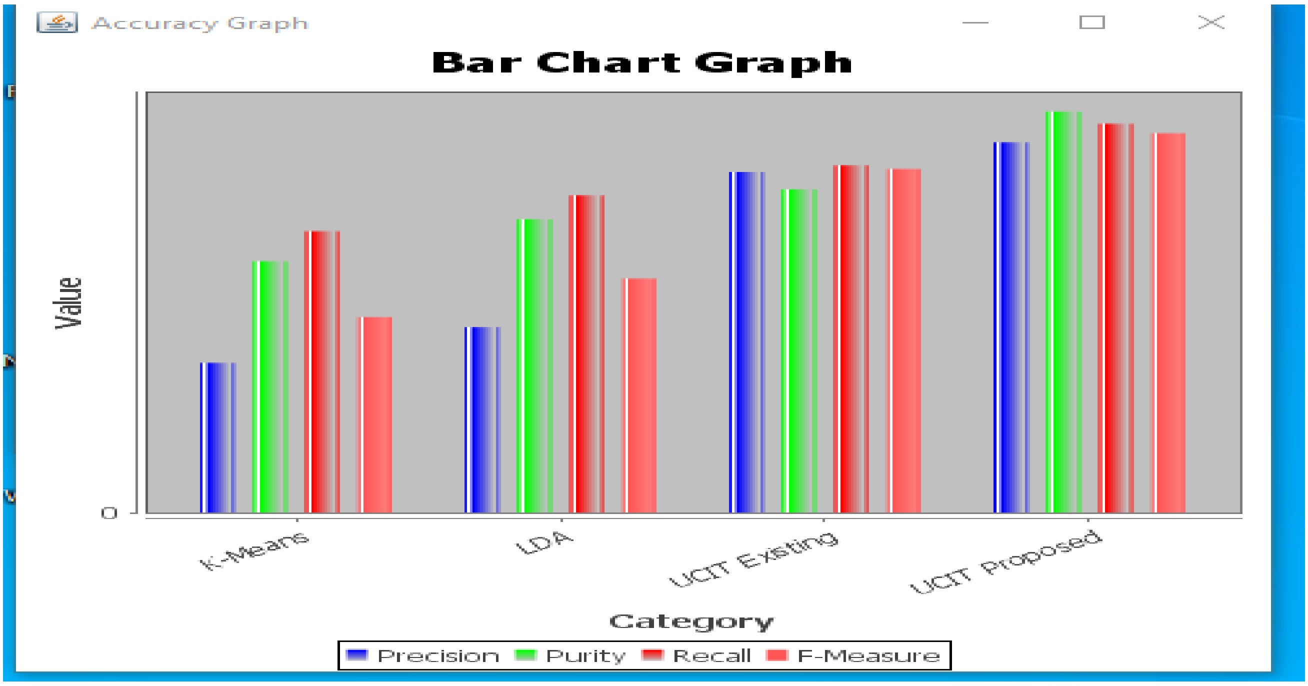

Figure 2:Graphical Representation of evaluated Results

Above results shows the difference between various methods implemented, out of which k-means & LDA has very less precision or accuracy. Existing system UCIT implemented has better results than the k-means & LDA due to collaborative clustering.

Enhanced UCIT has been proposed with possible updates of using se- mantic words calculation&generating clusters based on semantic words&collaborative clustering. It shows improvement in results&also improved the accuracy to understand&generate clusters based on similarity.

## V. CONCLUSION

Here, we have tried to analyze the issue of dynamically clustering users with respect to the short text streams. UCIT model was presented here.This technique has the ability to draw inferences & then find every user along with its follower's ever-changinginterests on social media for the purpose of user clustering. The UCIT will be able to manage not only textual sparsity of short documents but also ever-changingdemands of users'along withtheir followers' interests. That too, over a long period of time. The evaluation of effectiveness of this technique was also done here. The parameters for this evaluation included generalization effectiveness, topical representation as well as clustering.We also compared this methodology with state-of the-art models. With the help of user text stream with semantic analysis&location information, it would be helpful to generate more accurate&more relevant clusters groups based on user interest. Through experiments, we can conclude that UCIT is able to cluster the users in short text streams more efficiently.

## REFERENCES

[1] Shangsong Liang, Emine Yilmaz,&EvangelosKanoulas,"*Collaboratively tracking interests for user clustering in streams of short texts*" IEEE transactions on knowledge&data engineering, VOL. Xx, no. Xx, month 2018

[2] Akanksha Kapoor&Abhishek Singhal *"A Comparative Study of K- Means, K-Means++&Fuzzy C- Means Clustering Algorithms*"3rd IEEE International Conference on "Computational Intelligence&Communication Technology" (IEEE-CICT 2017)

[3] Lei-lei Shi, Lu Liu, Yan Wu, Liang Jiang, James Hardy, "Event Detection&User Interest Discovering in Social Media Data Streams,", IEEEAccess, DOI 10.1109/ACCESS.2017. pp 2675839

[4] Viktor Hozhyi&Bart Lamiroy, "Clustering of Users in Social Networks by Their Activity" IEEE First Ukraine

Conference on Electrical&Computer Engineering (UKRCON) ©2017 IEEE pp 978-1-5090-3006-4/17/

[5] Yukun Zhao, ShangsongLiangz, Zhaochun Ren,“*Explainable User Clustering in Short Text Streams.*”,16,Pisa, Italy c 2016 pp. 2911451.2911522.

[6] Ahmed Alsayat&Hoda El-Sayed, “Social Media Analysis using Optimized K-Means Clustering” 78-1-5090-0809-4/16/ copyright 2016 IEEE SERA 2016, June 8 10, 2016, Baltimore, USA

[7] Kuldeep Singh, Harish Kumar, Shakya Bhaskar Biswas, “Clustering of People in Social Network based on Textual Similarity Perspectives in Science” (2016), http://dx.doi.org/10.1016/j.pisc.2016.06.023

[8] W. Chen, J. Wang, Y. Zhang, H. Yan,&X. Li, “User based aggregation for biterm topic model” IEEE Trans. Pattern Analysis&Machine Intelligence” In ACL, 2015 pages 489494.

[9] P. Xie, Y. Pei, Y. Xie,&E. Xing, “Mining user interests from personal photosn” in AAAI, 2015, pp. 18961902.

[10] Dibya Jyoti Bora, Dr. Anil Kumar Gupta, “Performance of K-Means Algorithm: An Experimental Study in Matlab” International Journal of Computer Science&Information Technologies, Vol. 5 (2), 2014, pp 2501- 2506.

[11] Yung-Shen Lin&Yi Jiang, “A Similarity Measure for Text Classification&Clustering” July 2014IEEE Transactions on Knowledge&Data Engineering 26(7):1575-1590 DOI: 10.1109/TKDE.2013.19

[12] X. Yan, J. Guo, Y. Lan,&X. Cheng, “A Biterm Topic Model for Short Texts”, May 13–17, 2013, Rio de Janeiro, Brazil. ACM 978-1-4503-2035-1/13/05.

[13] G. Buscher, R. W. White, S. Dumais,&J. Huang, “Large-scale analysis of individual&task differences in search result page examination strategies” In WSDM, ACM, 2012 pages 373382

[14] Vivek Kumar Singh&Nisha Tiwari, Shekhar Garg, “*Document Clustering using K- means, Heuristic K-means&Fuzzy C-means*”International Conference on Computational Intelligence&Communication Systems Un. 2011

[15] K. A. Abdul Nazeer, M. P. Sebastian&S. D. Madhu Kumar, “A Heuristic k-Means Algorithm with Better Accuracy&Efficiency for Clustering Health Informatics Data” Journal of Medical Imaging&Health Informatics Vol. 1, 2011, pp 66–71.

[16] Blei, Ng,&Jordan Latent,“*Dirichlet Allocation Journal of Machine Learning Research*”Research 3 (2003) pp 993-1022

[17] Pankaj Jajoo, “*Document Clustering*” Indian Institute of Technology Kharagpur IEEE Trans. Neural Networks, vol. 14, no. 1, Jan. 2003, pp. 195-200.

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING