An Improved K-means based Text Document Clustering using Artificial Bee Colony with Support Vector Machine

Manpreet Kaur
manpreet16005@gmail.com
M.tech scholar, Department of CSE, IK Gujral punjab
technical university,jalandhar

Abstract: *The usage unsupervised clustering approaches in numerous applications has attracted the researchers as an innovative research area due to separation characteristics of unstructured format of documents in a similar group* contains *same data. Increasing requirement in numerous research fields and information technologies, led to an increase in the document clustering approaches with maximum efficiency. Therefore, in exiting works, researchers take a lot of time to find similar documents according to the requirement. During the document clustering, several types of issues faced by the researchers like best centroid localization, less uniqueness of clustered data, high clustering time and many more. To minimize these types of problems, most of the previous research focuses on similarity techniques for clustering but the clustering results are not satisfactory. In this research article, we developed an improved K-means based text document clustering using Artificial Bee Colony (ABC) along with the Support Vector Machine (SVM) as a concept of unsupervised mechanism. Here, SVM act as a machine learning technique and helps to classify the best centroid according to optimized centroid by K-means with ABC. Based on this mechanism, proposed work achieved better clustering performance that is validated by comparing with existing works based on the Accuracy, Normalized mutual information (NMI) and Adjusted Rand index (ARI).*

Keywords**:** *Document Clustering, Data Mining, Improved K-means, Artificial Bee Colony Similarity, Support Vector Machine, Performance Parameters.*

I. INTRODUCTION

There are lots of applications of documents clustering in the industrial or commercial real world and their goal is to provide high accuracy and low error rate. In the document clustering, weight calculation is a major factor to achieve better and accurate accuracy with minimum error. There are several researchers that have proposed different techniques for the weight extraction from the documents but in the weight calculation, the chances of similar weight are more during the weight extraction. If the weight is same for different types of documents, the chances of error is more, so, for the weight extraction and documents clustering, use of the supervised learning with unsupervised clustering techniques will be beneficial. Clustering refers to grouping of similar data and document clustering is one of the popular mechanism for researchers. Text document clustering is a technique of the unsupervised mechanism and aims to find out the usual grouping among text documents in such a way those documents within a cluster are similar to one another and are dissimilar to documents in other clusters. The process of clustering is shown in Fig. 1.

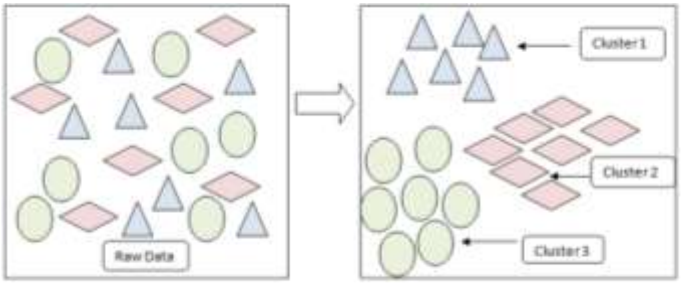


Fig. 1: Text Document Clustering Process

The basic process of clustering from raw text documents to the formation of the cluster is shown in the above figure. Here, we considered three different color of clustered documents representation like Red, Blue and Green with three different shapes. When unsupervised clustering algorithm is applied to the raw data, then data with similar features come under one cluster.

1.1 Motivation

Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. The goal is to create clusters that are coherent internally, but substantially different from each other. In plain words, objects in the same cluster should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in the other clusters. The following challenges gave us the motivation to use clustering of the news articles:

- ❖ The number of available articles was large.
- ❖ A large number of articles were added each day.
- ❖ Articles corresponding to same news were added from different sources.
- ❖ The recommendations had to be generated and updated in real time.

By clustering the articles we could reduce our domain of search for recommendations as most of the users had interest in the news corresponding to a few number of clusters. This improved our time efficiency to a great extent. Also we could identify the articles of same news from different sources. The main motivation of this work has been to investigate possibilities for the improvement of the effectiveness of document clustering by finding out the main reasons of ineffectiveness of the already built algorithms and get their solutions. Initially, applied the K-means and then Artificial Bee Colony (ABC) as an optimization method on the data and found that the results were not very satisfactory and the main reason for this was the noise in the data, created for the data. Thus, will tried for pre-processing of the data to remove the extra noises like stop words. Then applied ABC as a heuristic for removing the inter cluster and then applied the standard Support Vector Machine (SVM) for clustering to get much better results.

1.2 Contributions

Nowadays, the process of text document clustering is a basic requirement for lots of purpose like information retrieval, automatic topic extraction and document organization. High quality clustering algorithms plays a vital role efficiently in organizing, summarizing and navigating the unstructured documents. So, in this research we proposed an improved K-means based text document clustering using ABC with SVM and the main contributions are as follows:

- To study the existing documents clustering approaches with different algorithms.
- To develop a tokenization method in pre-processing steps with various additional data mining mechanism.
- To cluster the document, K-means will be used as an unsupervised approach using tokenized data of documents.
- To optimize the clustering output, Swarm-based ABC with SVM as classifier will be used with a novel fitness function.
- To validate the proposed document clustering model, performance parameters like Accuracy, Normalized mutual information (NMI) and Adjusted Rand index (ARI) will be calculated and compare with existing works.

This research article deals with improved K-means algorithms with a Meta heuristic ABC approach and SVM as a classifier to improve the text document clustering accuracy. The rest of this research article is systematized like: Segment 2 illustrates the survey of related work and Segment 3 describes the methodology of proposed document clustering mechanism, and its algorithms. Segment 4 presents the results and discussion based on the performance parameters whereas, Segment 5 discusses the conclusion with future direction text document clustering.

II. RELATED WORKS

A brief survey to analyze the existing work related to the proposed an improved K-means based text document clustering using ABC with SVM. There are lots of approaches already used to solve documents clustering problems using various approaches but in this section of research article, we focus to identify the challenging factors and problems regarding the proposed model. In 2019, ***R. Janani & Dr. S. Vijayarani*** had conducted a research on text document clustering using spectral clustering algorithm with Particle Swarm Optimization (PSO) and authors try to improve the text document clustering. By considering global and local optimization function, the randomization is carried out with the initial population. Main aims of research was combining the spectral clustering with swarm optimization to deal with the huge volume of text documents. The proposed algorithm SCPSO is examined with the benchmark database against the other existing approaches. The proposed algorithm SCPSO is compared with the Spherical K-means, Expectation Maximization Method (EM) and standard PSO Algorithm. The concluding results show that the proposed SCPSO algorithm yields better clustering accuracy than other clustering techniques [1]. After that, ***Qingchen Zhang et al.*** in 2019 had also conducted a research on secure weighted possibility C-means algorithm based clustering in big data for cloud environment. In this research, authors have designed an improved C-means algorithm based on the BGV encryption method for big data clustering. The properties of the proposed scheme with homomorphic encryption scheme have been integrated to improve the cloud computing clustering efficiency for big data without the disclosure of the private data. Main idea presented by authors is to use approximate functions for calculating the weight values and updating the membership matrix and the clustering centers as three polynomial functions. This leads to remove the division and exponential operations such that the presented scheme can obtain the correct clustering result on the encrypted data. Designed model experimental results have demonstrated that the clustering accuracy is good by using these concepts but the clustering time is high for experimental datasets. So, other encryption schemes such as homomorphic encryption schemes and garbled circuit have been investigated to implement the secure weighted possibility C-means algorithm, which is expected to further improve the clustering efficiency with fast and robust clustering without the disclosure of the private data on cloud [3]. ***Alguliyev et al.*** in 2019 have presented a two phase sentence selection system named as COSUM, which utilized clustering and optimization schemes. For clustering K-mean algorithm has been used to find out the topics in the given document. For the selection of relevant sentences from clusters, population based optimization approach known as Differential Evolution (DE) algorithm. The objection function based on the harmonic means of the selected sentences has been design. From the test results it has been concluded that using clustering with optimization scheme results in better sentence summarization [3]. ***Lazhar*** in 2019 has presented a fuzzy clustering approach for the identification of outlier document. During experiment, the researchers have assumed that the documents that are allocated to the clusters have similar values and are considered as applicant outlier. Initially, a semantic data model has been created by utilizing Doc2Vec structure and then fuzzy clustering algorithm has been applied. After that, applicant outlier document has been identified on the basis of various degrees of membership. At last, the objective function has been recomputed for every applicant outlier document and an applicant document is considered if the value of objective function seems to be increases. To determine the effectiveness of the designed algorithm, results have been examined on two different types of data set (i) real dataset and (ii) without outlier. The test results have indicated that without outlier the classifier provide better performance [4]. In 2018, ***Jose Maria Luna-Romera et al.*** have designed a technique to validity indices for clustering algorithms in Big Data. This research has been presented two clustering validity indices to handle the huge amount of data in small computational time. The proposed indices are based on redefinitions of traditional indices by simplifying the intra-cluster distance calculation. The experimental results have shown that both indices scheme can handle Big Data in a very low computational time with effectiveness similar to the traditional indices using Apache Spark framework. The authors have not considered the concept of other cluster validity indices that also obtained suitable results in their traditional version. Further research is needed to revise the outcomes because some of the big data cluster validity indices results were not enough clear [5]. ***Puja Shrivastava et al.*** in 2018 have conducted a research on Big Data clustering using the concept of expansion of K-means clustering algorithm. An expansion of K-means clustering algorithm is developed for Map-Reduce framework by using the concepts of genetic algorithm as an optimization approach. To overcome the problem of suboptimal solutions of K-means clustering algorithm the concept of chromosome formation, fitness calculation and crossover are used to reduce time complexity of K-means algorithm for big data. Designed improved algorithm by authors is not considering the concept of parent selection mutation steps to improve the performance time. Scientific part of this research is the novel scheme of utilizing the concept of genetic algorithm with K-means, which is an improve algorithm of computational intelligence. Implementation of proposed improved mechanism in multiple machine and comparison of it with existing Map-Reduce based on the K-means clustering algorithm is required to validate the designed work.

2.1 Issues in Existing Works

After studying existing research literature for text document clustering, the following points are highlighted as interferences drawn and verdict from the above state-of art as illustrated below:

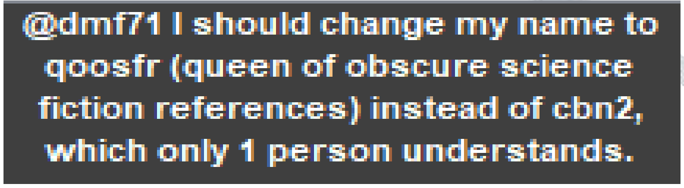
- In most of the existing work, the results obtained after clustering is not sufficient due to selection of cluster centroid.
- In previous work, due to lack of appropriate classification technique the classification of centroid is not acceptable. So, in proposed work, combination of clustering technique along with classifier will be proposed.
- In previous work, cosine similarity technique is used to find out the similarity between different types of text document. So, there is a chance to come out similarity index same for two or more than two text document. So, in proposed work, Euclidean distance based similarly, which return unique cosine similarity for each document.
- The data used in the previous work are standard and the researchers have not worked on user defined data. So, all existing works are non-real time model for text document clustering.

To solve such kind of problem in the proposed work, we use the concept of improved K-means using ABC with SVM to perform text document clustering and the methodology is described in the below segment of this article.

III. Model Methodology

In this section the designed and developed framework for the proposed an improved K-means based text document clustering using ABC with SVM are mentioned and the used steps for the development of the text document clustering is given as:

Step 1: Design a framework for text document clustering and upload documents datasets with different groups. The uploaded documents data is shown in the Fig. 2.

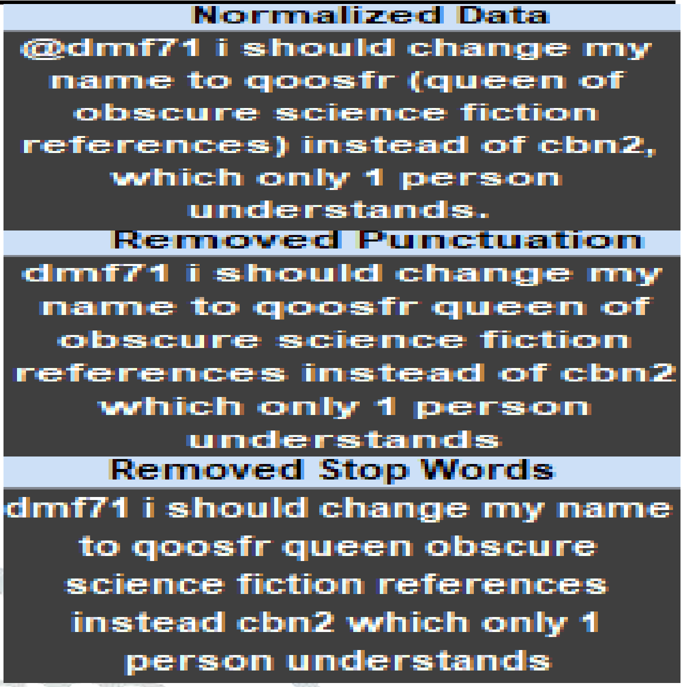


@dmf71 I should change my name to qoosfr (queen of obscure science fiction references) instead of cbn2, which only 1 person understands.

Fig. 2: Uploaded Text Document Data

Step 2: Apply pre-processing on uploaded text documents to convert text data into token value using tokenization method. In the pre-processing several basic steps are involved to make a data according to the requirements and these steps are very useful to achieve better clustering performance.

Step 3: In pre-processing, process of normalization is applied to normalize the uploaded text documents in same format and then remove the punctuations, stop words and at last calculate token value of text document data.



Normalized Data

@dmf71 i should change my name to qoosfr (queen of obscure science fiction references) instead of cbn2, which only 1 person understands.

Removed Punctuation

dmf71 i should change my name to qoosfr queen of obscure science fiction references instead of cbn2 which only 1 person understands

Removed Stop Words

dmf71 i should change my name to qoosfr queen obscure science fiction references instead cbn2 which only 1 person understands

Fig. 3: Pre-processing on Uploaded Data

Step 4: Develop a code for the finding the similarity between document's token values using the Euclidean Distance based similarity measurement technique. The algorithm of similarity measurement using Euclidean Distance is written as:

Algorithm: Euclidean Distance

```
Input: T-Data→ Document text as a raw data for similarity measurement
Output: ED-sim→ Euclidean Distance based similarity

Start
Create an empty array to store similarity, ED-sim = []
Data Sim-count = 0
For m = 1 → Length (T-Data)
    Current T-Data = Data (m)
        For n = m+1 →Length (T-Data)
            Calculate the distance
            DIST= SQRT| (Current T-Data (m) - T-Data (n)) $^2$|
            ED-sim [Data Sim-count, 1] = Current T-Data
            ED-sim [Data Sim-count, 2] = T-Data (n)
            ED-sim [Data Sim-count, 3] = DIST
            Data Sim-count = Data Sim-count + 1
        End – For
End – For
Return: ED-sim as an output in terms of Euclidean Distance similarity between data
End – Algorithm
```

Step 5: Apply K-means clustering approach with centroid for the clustering of text documents on the basis of similarity value and to validate the clustering next step will be used. The algorithm of K-means for text documents clustering is written as:

Algorithm: K-means

```
Input: T-Data→  Document Text Data for clustering
Output: C-Data and C→ Clustered Data and their centroids

Start
Initialize an estimated group (G)
Calculate size of Data in terms of [Row, Col.]
Define initial C-Data and random Centroid C = C1, C2…Cn
While clustering not performed
For i = 1→ Row
```

```
    For j = 1 → Col
        If Data (i, j) == C1
          C-Data 1=Data (i, j)
        Else if Data (i, j) == C2
          C-Data 2=Data (i, j)
          .
          .
          .
        Else Data (i, j) == Cn
          C-Data n=Data (i,j)
       End – If
       Adjust Centroid C using their mean
       C = Mean (C-Data 1, C-Data 2… C-Data n)
    End – For
End – For
End – While
Return: C-Data and C as a Clustered Data and their centroids respectively
End – Algorithm
```

Step 6: Initialize ABC to optimize clustering index and remove the unwanted text document from one to another cluster using their fitness. To check the fitness of document data a novel fitness function is design in ABC algorithm.

Algorithm: ABC for Clustering

```
Input: C-Data and C → Clustered Data and their centroids
Output: OC-Data and C → Optimized Clustered Data and their centroids

Start
To optimized the C-Data, ABC Algorithm is used
Set up basic parameters of ABC:
Population of Bee (B) – Number of Sensor Nodes
Defined Fitness Function:
```

$$F(f) = \begin{cases} 1; & \text{if } N_{PROP} < Threshold_{PROP} \\ 0; & Otherwise \end{cases} \quad \ldots (1)$$

```
In the fitness function, $N_{PROP}$ : is feature of current text data and
$Threshold_{PROP}$ is the threshold properties of all data which is
define on the basis of average values
Calculate Length of C-Data in terms of Len
Set, Optimized Clustered Data, OC-Data = []
For k = 1→ Len
     $E_{BEE}$ = C-Data (k) // Current Bee from B
     $O_{BEE}$ = mean ($E_{BEE}$) // // Mean of all B
     F (f) = fitness ($E_{BEE}$, $O_{BEE}$)
     OC-Data = ABC (F (f), C-Data (k))
 End – For
 Returns: OC-Data and C as an optimized clustered data and their centroids
 End – Algorithm
```

Step 7: Apply SVM on text document data according to the optimization to train the database using following steps:

- Select optimized feature set as an input of SVM for training and testing data.
- Compute the total categories which are generated by the training of optimized data using SVM.

Step 8: After that in the classification section, classify accurate cluster for the query text data according to the trained SVM structure.

Algorithm: SVM

```
Input: OC-Data → Optimized clustered data as training data and their centroids as a target or category
Output: SVM-Structure→ Trained SVM structure with Updated-C (Centroids)

Start
Initialize the SVM with training data OC-Data with RBF as Kernel function
For I = 1→ Length (OC-Data)
     If Cluster of C-Data (I) == C1
        Defined the Cat as a categories of training data
        Cat (1) = C-Data (I)
    Else
        Cat (2) = C-Data (I)
    End – If
End – For
SVM-Structure =SVMTRAIN (T, Cat, Kernel function)
Verify the Test Text Document (T-data) clustering
Updated-C = SVMCLASSIFY (SVM-Structure, T-Data)
If Updated-C == True
   Done!!!
Else
  Check Next T-Data
End – If
Return: SVM-Structure as a Trained SVM structure with Updated-C
End – Algorithm
```

Step 9: At last of the module, the performance parameters of proposed work like Accuracy, Normalized Mutual Information (NMI) and Adjusted Rank Index (ARI) will be calculated and compare with existing work.

Accuracy: It is sued to check the system efficiency according to the ratio of accurate matching feature during clustering by considering the True Positive (TP) that is the feature between two documents in the same cluster. Dissimilar features allocates two dissimilar documents to different clusters is known as a True Negative (TN) and False Positive (FP) represents the assignment of two dissimilar documents to the same cluster. Whereas, False Negative (FN) represents the assignment of assigns two similar documents to different clusters and based on these factors, accuracy of system is calculated using the equation 2.

$$Accuracy = Sum\ (TP, TN) / Sum\ (TP, TN, FP, FN) \quad \ldots (2)$$

NMI: It is a parameters used as an external measure to check the quality of clustering algorithm. Suppose, C is a set of data clusters with L labeled and the entropy is represented by H. so, to find out the Mutual Information (MI) between L and C the given equation 3 & 4 is used.

$$MI\ (L: C) = H\ (L) - H\ ((L \mid C)) \qquad \ldots (3)$$

$$NMI\ (L: C) = 2 \times MI / \{H\ (L) + H\ (C)\} \qquad \ldots (4)$$

ARI: It is sued to represents the adjusted rand index during the text documents clustering and the flow of proposed model is shown in the Fig. 2. The below flowchart represents the procedure of proposed work which is used to validate the efficiency of proposed web documents clustering using K-means with concept of ABC along with SVM as a classifier. Here, the combination of ABC and SVM is used to find out optimal clustering of text document based on their fitness evaluation. Algorithm here will readjust the elements of clusters to minimize the intra cluster and maximize the inter cluster distances and to validate the system, we will perform several experiments with this procedure on several query data. In the proposed system, there are several steps used for clustering of query text data. The methodological flowchart of proposed work is given below:

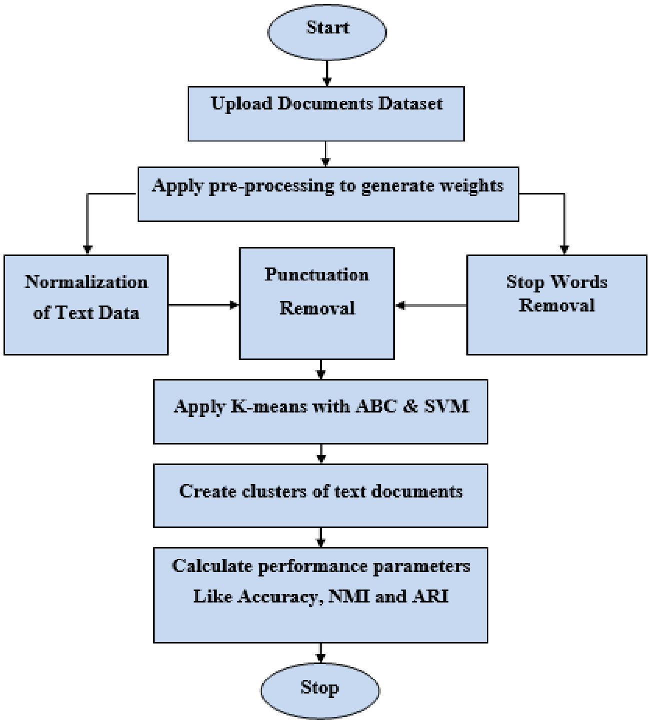

Start
Upload Documents Dataset
Apply pre-processing to generate weights
Normalization of Text Data
Punctuation Removal
Stop Words Removal
Apply K-means with ABC & SVM
Create clusters of text documents
Calculate performance parameters Like Accuracy, NMI and ARI
Stop

Fig. 2: Flow of Text Document Clustering

Based on the above mentioned methodology and algorithm, we simulate the proposed model for several time using the combination of three different datasets name as:

Reuters Dataset: It is the most useful dataset for the various text document clustering purpose and it contains to 21,578 text files. In 9187, the Reuters dataset was collected from the Reuters Newswire for the research purpose.

20 Newsgroup Dataset: It is also a text dataset and it was collected from twenty different news groups. It has also total 20 categories of data and total size is near to 20K numbers of text documents.

Topic Detection and Topic Tracking (TDT2) Dataset: It is a text document data repository and used for various document related research like topic detection and tracking, news classification, clustering etc. here, total 64,527 number of documents with 96 different categories are stored from different news agencies like America World News (AWN), American Broadcasting Company (ABC) World News, Cable News Network (CNN) News and many more news channels.

IV. SIMULATION & RESULTS

In this section we describe the simulation of proposed work with obtained results with several data. In this research, we have used a total number of three different text document datasets with two semantic groups or categories in a combined form. For all the datasets, we applied same process and steps mentioned in the third section of article like a data uploading, pre-processing, feature extraction and optimization using ABC and SVM as a classifier at the last. The obtained simulation results for proposed model in terms of Accuracy, NMI and ARI is shown in the Table 1:

Table 1: Performance of Proposed Text Document Clustering Model

No. of Documents	***Accuracy of Model***	***NMI of Model***	***ARI of Model***
10	94.832	0.95967	0.93525
20	91.688	0.98915	0.90935
30	98.179	0.99262	0.93638
45	90.283	0.94626	0.92301
50	89.383	0.93245	0.92605
60	97.539	0.98662	0.89586
70	98.835	0.99317	0.88037
80	94.576	0.96485	0.90546
90	94.894	0.98515	0.93525
100	96.933	0.98661	0.91912
Average	***97.714***	***0.97365***	***0.91606***

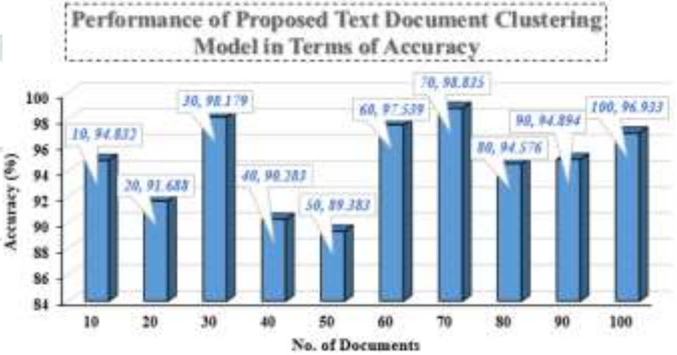


Fig. 3: Accuracy of Proposed Model

Above Fig.3 represents the achieved accuracy of proposed improved K-means based text document clustering model using the concept of a metaheuristic ABC as on optimization technique with SVM as a unsupervised classifier and we observed that the achieved accuracy is near to 97.97%. Above experimentation is done based on the average accuracy for group of ten sample at a time and simulation were done for total ten iterations (No. of documents).

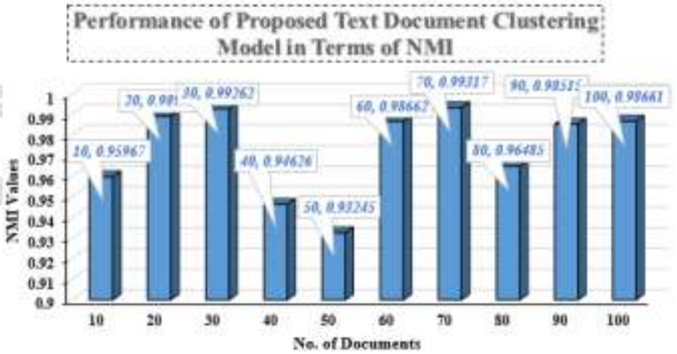


Fig. 4: NMI of Proposed Model

Above Fig.4 represents the obtained value of NMI that represents the mutual normalized information for proposed improved K-means based text document clustering model using the ABC with SVM and we observed that the achieved rate of NMI is more than 97% that indicates the utilization of improved K-means is doing better for ten number of group documents.

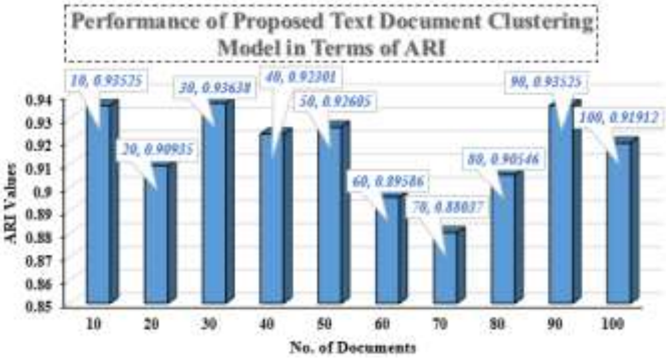


Fig. 5: ARI of Proposed Model

The rate rank index in terms of ARI is shown in the above Fig.6 with ten number of grouped document for clustering evaluation and the achieved ARI value indicates that proposed improved K-means based text document clustering model is far better than the existing traditional approaches. So, the comparison of proposed work with existing work by ***R. Janani & Dr. S. Vijayarani*** is given in Table 2 on the basis of different parameters like Accuracy, NMI and ARI for 100 text documents for clustering using the proposed an improved K-means based text document clustering using ABC with SVM as unsupervised classifier.

Table 2: Comparison of Proposed Work with Existing Work based on Accuracy, NMI and ARI

Parameters	Proposed Work	Existing Work
Accuracy	**0.97714**	**0.84130**
NMI	**0.97365**	**0.87970**
ARI	**0.91606**	**0.81230**

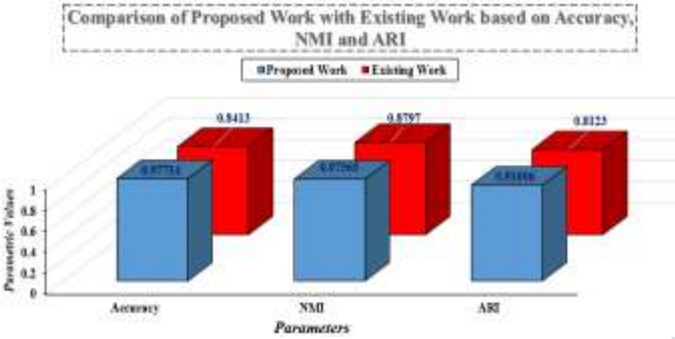


Fig. 6: Comparison of Proposed with Existing Work

Above Table 2 and Fig. 6 represents the comparison of proposed work with existing work by ***R. Janani & Dr. S. Vijayarani*** in 2019 and they worked with the concept of Spectral Clustering algorithm with PSO for text document clustering. From the figure, it is clearly seen that the, proposed an improved K-means based text document clustering model using the concept of a metaheuristic ABC as on optimization technique with SVM as a unsupervised classifier. The comparative results shows that the proposed work is far better than the existing work and total improvement is near to 13.9% by utilizing the concept of ABC with SVM for text documents clustering. So, we can say that the ABC as an optimization technique better than PSO for text document.

V. CONCLUSION

In this section, we presents the concluded points about the proposed an improved K-means based text document clustering using ABC with SVM as a classifier because the text document clustering is an open research area for researchers due several challenging factors. The problem of traditional K-means text document clustering has been considered when the documents data is huge for storing in the memory and should be implemented accordingly like from the disk, and from the point where the user should utilize small memory if feasible. The proposed algorithms are dependent on the new hypothetical outcome with major enhancement for developing it practical using ABC with SVM. To solve the above mentioned problem of document clustering in section2, we use the concept of ABC along with SVM based on the Euclidean distance to determine the similarity between the text documents and then optimize the clustering results is a better option. Based on the comparison, we founded that the proposed model is better in terms of accuracy, NMI and ARI whereas, the accuracy of system is improved by the 13.9%. Proposed model NMI and ARI rate is near to the 0.97 and 0.91 and it is a big achievement by utilizing the ABC as an optimization approach with fitness function along with SVM classifier. In the future, the proposed model with ABC and SVM will be extended by utilizing the concept of deep learning as a classifier to minimize its processing time for the semantic knowledge-based clustering with hybridization of similarity measure techniques.

REFERENCES

[1]. Janani, R., and S. Vijayarani. "Text document clustering using spectral clustering algorithm with particle swarm optimization." Expert Systems with Applications 134 (2019): 192-200.

[2]. Zhang, Qingchen, et al. "Secure weighted possibilistic c-means algorithm on cloud for clustering big data." Information Sciences 479 (2019): 515-525.

[3]. Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. Expert Systems, 36(1), e12340.

[4]. Lazhar, F. (2019). Fuzzy clustering-based semi-supervised approach for outlier detection in big text data. Progress in Artificial Intelligence, 8(1), 123-132.

[5]. Luna-Romera, José María, et al. "An approach to validity indices for clustering techniques in big data." Progress in Artificial Intelligence 7.2 (2018): 81-94.

[6]. Shrivastava, Puja, et al. "AKM—Augmentation of K-Means Clustering Algorithm for Big Data." Intelligent Engineering Informatics. Springer, Singapore, 2018. 103-109.

[7]. Zhao Jia et al. "A novel clustering-based sampling approach for minimum sample set in big data environment." International Journal of Pattern Recognition and Artificial Intelligence 32.02 (2018): 1850003.

[8]. Gu, Ziyuan, et al. "A big data approach for clustering and calibration of link fundamental diagrams for large-scale network simulation applications." Transportation Research Part C: Emerging Technologies 94 (2018): 151-171.

[9]. Al_Janabi, S., Salman, M. A., & Mohammad, M. (2018, November). Multi-level network construction based on intelligent big data analysis. In International Conference on Big Data and Smart Digital Environment (pp. 102-118). Springer, Cham.

[10]. Kanimozhi, K. V., & Venkatesan, M. (2018). A novel map-reduce based augmented clustering algorithm for

big text datasets. In Data Engineering and Intelligent Computing (pp. 427-436). Springer, Singapore.

[11]. Fan, T. (2018). Research and implementation of user clustering based on MapReduce in multimedia big data. Multimedia Tools and Applications, 77(8), 10017-10031.

[12]. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information systems, 47, 98-115.

[13]. Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, T. (2014, June). Big data clustering: a review. In International conference on computational science and its applications (pp. 707-720). Springer, Cham.

[14]. Sardar, T. H., Faizabadi, A. R., & Ansari, Z. (2017, July). An evaluation of MapReduce framework in cluster analysis. In 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) (pp. 110-114). IEEE.

[15]. Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).

[16]. Zhang, Qingchen, et al. "Secure weighted possibilistic c-means algorithm on cloud for clustering big data." Information Sciences 479 (2019): 515-525.

[17]. Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. Expert Systems, 36(1), e12340.