

RESEARCH ARTICLE

A Novel Hybrid Clustering Approach Based on Black Hole Algorithm for Document Clustering

FAZILA MALIK¹, SALABAT KHAN^{1,2}, ATIF RIZWAN²,
GHADA ATTEIA³, AND NAGWAN ABDEL SAMEE^{3,4}

¹Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan

²Department of Computer Engineering, Jeju National University, Jeju-si, Jeju Special Self-Governing Province 63243, Republic of Korea

³Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

⁴Computer Engineering Department, Misr University for Science and Technology, Giza 12511, Egypt


Corresponding authors: Salabat Khan (salabat.khan@cuatk.edu.pk) and Ghada Atteia (geatteiaallah@pnu.edu.sa)

ABSTRACT In information retrieval and text mining, document clustering is a big challenge because the amount of document collection has been increasing, day by day. The problem of clustering is NP-hard, use of meta-heuristic algorithms to solve these problems could be an effective method. When the solution space is large, traditional methods are unable to find a solution in a reasonable amount of time. K-means is a heuristic clustering algorithm, two main issues with heuristic algorithms are the early convergence and trapping in local optima. Moreover, finding the right number of clusters is one of the main drawbacks of the k-means algorithm. The correct value of k is always confusing, different researchers used different methods to solve this problem. To overcome these mentioned problems, this study presents a novel Hybrid approach for document clustering. One of the challenges in existing BH algorithm is the input data type. Recently, the algorithm was only accepting textual data. Another flaw in the existing model is that it doesn't choose how many clusters k to form automatically, and the centroids are chosen at random in it. In this paper, we have constructed a Hybrid cluster identification approach which consists of the Elbow method and Silhouette score for cluster k identification. This paper mainly offers three novel combination of model to represent text documents, namely i) K-mean++ - BH + TF-IDF with fix k ii) K-mean++ - BH + W2V with fix k iii) Hybrid Black Hole with automated k. The proposed improvements have validated on the document clustering problem. Cluster analysis based on two evaluation measures, external (Purity) and internal measures (Silhouette score) are used to report the findings. Experiments have been carried out on the four alphanumeric datasets (Doc50, Reuters, WebKB and News20) as well as on two numeric datasets (Iris and Wine) respectively. The complete result analysis is reported in detail with respect to each research contribution to compare the performance of the proposed algorithm with existing clustering methods. Result shows that the proposed Hybrid BH algorithm outperforms better than the existing clustering methods for all datasets. The clustering of data with and without stop words is examined; additionally, the two alternative word embedding used for data exploration in conjunction with proposed model are also evaluated. In the present study, proposed Hybrid BH algorithm handles the optimal value of k efficiently. This is one of the major contributions of the paper, concluded that Hybrid Black Hole is an effective algorithm for cluster analysis.

INDEX TERMS Document clustering, black hole algorithm, k-mean, data mining, comparative analysis.

I. INTRODUCTION

With rapid progress in technology, we can now collect large amounts of data of multiple types. These are unstructured

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh .

data, and cannot be analyzed quickly, as a result, we are unable to find a good solution to our query in search engines. Data mining is the process by which useful information is collected from large amounts of data. Data mining techniques have been used to solve a variety of real-life problems like clustering [1]. In clustering categorizing a population N data

points into K subgroups so that data points in one group are more similar to data points in other groups. Grouping data into groups of related data objects can provide meaningful structure to the data [2]. The higher the resemblance inside a group and the larger the variance between groups, the better or more definite the clustering. It is easy for the data analyst to process the data and discover new information from it. With reduced data dimensions, we efficiently minimize the amount of time, a computer takes to collect the requested information [3].

Document clustering is an application of data mining that is widely used in search engines. There has been a continuous increase in the number of documents. With the increase in the number of electronic documents, it is hard to organize, analyze and present these documents efficiently by putting a manual effort. Although humans can determine clusters in two and three dimensions but, when the data is in large amount, algorithms are required for high-dimensional data. There must be some way to organize data in such a way that the desired documents can be easily retrieved and located. So there is a need for an effective and efficient grouping of text documents automatically [4]. Given a group of N documents, the task is to divide this set N into a fixed number of K sub-groups g_1, g_2, \dots, g_k , so the documents that belong to the same sub-groups of documents have a high degree of similarity than those of other sub-groups. Grouping of documents is decided by the occurrence of words in each document set [5]. Document clustering can be used in document sorting, document retrieval, data visualization, document analysis, and document tag clustering, etc. [6].

Clustering algorithms are used for finding groups having a high degree of similarity based on maximum similar words among the documents [7]. Using cluster analysis, a user can get a good insight of a data (major properties) without any previous knowledge. However, cluster analysis is usually challenging due to the large number of input parameters required by most clustering algorithms. For K-means [8], [9], input parameters include the initialization of centroids and the number of clusters. Finding a suitable input configuration of an algorithm is often difficult without prior knowledge of the data. Parameters more often are adjusted using a time consuming trial-and-error method. It cannot be guaranteed that this will lead to the detection of useful parameter estimation. The performance of various standard clustering algorithms, such as K-means, is also influenced by user-defined parameters such as how initial points are chosen or which distance measure is used to compute data point similarity [10].

A. CONTRIBUTION

Several factors make document clustering a difficult task which are listed below;

- 1) Firstly, text documents suffer from high dimensionality and feature sparsity in representation. The data contains much fewer informative features than the original feature space. Furthermore, the number of words in

different document sets can vary significantly. Therefore, before using the clustering method, it is required to perform a proper text pre-processing step.

- 2) Secondly, the choice of initialization technique for centroid selection in k-means is important.
- 3) Thirdly, the correct identification of k-value is still a challenging task while performing document clustering.

The focus of this research is on an optimization based approach to clustering problems. We will use qualitative research to find the value of clusters, how many clusters are formed from the collected data. To the best of our knowledge, the hybridization of black hole algorithm [11] with heuristic algorithm (k-mean++) [12] has not been used to cluster documents. Their stochastic nature improves clustering by recovering from poor solution initialization and avoiding local optima.

In this paper, we propose a novel hybrid clustering approach based on black hole algorithm for document clustering. The complete paper is presented in the following order; Section 2 covers the literature review on existing approaches of document clustering. The methodology of proposed work and a detailed description of each module is explained in Section 3. Section 4 focuses on the result and also covers the answers to the research questions of this study. Section 5 concludes the whole research with conclusion, enhancements, and possible future work.

II. RELATED WORKS

An analysis of several pieces of literature on document clustering not only gives good knowledge but also helps to identify emerging challenges in the area of clustering [13]. There are numerous methods to solve the document clustering problem.

Lakshmi and Baskar [14] offered a novel DIC-DOC k-means algorithm (dissimilarity-based Initial Centroid selection for document clustering using k-means). Using this suggested method, the document with the lowest standard deviation of term frequency is selected as an initial centroid. The remaining initial centroids are picked based on how dissimilar they are to the centroids that have already been chosen. In this study, WebKB and Reuters 8 are the two data sets used to validate the value of clusters. Two documents are compared by using a cosine similarity measure. Using three external measures: entropy, purity, and F-measure, efficiency of proposed algorithm is compared to different clustering algorithms over a range of k values. The identification of k-values is not addressed in this work. Abdolreza [11] offered a novel algorithm based on black hole phenomenon which is used to solve the clustering problem. This research is conducted out on six numeric datasets: Iris, Vowel, Wine, Glass, Cancer, and CM, using error rate and intra-cluster distance as evaluation measures. The findings of the experiment, uses six benchmark datasets, indicates that the proposed black hole (BH) algorithm surpasses existing test algorithms (PSO,

K-means, and GSA). The presented mathematical idea of BH algorithm can be used in combination with other algorithms, which is much more successful than using it individually.

K-means is one of the most useful heuristics-based partition clustering algorithm. It has a good convergence speed however it often gets stuck into the local optima. This is because the performance of K-means is dependent on initial centroids chosen. The quality of clusters formed is highly influenced by initial centroid values [15]. In the last few years, many attempts [16] have been made by researchers to overcome this drawback. Among them, one of the successful attempts is to integrate heuristic (K-means) clustering algorithm with meta-heuristic (nature-inspired) algorithms [17]. Nature-inspired algorithms are non-deterministic optimization techniques. Their exploration and exploitation ability provide a near-optimal solution to non-linear, high dimension, and complicated problems within acceptable time limits [18]. Muhammad *et al.* [19] provided a soft computing-based method for document clustering. The implementation of Black hole algorithm is performed in this work. The random heuristic algorithm is embedded in black hole algorithm to produce the best results. For parameter variations, local and global search optimization are used. Experiments are performed on text mining datasets named Reuters, WebKB, Doc50, and News20, and results are calculated based on silhouette and purity index. The proposed method outperforms the simple k-mean method and produces a near-optimal solution. The pre-determination of a number of clusters k is not handled automatically and centroid initialization in k-means is random.

Chouhan and Purohit [20] proposed a method for document clustering that combined K-means and PSO (Particle Swarm Optimization). To determine initial cluster centroids for the K-means method, PSO is used before K-means. The results of clustering methods are examined on four datasets (BBCSports, FOX, BBC, and CNN). To validate the performance of the suggested algorithm, three evaluation measures (cohesion, entropy, and separation) are used.

Lakshmi *et al.* [21] used the Crow Search Algorithm with K-means (CSAK) to discover the optimum global solution. Six benchmark datasets (Breast Cancer, Contraceptive Method Choice (CMC), Iris, Glass, Haberman's Survival, and Wine) are used to determine the performance of the proposed CSAK-means algorithm. These data are obtained from the UCI machine repository [22]. The validity of CSA-KM is estimated with internal (Silhouette Score) and external (Purity, Rand Index, Normalized Mutual Information, Precision, F-Measure, and Recall) measures. The result of suggested algorithm is compared to the results of the other algorithms (PSOK-means, K-means, Genetic k-means, K-means++).

The fitness function used to analyze the CSAK-means algorithm is the Mean Square Error. The CSAK method outperforms other algorithms in test experiments. There is a need to automatically decide the number of clusters.

Mohammad *et al.* [23] introduced a new hybrid-mean algorithm that combines the Black hole (BH) algorithm with bisecting k-means algorithm (BK). The presented hybrid algorithm (BH+BK-means) combines the global searching ability of BH algorithm with the quick convergence capability of K-means algorithm. Experiments on various real datasets (CMC, Glass, Iris, Vowel) have shown that using a composite solution with bisect k-mean and black hole algorithms to find cluster centers is better than using single k-mean and black hole algorithms. Maintaining the sequence of the hybrid algorithm (BH-BK) is highly useful. The overall search performance and efficiency of BH algorithm are reduced when BK-means clustering is performed before BH clustering module. The average intra-cluster distance and error rate are used to determine and compare the performance of the provided algorithm. Experiments on real datasets indicate that the novel hybrid BH+BK-means method exceeds individual algorithms for finding cluster centers. The pre-determination of the value of k is not handled.

Yogesh and Ashish [24] utilized the particle swarm optimization (PSO) approach with K-harmonic means (KHM) for clustering. To overcome KHM's limitations, such as the local optimum problem, PSO is made adaptive with the use of fuzzy logic. Comparison of suggested method named Enhanced fuzzy PSO-based clustering method with K-harmonic means (EFPSOKHM) shows that the proposed algorithm produces better clusters than existing algorithms. Five numeric benchmark datasets (Cancer, Iris, Wine, CMC, and Glass) are used to validate the effectiveness of proposed approach. The pre-determination of the value of k is not handled. For addressing the exploration issue in the original black hole, Haneen *et al.* [25] suggested a new clustering algorithm named levy flight black hole. In this algorithm, the movement of all stars generally depends on step size, produced via Levy distribution. This novel clustering approach was tested on six datasets namely Iris, CMC, Glass, Cancer, Wine, and Vowel collected from the UCI machine learning laboratory [22]. The algorithm performance is tested via two evaluation measures sum of intra-cluster distance and error rate. Experiment results demonstrated LBH approach escape easily from the local optima and clustered data objects efficiently. The number of clusters k is not handled.

Literature illustrates that several algorithms have been developed to deal with document clustering (NP-hard) problems but optimal solutions are not guaranteed. There exists no algorithm that finds the optimal solution to NP-hard problems. Many problems are solved by hit or trial method but it does not work for all types of problems. For example, K-means clustering algorithm is treated as an optimization algorithm but it could not find optimal clusters as it depends on initial centroids chosen. These centroids are selected randomly through hit and trail method. To cope up with the NP-hardness of clustering problem, researchers have drawn their inspiration from nature [26], [27], [28], [29], [30], [31]. Since decades, nature has been a rich

source of inspiration for developing new algorithms termed nature-inspired optimization algorithms. The various studies discussed above precisely indicate that various meta-heuristic algorithms i.e. nature-inspired algorithms have been integrated with K-means algorithm to improve clustering efficacy. Nevertheless attaining global optimal solution and pre k-value identification remains a challenge. Clustering algorithms developed in the literature provide a near-optimal solution. Hence, there is scope for improvement.

To sum up, existing literature provides the following research gaps respectively Thus, it can be concluded from the above research observations that none of the nature-inspired algorithms is applied to cluster data for all types of inputs. Many of the studies just perform experiments on classification datasets [11], [21], [23], [24], [25], which shows limited suitability for analyzing clustering algorithm performance. Many methods have been proposed and used to improve the effectiveness of text document clustering, but still, now there are many challenges in text document clustering, such as Documents Representation, High Dimensionality, Efficient Initial Seed Selection, and Semantic Relationship between words, k-value identification, and effective clustering algorithm. Many document representation models [32] have been used the bag-of-words and term frequency. They do not consider the semantic relationship between words and also face the high dimensional problem. Need an efficient approach to calculate the semantic relationship between terms of a document and grouping similar terms based on the semantic relationship. To overcome this problem, we use the word2vec [33] model in our method. Lack of finding quality clusters, due to random initial seed selection techniques. we have use k-mean++ in our approach and for k-value identification, we make a hybrid approach with the help of two well-known cluster identification measures elbow and silhouette score and for performance validation of our proposed work, we will use evaluation measures purity and silhouette. To achieve good performance of clustering method by improving feature extraction and feature representation, optimizing solution for all types of inputs data, and finding the best K value for clustering is a challenging problem. This forms the motivation of our problem statement i.e. “A novel hybrid approach based on Black Hole algorithm for document clustering” In this research work, the document clustering problem has been formulated as an optimization task and is solved using a hybrid approach based on meta-heuristic algorithm (Black Hole) and heuristic algorithm (k-mean++).

III. MATERIALS AND METHODS

A. DATASET COLLECTION

We presented results on four standard alpha-numeric text datasets [19]: Doc50, News20, WebKB, and Reuters, and two numeric datasets: Iris [34] and Wine [34] respectively, collected from UCI Machine Learning Repository.¹

1) ALPHA NUMERIC DATASETS

a: Doc50

Doc50 is a subset of news20 dataset which contains 50 documents. It is the most basic dataset having a minimum possible unique tokens. Some of them are lengthy emails, while others are simply e-mail chunks. They have a lot of stop words and special characters in them.

b: NEWS20

News20 dataset contains documents from the newsgroups dataset. The data is organized into 20 newsgroups, each with its own topic. Some newsgroups are closely associated, while others are completely unrelated. It is the famous dataset for machine learning research in text applications.

c: WebKB

WebKB dataset is a collection of web pages of four universities. It collects details from four different universities of computer science departments, including students, faculty, projects, staff, courses, and additional details of the department. We have tested the proposed algorithm in the courses section. In this category, there are 930 documents and 5 classes.

d: REUTERS

Reuters is a subset of the original Reuters21578 dataset which consists of 12 classes. Each class includes documents on a particular topic. In each class, the total number of documents ranges between 50 to 100. It is a set of documents containing news articles.

2) NUMERIC DATASETS

a: IRIS

The Iris dataset is divided into three classes, every 50 instances related to a different species of iris plant. There are 50 samples in the Iris dataset, each one with four different characteristics (sepal and petal length and width). Iris dataset is commonly used in data mining, classification, and clustering purposes and also for algorithm testing.

b: WINE

The wine dataset includes the findings of a chemical examination of wines manufactured in a single Italian region. The 178 samples represent three types of wine, with the results of 13 chemical tests performed on each sample. There are no missing values in the data, it is entirely numerical and classified using a three class target variable.

3) DATASET COMPLETE STATISTICAL INFORMATION

Complete statistical information, as well as the difference in dimensionality between these datasets, is given in Table 1 and Table 2.

Table 1 and Table 2 show the number of documents in each dataset, the number of terms in each document, and the ground truth values that are the number of clusters it has.

¹<http://archive.ics.uci.edu/ml>

TABLE 1. Dataset Information (Alpha-Numeric).

Datasets	No.of Documents	No. of Terms	No. of clusters
Doc50	50	3462	5
News20	813	52228	20
WebKB	930	11286	5
Reuters	786	6609	12

TABLE 2. Dataset Information (Numeric).

Datasets	No. of Documents	No. of Terms	No. of clusters
Iris	4	150	3
Wine	13	178	3

B. FLOWCHART OF PROPOSED APPROACH

Proposed Approach has three main module which are explained in Figure 1.

As presented in Figure 1, we have divided our methodology into three main modules which are listed below:

Module 1: Data Pre-Processing

Module 2: Cluster Identification

Module 3: Hybrid Black Hole Algorithm

Each module is further subdivided into phases, which will be discussed in detail as follows.

1) MODULE 1: DATA PRE-PROCESSING

This module is sub-divided into various phases: optimizing for all inputs, feature extraction, collection, and document representation.

a: PHASE 1: OPTIMIZING FOR ALL INPUTS

In this phase, we have pre-process the data in any format (alphanumeric or numeric) and convert it to a numeric format that can be used as input to the algorithm. This phase includes several sub-processes, such as deciding whether or not to normalize the data and whether or not standardization is necessary. In our case, we have made two separate functions to read data. One is for the alpha-numeric dataset and the other is for the numeric dataset. Using the Python's Pandas library, drop the last column in the dataset as it contains ground truth values of the clusters that are to be formed. The next step for numeric datasets is to standardize the data, but for alpha-numeric datasets, we will use word embedding (word2vec and TF-IDF) to create features. To standardize numeric data, we may use min-max scaling rather than word2vec or TF-IDF embedding.

b: PHASE 2: FEATURE EXTRACTION

During this phase, we performed some initial cleaning steps on our dataset. Cleaning the data is a very important step in any sort of analysis. By converting all characters to lower-case, eliminating punctuation marks, and removing stop words and typos, it is possible to remove unhelpful sections of the data, or noise. In the feature extraction process, we parse each document to generate a collection of features while excluding a list of pre-defined stop terms that

are meaningless. We have performed the selected cleaning on our used dataset. The cleaning processes will be as follows which are listed below:

- 1) Remove all stop words
- 2) Remove all punctuation
- 3) Remove all lower case and blank spaces
- 4) Lemmatization the words

By applying the word embedding technique, we started encoding collected datasets using python. We convert Alpha-numeric data into readable form, remove unnecessary information by using the word embedding techniques (Word2vec and TF-IDF). Correct feature selection decreases the high dimensionality of the feature space and improves data comprehension, resulting in improved cluster creation. We compare our findings with and without stop words in our work, so we have set stop word removal as an optional parameter to check its impact on the results. In our case, we have performed results analysis with and without using stop words in data

c: PHASE 3: DOCUMENT REPRESENTATION

The first exploratory step in the clustering process is to represent the text documents uniformly. The goal is to organize the documents coherently. The machine learning algorithms are not capable of working directly with raw text, therefore the unstructured form of documents must be transformed into a vector of numbers. In the document representation phase, each document is represented by k features with the highest selection metric score. The word embedding technique word2vec stores the relationship between words, every word is represented in a 32-bit vector. Word2vec consists of two models CBOW (continuous bag of words) and Skip-gram. In this work, the CBOW model is used. For document representation in W2V model, we have used CBOW (continuous bag of words) because it is much quicker to learn a model than skip-gram and also has better accuracy for common words. For numeric data, we will not be using word2vec, for that we will only use min-max scaling to normalize all the inputs.

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Eq.(1) shows the equation of the min-max scalar for normalizing all the inputs (x_{norm}), where x is the value of each instance in the specific column. Here $\min(x)$ and $\max(x)$ are the minimum and maximum values of the feature respectively.

d: PHASE 4: STANDARDIZED DATA

After the successful vector formation for all types of data inputs, the pre-processing module (module 1) is now complete. Now, data is in a standardized format. Next, we will discuss the cluster identification phase, in which we determine how many clusters are to be chosen.

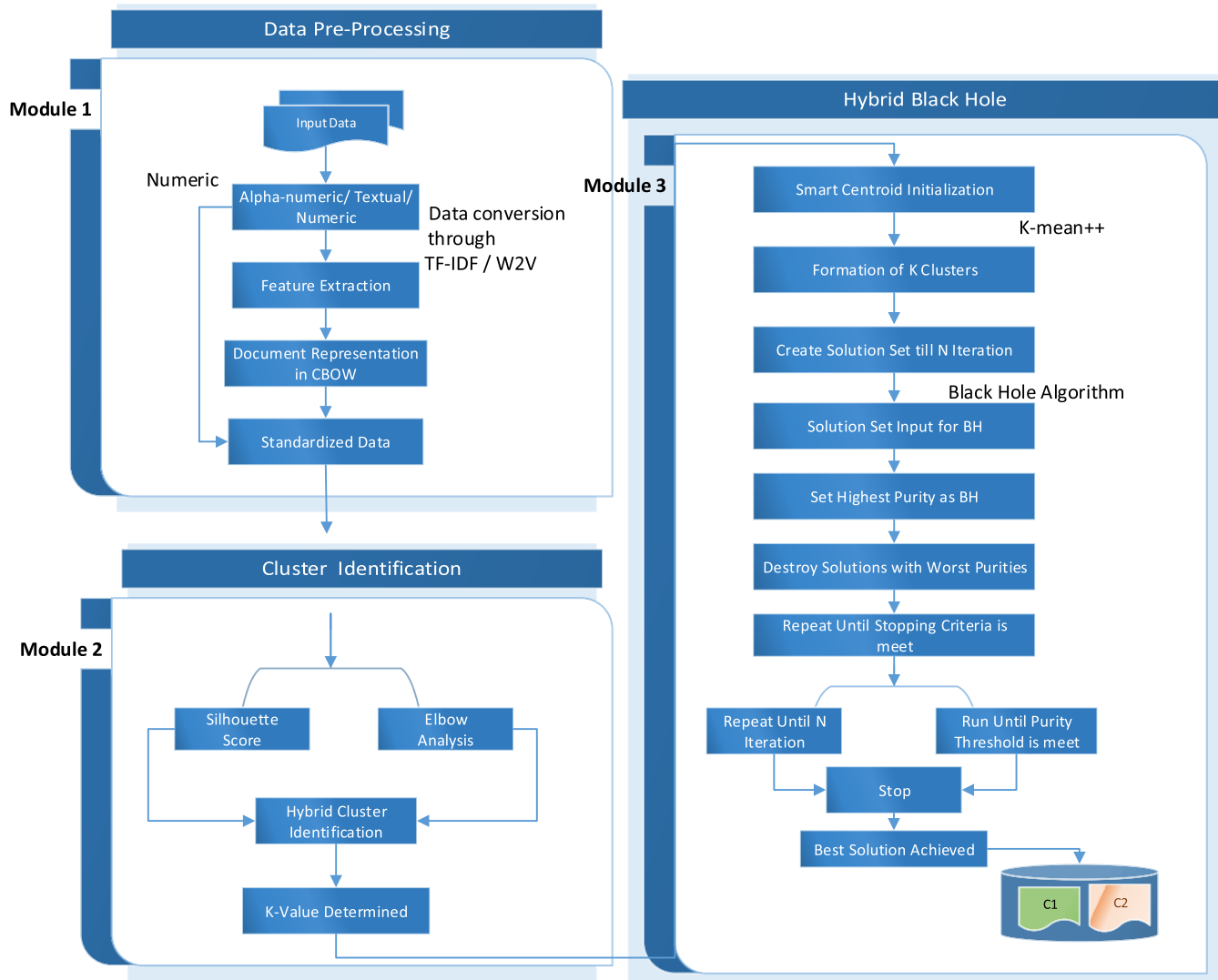


FIGURE 1. Flowchart of Proposed Approach.

2) MODULE 2: CLUSTER IDENTIFICATION

In this module, the number of clusters will be determined. One of the method for determining the correct number of k for the k-mean algorithm is to use the Elbow method. Although, it produces generally good results and is easy to understand and implement, but it involves biased judgment to decide where the actual elbow is found. Silhouette score is another method used to check the validity of clustering. The base K-mean and any other K-mean variant does not select the number of clusters automatically. We have devised a hybrid technique that averages out the result of silhouette score and elbow analysis to determine the most optimal number of clusters for a specific dataset.

We have used two methods in our hybrid approach. The first one is the commonly used Elbow analysis method and the second one is the Silhouette score. The Silhouette score computes the dissimilarity between clusters. In the present study, proposed Hybrid BH algorithm handles the optimal

values of k efficiently. This is one of the major contributions of the paper. In proposed hybrid approach, we take the average of both method findings (Elbow analysis method and Silhouette score) and then proceed. There are two possibilities in this module. First, if the user knows how many clusters are required, they can manually enter the number of clusters they needed. In the second situation, when the number of clusters to be selected is unknown, then proposed hybrid approach is used to automatically determine the number of clusters.

a: PHASE 5: SILHOUETTE SCORE METHOD

The Silhouette score method is used to select the optimal number of clusters present in the data. The cohesion is measured based on the distance between all the points in the same cluster and the separation is based on the nearest neighbor distance. It is recommended that user provide this number if he/she already knows the number of clusters. If not, the best number of clusters is selected by proposed hybrid function.

We select the cluster on which we have the best silhouette score.

b: PHASE 6: ELBOW ANALYSIS METHOD

The Elbow analysis method is a very common method for determining the optimum k value of a cluster. This method is used to calculate the distance via cosine similarity. The equation of the cosine similarity is shown below

$$Cosine_Similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \times \|B\|} \quad (2)$$

where in Eq.(2) Ai and Bi are vector elements. Cosine similarity measures similarity between vectors of two documents. The distances are first computed between vectors, and the cosine function calculates the similarity.

Silhouette method is also used for K value identification. The silhouette method is considered better for K value identification as compared to elbow because it is used to study the separation distance between the resulting clusters. The elbow method only calculates the distance, while on the other hand silhouette consider other variables such as high-level difference, variance, and skewness.

c: PHASE 7: HYBRID CLUSTER IDENTIFICATION

After performing the hybrid cluster identification approach for k-cluster identification, an optimal value of k is determined. After running both the methods (Silhouette score method and Elbow analysis method), we have compared the results and take the average of both the methods.

For example, if elbow analysis suggests that the best elbow is on cluster 3 and silhouette co-efficient suggests that the best score is on cluster 5, then we simply take the average of the resultant of two methods which will result in k=4 clusters. After the optimal k value determination, the cluster identification module (module 2) is now complete. Next, we will discuss the hybrid black hole phase, in which we achieve the best solutions for cluster formation.

3) MODULE 3: HYBRID BLACK HOLE ALGORITHM

In this module, we have an input of k locally optimized k-mean++ solutions into the black hole algorithm. The first step is to identify the black hole (cluster with the best local optima). Now, this black hole will attract the nearest star (clusters) and destroy it. When this star is destroyed, a new star will be created. The k-mean++ algorithm is responsible for the formation of new stars (clusters). Whenever a star is destroyed, k-mean++ is invoked to generate a new star. We will repeat this process until the mentioned stopping criteria will meet that is, run until N iterations and run until purity threshold will meet. If any of the mentioned stopping conditions will meet, the algorithm stops and we report the best solutions of k cluster formation.

We proposed a global optimal solution by embedding a k-mean++ solution to Black Hole Algorithm. Proposed approach uses the global optimal property of the Black Hole algorithm. We simulated the idea of event horizon in our

algorithm using inspiration from real-world black hole phenomena. We have used purity and silhouette score as the event horizon in our algorithm.

Mathematically, a multi-objective clustering problem can be written as shown in Eq.(5). If $D = d1, d2, \dots, dN$ are the 'N' documents, the problem is to find the k document-clusters $c1, c2, \dots, cK$ then $ci = di1, di2, \dots, di n$, and $T = t1, t2, \dots, tm$ is set of 'm' different terms which occur in D. Here n shows number of documents in cluster 'i', ' $d^{i n}$ ' represents 'nth' document of cluster 'i'. Here $ci \cap ck = \phi$ for all $i \neq n$.

Vector representation of documents $di n$ is as follows in two ways:

$$d^{i n} = tfidf(dn, t1), tfidf(dn, t2), \dots tfidf(dn, tm) \quad (3)$$

$$d^{i n} = wv(dn, v1), wv(dn, v2), \dots wv(dn, vm) \quad (4)$$

Here 'tfidf' represents term frequency and inverse document frequency of each documents and 'wv' shows word vector of word2vec embedding.

Where objective functions is

$$Maximizef(K) = (f1(k1), f2(k2), \dots, fm(kn)) \quad (5)$$

$ki = (k1, k2, \dots, kn) i \in k$.

Here K is the set of possible solutions (in terms of purities/ silhouette score) generated by defined k-means++. It is nearly impossible to maximize all of the objective functions at the same time with a single solution k in K because the objective functions usually vary. The set B is the set of potential solutions k in K for which no other survivable solution is as good as k in all objective functions and completely better than k in at least one objective function.

Updation of next star position is based on,

$$ki(t + 1) = ki(t) + rand * (kBH - ki(t)) \forall i = 1, 2, \dots N \quad (6)$$

where 'ki(t)' is the current position of a star at iteration 't', ' $ki(t + 1)$ ' is the next position of stars at iteration '(t + 1)' and 'kBH' is the best solution among all at each iteration. For $j = 1, \dots, m$, the set B is explicitly defined as in equation 7. Which satisfies the following mentioned criteria:

$$B = \left\{ \begin{array}{l} (k \in K : fi(k) \geq fi(b) \forall b \in K, i \in j \text{ and}) \\ fi(k) < fi(b) \text{ for some } i \in j \end{array} \right\} \quad (7)$$

$fi(b)$ = best global solutions achieved.

After the successful completion of module 3, the best solutions of k cluster formation is reported.

A general pseudo code of proposed Algorithm is mentioned in Algorithm.

C. EVALUATION MATRIX

The cluster evaluation measures are used to ensure the quality of the results produced by our proposed algorithm. They enable us in identifying correct cluster findings [35]. The cluster validity indices can be divided into two major categories:

- i) Internal
- ii) External

Internal indices [36] are used to measure the goodness of the cluster structure by depending on the implicit knowledge of the data. Besides that external indexes [37] are used to analyze clustering results by comparing cluster memberships given to a clustering algorithm with existing information, such as an externally provided class name. We have used two evaluation measures for the validation of our approach. One is the Purity index [38] and the other is the Silhouette score [35]. Each measure has a range of values associated with it.

1) EXTERNAL MEASURE (PURITY INDEX)

Purity is a measure that determines how closely a group of documents belongs to the same class. It is also known as homogeneity or purity of class. The purity of a cluster is used to determine its homogeneity. It has a value between [0, 1], where 0 is the worst and 1 is the best clustering solution. When we have a class label available, we evaluate the clustering results using that class label, then purity is the best measure [39]. In this measure, each cluster is given a label based on the most widespread class within it. The purity measure of a cluster indicates how much data from a single class it contains.

Algorithm 1 Data Clustering

Data: Dataset X,Y,Population size(i.e number of stars),maxiter,ninit
Result: K Partition of data
 initialization;
 data \leftarrow (Request(ReadCSV));
if Text document **then**
 TFIDF and Word2vec to extract the numeric features from text data;
if Nemerical data (Wine and IRIS) **then**
 standardize the data by Eq.(1),(3) and (4) ;
for each $d=1$ to dn **do**
 Compute Hybrid Cluster
 $k \leftarrow \frac{\text{ElbowCurveK} + \text{silhouetteK}}{2}$
 return K
for each $d=1$ to dn **do**
 input K Number of clusters to Hybrid BH
 Initialize position of stars as encoded by local search (K mean++) algorithm
 The position of stars is updated using Eq.(6)
 The objective function of new position of star is updated by Eq. (7)
 if stopping criteria in Eq.(5) and maxiter is not met **then**
 Repeat the above steps ;
 Best global solutions achieved

The purity is then calculated by dividing the number of correctly matched class and cluster labels by the total number

of data points. It represents the degree of homogeneity among clusters.

External Measure (Purity) is defined as

$$\text{Purity Index} = \sum_j \frac{n_j}{n} \times p(j) \quad (8)$$

Here in Eq.(8), ' n_j ' is the number of documents in cluster ' j ', ' n ' is corpus size and ' P_i ' is the ratio of the majority class in that cluster $p(j) = 1/n_j \max(n_{ij})$ where ' n_{ij} ' is the number of documents of class i in cluster j .

2) INTERNAL MEASURE (SILHOUETTE SCORE)

Peter J. was the first to propose the Silhouette method [40]. The silhouette score measures how close an object belongs to its own cluster (cohesion) as compared to other clusters (separation). Silhouette score is a method used to check the validity of clustering. It combines two factors cohesion and separation. The similarity between the object and the cluster is referred to as cohesion. It's referred to as separation when compared to other clusters. The Silhouette Score is used to determine how good a clustering technique is. Its value ranges from -1 to 1 . The silhouette plot shows how close each cluster's point is to its neighboring cluster points. When a class label is unknown, the silhouette coefficient is a more relevant estimator. The Silhouette value is close to 1 , indicating that the object and the cluster have a close relationship. A value of 0 specifies that the object is on or near the decision boundary between two neighboring clusters, while negative values indicate that the objects may have been assigned to the incorrect cluster [41].

1 : Indicates that clusters are well separated and distinct from one another.

0 : Indicates that clusters are unrelated, or that the distance between clusters is not significant.

-1 : Clusters have been assigned incorrectly.

It is defined as:

$$\text{Silhouette Score} = \frac{1}{k} \sum_{j=1}^k S_j \quad (9)$$

Here in Eq.(9), the Silhouette value of the i th vector in the cluster S_j is given by

$$S_j = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (10)$$

Here in Eq. (10), ' $a(i)$ ' is the mean distance between ' i ' and all other data points in its own cluster and ' $b(i)$ ' is the mean distance between ' i ' to all data points in other cluster centroids [42].

IV. RESULTS AND DISCUSSION

We explain our experimental results of the conducted research in this section.

A. PARAMETER SETTINGS

In this experiment, selected hyper-parameters of local and global search algorithm are used by first investigating data

TABLE 3. Hyper-parameters sett.

Parameter Name	Values
Number of Stars (Population Size)	150(BH*K)
Global Search	5-15
Local Search	300
Rand	1
Number of Features	Vector length

TABLE 4. Comparison of purity results (without stop words).

Datasets	k-mean	k-mean++
Doc50	0.82	0.86
Reuters	0.87	0.88
WebKB	0.76	0.77
News20	0.6	0.64

and search space for the best hyper-parameters. Table 3 showing the parameter setting of values for proposed model.

Table 3 presents the parameter setting; algorithm is run for each iteration and for each dataset’s evaluation. Meta-heuristic optimization algorithm can quickly produce a global optimal solution using heuristic algorithm. The phenomena of global search and local search optimization are used as parameters adjustments of proposed algorithm. Tested different parameters on these datasets. We have chosen those parameters in which proposed model performed best. To our knowledge, these selected parameters performed best and gave better results than others.

B. PERFORMANCE COMPARISON

This section presents the result of the conducted research. In this section, evaluate the performance efficiency of proposed algorithm in terms of two measures i) Purity and ii) Silhouette score for evaluating clustering quality. Results are listed in tabular form as well as graphically represented. Complete result analysis are performed in two context: one is without stop words and another is with stop words to answer our research question which is to find the impact of stop words on results

1) PERFORMANCE ANALYSIS BASED ON EXTERNAL MEASURE

α: PERFORMANCE ANALYSIS OF HEURISTIC METHODS ON ALPHA-NUMERIC DATASETS

Step by step presented discussion on results according to research questions which are listed below:

- 1) Which variant of k-mean (k-mean vanilla or k-mean++) performs the best inside the black hole algorithm?
- 2) What will be the effect on the findings before and after removing stop words from the dataset?

To answer these above mentioned research questions, firstly compared the results of both heuristic algorithms, k-mean and k-mean++ individually with or without using stop words in data.

From Table 4 it is observed that k-mean++ performs better than k-mean in all datasets. The baseline clustering algorithm

TABLE 5. Comparison of purity results (with stop words).

Datasets	k-mean	k-mean++
Doc50	0.8	0.82
Reuters	0.7	0.74
WebKB	0.75	0.76
News20	0.5	0.57

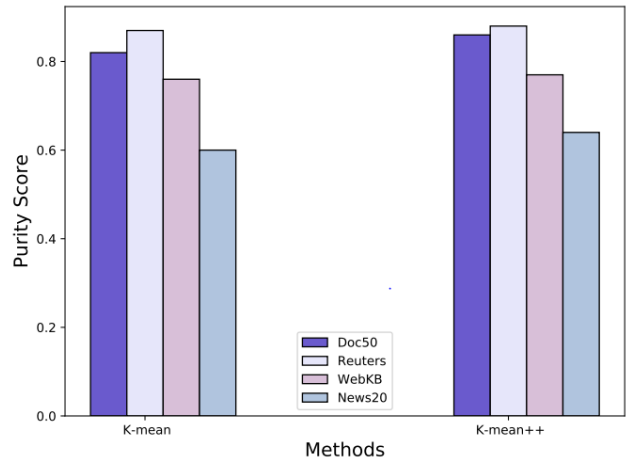


FIGURE 2. Purity result comparison of heuristic methods without stop words.

k-means select the initial centroids randomly. Due to this nature of initialization sensitivity in k-means, the clustering algorithm trend the following problems; (i) To affect the final formed clusters offers low quality clustering solutions and (ii) provide solutions with local optima because initial set of center are not distributed over the dataset.

To avoid this problem of initialization sensitivity in k-mean, k-mean++ algorithm is used and enhancement in results by considering the four datasets are shown in Table 4. K-mean++ is a smart centroid initialization technique based on probability distribution instead of randomly picking all the centroids. It yields a much better performance as compared to baseline algorithm.

Table 5 clearly depicts that data with stop words badly effects the algorithm results as compared to data without stop words. Remove stop words is determined by the nature of data. In proposed work, datasets (Doc50, Reuters, WebKB and News20) are used which are based on emails, webpages, university courses and news document respectively. Not eliminating stop words from data curse to degrade performance of clustering algorithm. In both context, results comparison of Table 4,5, k-mean++ performs more efficiently than base algorithm k-mean.

The pictorial representation of comparison based on purity result of heuristic algorithms (k-mean and k-mean++) with and without using stop-words on four datasets are shown in Figure 2 and Figure 3.

Figure 2 shows the results of K means and K means++ without stop words on all alpha numeric dataset. It shows that

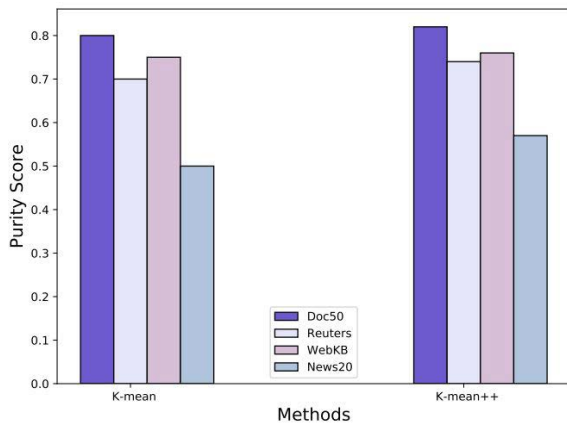


FIGURE 3. Purity result comparison of heuristic methods with stop words.

K mean++ perform better on all dataset set as compared to K mean clustering algorithm.

Figure 3 shows the result of K mean and K mean++ with stop words. It shows that K mean++ perform better as compare to K means with stop words. However on the other hand, K mean++ achieved better results without stop words as compared to with stop words.

The results presented in Figure 2 and 3 differentiate two concepts; one is that k-means++ outperforms the k-means and the reason for this is due to the fact that the k-means identifies the first initial centroid randomly, while the k-means++ algorithm [43] selects the second initial centroid through probability proportional to the square of distance over summation of square distance for the current point and second by removing stop words / low level information from data, enable us to focus more on the important information which helps to improve the performance of clustering algorithms. Removing stop words helps to enhance performance of clustering algorithms. The performance is not too encouraging but compared to the baseline method the used algorithm k-mean++ have been able to show improvement in purity with 4% in some datasets and 1% on other datasets which are considered an extent enhancement as compared to the baseline method results [19]. Based on these findings, it is recommended that it is better to choose heuristic algorithm (k-mean++) inside the BH algorithm to enhance results.

b: PERFORMANCE COMPARISON OF EXISTING BH AND PROPOSED (K-MEAN++ - BH) ALGORITHM

According to proposed conducted analysis in Table 4,5 it is clearly evident that heuristic algorithm k-mean++ performs better than k-mean. So due to this reason, embedded k-mean++ inside BH algorithm and compare its results with existing BH [19] model using same feature extraction technique i.e. TF-IDF.

The main reason for the selection of heuristic algorithm with a combination of meta-heuristic (BH) algorithm is, its powerful exploration ability i.e. local search. Black Hole algorithm [44] explores entire search space effectively to

TABLE 6. Comparison of purity results (without words).

Datasets	BH	k-mean++-BH
Doc50	0.94	0.96
Reuters	0.90	0.94
WebKB	0.80	0.82
News20	0.63	0.78

TABLE 7. Comparison of purity results (with stop words).

Datasets	BH	k-mean++-BH
Doc50	0.84	0.9
Reuters	0.86	0.88
WebKB	0.79	0.8
News20	0.61	0.75

determine optimum solution. Focuses on the shortcomings of heuristic algorithm and explores the search space effectively, proposed hybridization of algorithm is a best choice for cluster analysis.

Table 6 shows the results of existing black hole and Black hole with K mean++ for alpha numeric dataset without stop wrds. It shows that existing existing black hole improves the results as compared to the existing black hole on all dataset.

In Table 7 the results of existing black hole and black hole with k mean++ is presented with stop words for all text dataset. Black hole with K mean++ is also perform better on all dataset.

Table 6 and 7 illustrates that proposed hybridization of algorithms (k-mean++-BH) performs better than the existing BH algorithm in all datasets also by eliminating an unwanted information from data improves model performance. Due to the problem of random selection of initial centroids in existing BH algorithm [19], takes a large number of iterations for each datasets in comparison to the optimization based K-means++ - BH clustering algorithm. The improvement in the results show that it possess the capability of greater convergence in objective function values. There has been 2% improvement observed on Doc50 and WebKB datasets, 4% on Reuter’s dataset and News20 dataset 15% improvement has reported respectively. 6, 7 Tables expresses that that using this hybridization of methods (K-means++- BH) generate higher compact clustering than either using each algorithm individually.

c: RESULT ANALYSIS BASED ON DIFFERENT WORD EMBEDDING

This section is prepared to perform deep analysis of the impact of two different word em-embedding on results. The detailed result analysis is specified in terms of two different word embedding; one is TF-IDF and another is W2V which are mentioned in Tables 8,9.

Table 8 shows the results of Proposed method without stop words using TF-IDF and W2V embedding techniques. It shows that Proposed method achieved better result with TF-IDF as compared to W2V.

TABLE 8. Impact of word embedding (without stop words).

Datasets	TF-IDF	W2V
Doc50	0.96	0.9
Reuters	0.94	0.83
WebKB	0.82	0.75
News20	0.78	0.65

TABLE 9. Impact of word embedding (with stop words).

Datasets	TF-IDF	W2V
Doc50	0.9	0.84
Reuters	0.88	0.72
WebKB	0.8	0.74
News20	0.75	0.62

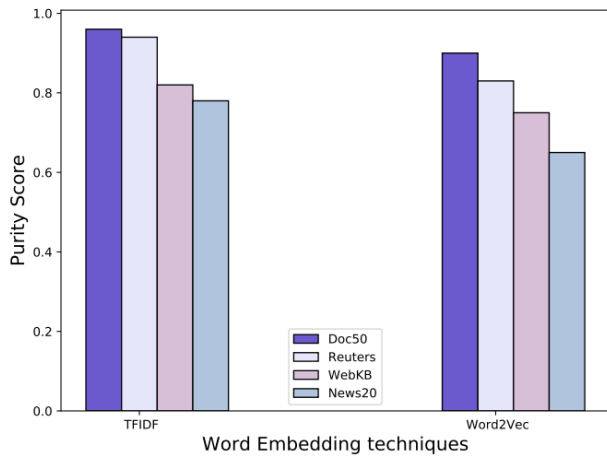


FIGURE 4. Result comparison of different word embedding (without stop words).

The comparison results of two embedding method is presented in Table 9 with stop words. It shows that proposed method has highest purity score with TF-IDF as compared to W2V.

Table 8 and 9 presents purity results of hybridization of algorithm without using stop words and with stop words for all datasets (Doc50, Reuters, WebKB and News20) with respect to different word embedding techniques. The result states that k-means++ - BH using TF-IDF has obtained 0.96 % purity in Doc50 dataset 0.94% on Reuters dataset, 0.82 on WebKB dataset and 0.78% on News20 datasets respectively. Whereas, from the combination of k-means++ - BH using W2V consumes 0.9% purity in Doc50 dataset 0.83% on Reuters dataset, 0.75 on WebKB dataset and 0.65% on News20 datasets. Results clearly shows that combination of k-means++ - BH using TF-IDF embedding performs much better on these datasets as compared to W2V.

The pictorial representation of result comparison based on impact of two different embedding in terms of purity measure, shown in Figure 4,5.

Figure 4 shows that k-means++ + BH using TF-IDF has obtained 96% purity in Doc50 dataset 94 % on Reuters dataset, 82% on WebKB dataset and 78% on News20 datasets

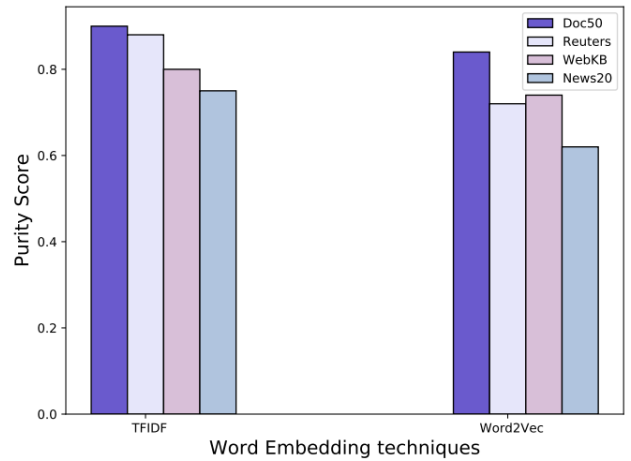


FIGURE 5. Result comparison of different word embedding (with stop words).

respectively whereas from the combination of k-means++ + BH using W2V consumes 90% purity in Doc50 dataset 83 % on Reuters dataset, 75% on WebKB dataset and 65% on News20 datasets without stop words.

Figure 5 shows the results of TF-IDF and W2V word embedding techniques with stop words. It shows that TF-IDF has highest purity results on all dataset as compared to W2V on all dataset.

Tables 4, 5, 6, 7, 8 and 9 have reported two key findings:

i.) First, is using word embedding whether with or without stop words, has a considerable impact on the result.

Working without stop words in documents reduces the number of features, which could result in a slight computational benefit. However, eliminating stop words and keeping stop words in data mainly depends upon the used datasets and the addressing problem but they should be removed if they are overused in data and reduces the effect of other important terms.

ii.) Second, is word embedding TF-IDF give much better results as compared to W2V embedding.

Due to the reason of used data with less semantic information, W2V performs well on data having terms which are included in its pre-trained model. Whereas TF-IDF gives results based on keyword occurrence in data. Deciding which embedding method to use mainly depends on the datasets as well as the problem being tackled. It has been evidence to literature that TF-IDF achieves better results than W2V embedding.

2) HYBRID CLUSTER IDENTIFICATION

Now with the results of cluster identification part following mentioned research question is considered for all datasets;

3) What would be the optimal value of the cluster k after the word embedding has been performed?

In this phase, for k-value identification two measures are used; one is Elbow method [45] and other is Silhouette score and then resultant is divided by 2 and get proposed k-value.

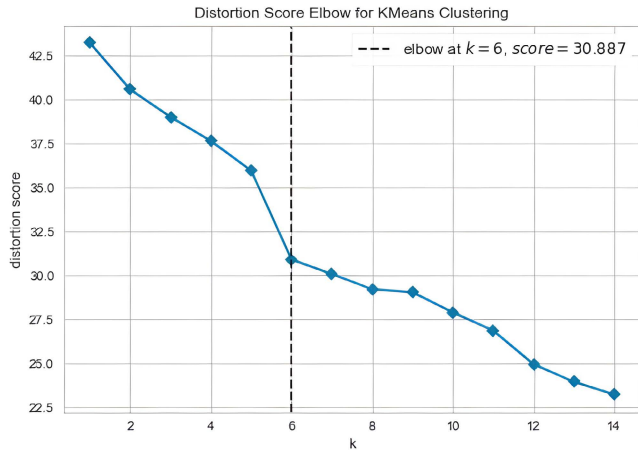


FIGURE 6. Elbow analysis of Doc50 Dataset.

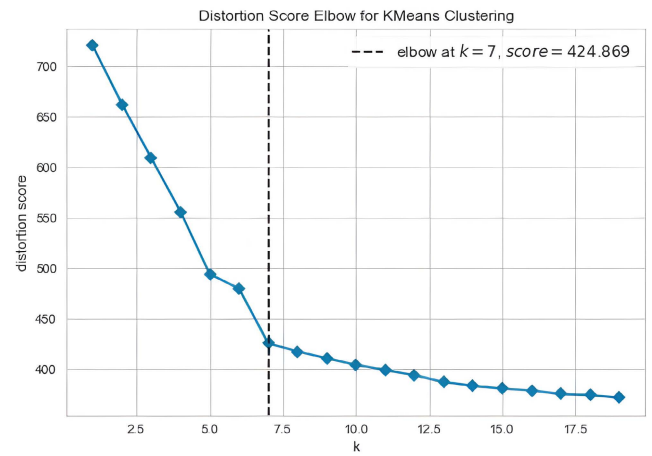


FIGURE 8. Elbow analysis of Reuters Dataset.

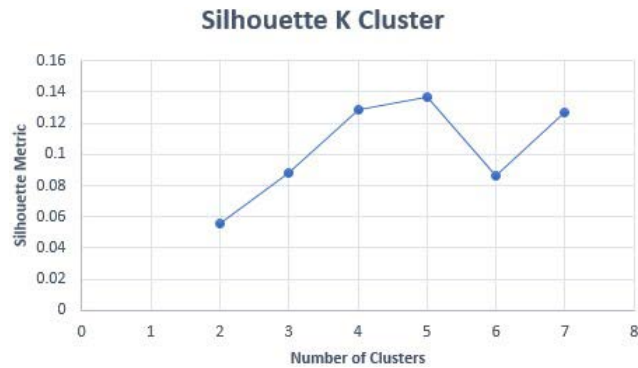


FIGURE 7. Silhouette analysis of Doc50 Dataset.



FIGURE 9. Silhouette analysis of Reuters Dataset.

As reported, research proves that without stop words, data produces better result as compared to data with stop words. So, in presented study, we have find k-value identification of data having no stop-words in data.

α: K-VALUE ANALYSIS ON ALPHA-NUMERIC DATASETS

K-value identification is performed by using two methods: Elbow method and Silhouette method for all datasets.

Figure 6, obtained the value 6 as the most optimal number of clusters by the Elbow method. As the number of clusters increases, the distortion score will start to decrease in a linear manner. The graph begins to move almost parallel to the X-axis at this point. The optimal k-value is the one that corresponds to this point. Therefore, for the given Doc50 dataset, concluded that the optimal number of clusters is 6.

Figure 7, find the value 5 as the most optimal number of clusters for a given Doc50 dataset, as it has the maximum silhouette score. Elbow methods determine the k value 6 and Silhouette score identify k-value as 5. In this case, proposed model use k=5 as an optimal number of clusters.

As the number of clusters increases, the distortion score will start to decrease in a linear manner. At point 7 in Figure 8, the graph will rapidly change and create an elbow shape. Therefore, for the given Reuters dataset, concluded that the

optimal number of clusters is 7. From Figure 9 obtained the value 19 as the most optimal number of clusters as it has the maximum silhouette. Figure 8 and 9 illustrates that, k-value identification is performed by using hybrid methods. In this case, proposed model use k=13 as an optimal number of clusters.

Figure 10, obtained the value 3 as the most optimal number of clusters by the Elbow method. At point 3 elbow found, concluded that the optimal number of clusters is 3.

From Figure 11, obtain the value 9 as the most optimal number of clusters as it has the maximum silhouette score. Figure 10 and Figure 11 displays, k-value is 3 by the Elbow methods and Silhouette score identify k-value as 9. In this case, proposed model gives k=6 as an optimal number of clusters.

The same K value identification procedure is performed on news 20 dataset and shown in Figure 12 and 13.

Figure 12, obtained the value 13 as the most optimal number of clusters by the Elbow method. At point 13 elbow found, concluded that the optimal number of clusters is 13.

From Figure 13, obtain the value 29 as the most optimal number of clusters as it has the maximum silhouette score.

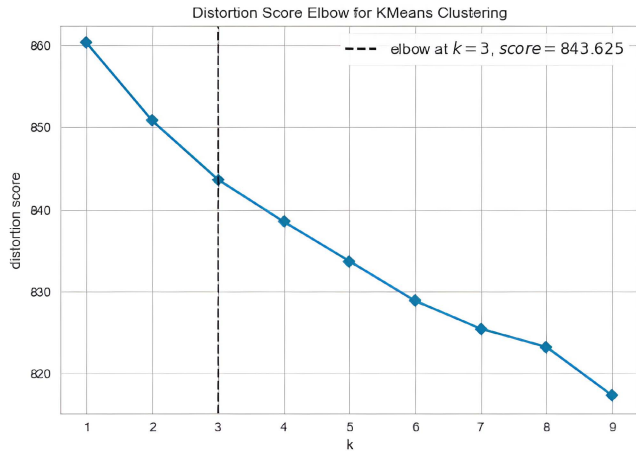


FIGURE 10. Elbow analysis of WebKB Dataset.

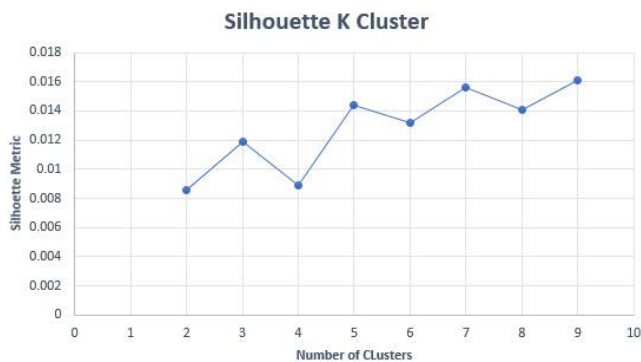


FIGURE 11. Silhouette analysis of WebKB Dataset.

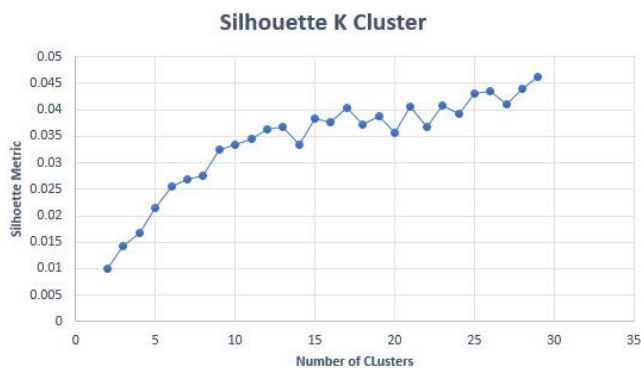


FIGURE 12. Silhouette analysis of News20 Dataset.

Figure 12 and 13 displays, k-value is 13 by the Elbow methods and Silhouette score identify k-value as 29. In this case, proposed model gives $k=21$ as an optimal number of clusters.

b: K-VALUE ANALYSIS ON NUMERIC DATASETS

This study also performed K value analysis on numeric dataset to find the number of clusters in each dataset. The proposed method use Elbow analysis and Silhouette analysis to find the optimal number of K in numeric data. The graph of

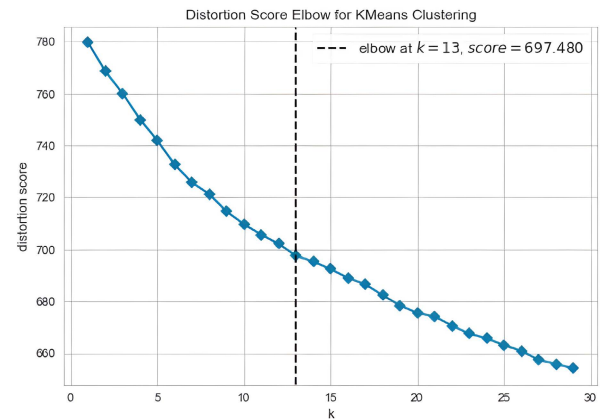


FIGURE 13. Elbow analysis of News20 Dataset.

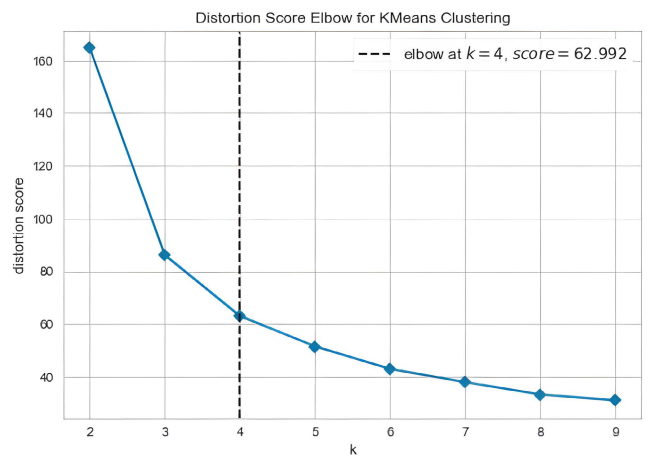


FIGURE 14. Elbow analysis of Iris Dataset.

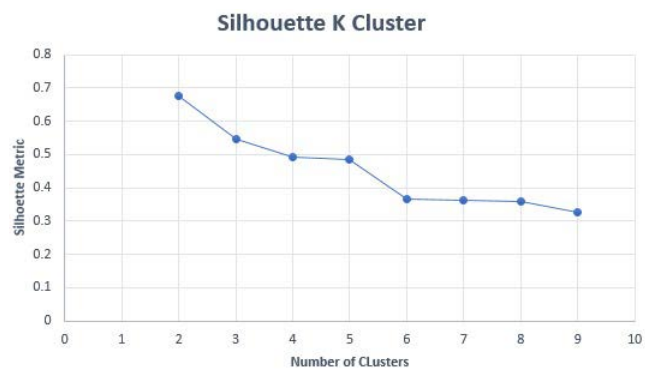


FIGURE 15. Silhouette analysis of Iris Dataset.

elbow and silhouette analysis are shows in Figure 14,15,16, and 17.

Figure 14, obtained the value 4 as the most optimal number of clusters by the Elbow method. At point 4 elbow found, concluded that the optimal number of clusters is 4.

From Figure 15, obtain the value 2 as the most optimal number of clusters as it has the maximum silhouette score. Figure 14 and 15 displays, k-value is 4 by the Elbow methods

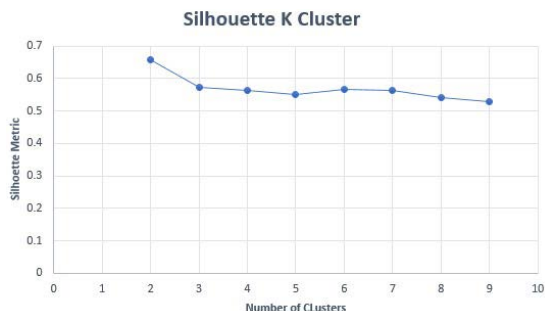


FIGURE 16. Silhouette analysis of Wine Dataset.

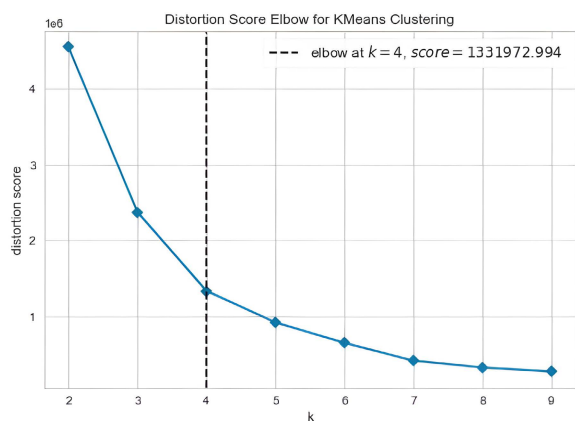


FIGURE 17. Elbow analysis of Wine Dataset.

TABLE 10. Result of Various method on each dataset.

Datasets	k-mean	k-mean++	BH	k-mean++-BH	Hybrid BH
Doc50	0.82	0.86	0.94	0.96	0.92
Reuters	0.87	0.88	0.90	0.94	0.95
WebKB	0.76	0.77	0.80	0.82	0.83
News20	0.6	0.64	0.63	0.78	0.81

and Silhouette score identify k-value as 2. In this case, proposed model gives k=3 as an optimal number of clusters.

From Figure 16, obtain the value 2 as the most optimal number of clusters as it has the maximum silhouette score.

Figure 17, obtained the value 4 as the most optimal number of clusters by the Elbow method. At point 4 elbow found, concluded that the optimal number of clusters is 4. Figure 17 and 16 displays, k-value is 4 by the Elbow methods and Silhouette score identify k-value as 2. In this case, proposed model gives k=3 as an optimal number of clusters.

c: PURITY RESULT COMPARISON OF DIFFERENT METHODS FOR ALPHA-NUMERIC DATASETS

Table 10 presents the comparative analysis of the performance of different methods with proposed Hybrid BH algorithm. The four alpha-numeric datasets are used for evaluating the results of proposed algorithm. For every dataset, each algorithm run individually according to the mentioned parameter setting in 3 Table. It is revealed that hybridization of algorithm (k-mean++-BH) achieves maximum purity for

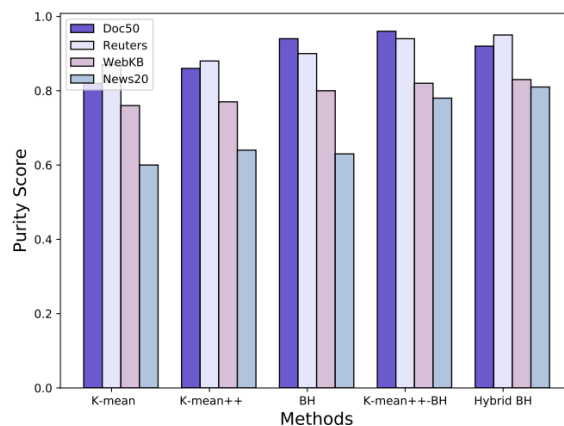


FIGURE 18. Purity result comparison of different methods for alpha-numeric datasets.

TABLE 11. Improvement percentage of proposed method over existing Black hole.

Dataset	Percentage of improvements
Doc50	2%
Reuters	4%
WebKb	3%
News20	15%

all datasets. In proposed algorithm, two improvements are inculcated to address the issues related to traditional BH algorithm [19].

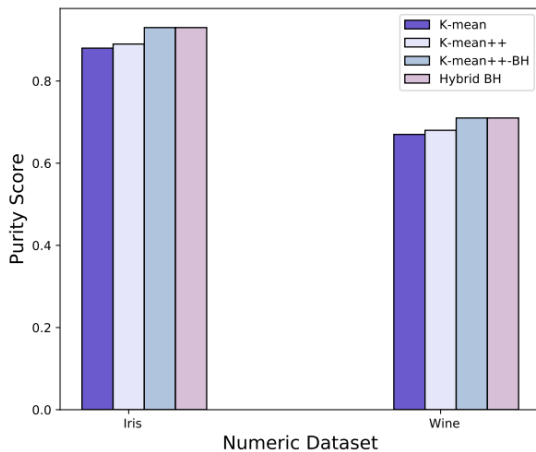
These issues are convergence rate and diversification. Every execution of k-mean++-BH algorithm consists of k-mean++ algorithm followed by BH and finally optimal solution is generated after specified number of parameter setting. In proposed algorithm, candidate solutions are generated by heuristic algorithm, exploration process of Hybrid Black Hole explores search space efficiently. Recently used BH uses global optimal solution through standard k-mean (locally optimum) solutions. However, sometimes locally optimal solution cannot converge on globally optimal solution. To improve diversification and obtain global optimum solution, proposed method provide an optimal solution through the interaction of multiple local best solutions. Every local solution interprets as a star, and the best solution among all the best local solutions is selected called black hole. Further, proposed Hybrid BH algorithm is used to optimize the candidate solution of heuristic algorithm and determines the global best solution.

Figure 18 displays the graphical view of the results by each method. Experimental analysis represents that proposed method performs better than existing methods.

Table 11 shows the overall percentage of improvement of our proposed method on all dataset as compared to existing method. It shows that performance on each dataset improves significantly. The performance of proposed method is increased 13% on news20 dataset that is highest improvement of our proposed method.

TABLE 12. Comparison of purity results of different methods (numeric datasets).

Datasets	k-mean	k-mean++ (Fix K)	BH	k-mean++- BH(unknown k)
Iris	0.88	0.89	0.93	0.93
Wine	0.67	0.68	0.71	0.71

**FIGURE 19.** Purity result comparison of different methods for numeric datasets.

d: PURITY RESULT COMPARISON OF DIFFERENT METHODS FOR NUMERIC DATASETS

Table 12 shows purity results of different existing methods with proposed model. It is noticed that all clusters of Wine and Iris datasets are non linear in nature. Due to non linearity, the clusters are not well separated by heuristic algorithms individually. The proposed hybrid algorithm effectively assigns the data objects to clusters. One of the challenge in existing BH algorithm [19] is the input data type. Recently, the algorithm was only accept textual data. Another flaw in the existing model is that it doesn't choose which cluster k to form automatically.

Figure 19 shows the graphical view of the results achieved by each method. Experimental study depicts that proposed method performs better than existing methods.

C. PERFORMANCE ANALYSIS BASED ON INTERNAL MEASURE (Silhouette Score)

The importance of a clustering result can be hard to determine, especially for vectors representing word. Clustering on labelled data is the best way to determine whether or not a clustering method is valid. After this, the original labels and the estimated labels can be compared. The problem with text data clustering is that in most of the cases labelled data is unavailable. Moreover, on some datasets determining what makes good clustering is extremely challenging. Silhouette score is a measure of how close each point in one cluster is to the points in other clusters. This metric is the most important performance metrics because, in a real scenario

TABLE 13. Silhouette score of each datasets using proposed model.

Datasets	Hybrid BH(TF-IDF)	Hybrid BH(W2V)
Doc50	0.14	0.085
Reuters	0.43	0.39
WebKB	0.18	0.064
News20	0.045	0.105

TABLE 14. Silhouette score of numeric datasets using proposed model.

Datasets	Hybrid BH
Iris	0.6
Wine	0.57

of clustering, true labels are not available to us. Silhouette co-efficient mainly determines the quality of clusters without requiring external labels. For all the used datasets, the values of this metric for the proposed clustering model are reported in Table 13 and 14, respectively.

1) PROPOSED MODEL SILHOUETTE SCORE FOR ALPHA-NUMERIC DATASETS

Table 13 presents Silhouette score of proposed model on the four alpha-numeric datasets, respectively. This measures is used to calculate the dis-similarity of clusters.

2) PROPOSED MODEL SILHOUETTE SCORE FOR NUMERIC DATASETS

From Table 14, we observed silhouette score of numeric datasets, Silhouette score of 0.6 is reported for Iris dataset whereas Silhouette score of 0.57 is achieved on Wine dataset. The results clearly show the compactness of formed cluster by proposed model.

V. CONCLUSION

With the rapid growth of document collections available in the field of information retrieval, organizing a large number of text documents is a core problem in the field of data mining. The process of grouping documents with similar properties/content, known as document clustering, is an important part of document organization and management. In document clustering, the documents are organized without the intervention of human, fast information retrieval, topic extraction and filtering, so it is similar to data clustering. The most well-known algorithm used for clustering is k-means but due to the certain problems like the efficacy of k-means is dependent on initial seeds chosen for clustering. Another problem is, k-means do not guarantee to form global optimal clusters. It easily gets trapped into local optimal clusters formed and hence could not improve results thereafter and determination of number of clusters k is not handle automatically and centroid initialization in k-means is random so clustering result under this method is less efficient. Document clustering is gaining popularity as an important and needed technique for un-supervised document organization and faster information retrieval. In this work our aims to automatically group

related documents into clusters. To evaluate the performance of clustering, two measures, Purity and Silhouette score are calculated and then, based on the results of external and internal measures, the performance of clustering is compared with base paper results. Experimental results are reported in two context with and with-out using stop words in data. Based on the type of cluster analysis used in this study, it can be concluded that the final results are primarily influenced by three factors which are: the document representation, the distance or similarity measures considered, and the fine hyper parameter tuning of the clustering algorithm itself. The research findings confirms that the proposed algorithm outperforms the previous black hole algorithm and offers the optimal global solution or close to optimal global. The limitations of this work are; we have listed results on using default distance measure of k-means same as our base paper. Proposed work can be extended to analyze the performance of text document clustering algorithms for different similarity measures with different document datasets and to provide the best combination of clustering algorithm with suitable similarity measures for different datasets. The above mentioned concern, have a significant impact on any text clustering algorithms. The presented work is experimented with two different document representation techniques. Further enhancements such as using a more advanced and complex word embedding can also be used. This might result in dealing with more complex documents such as literary articles/poems and dramas, online news, scientific papers and blogs. The future studies can also apply this work with different cluster evaluation measures. Now a days many organizations produce, collect, and analyze the huge amount of data. This huge amount of data has the characteristics such as variety, volume, and velocity etc. The K mean++ with other variant of BH can be used for cluster analysis to find the patterns in Big Data.

REFERENCES

- [1] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, p. 664, Feb. 2021.
- [2] Y. Fakir and J. E. Iklil, "Clustering techniques for big data mining," in *Proc. Int. Conf. Bus. Intell. Cham, Switzerland: Springer*, 2021, pp. 183–200.
- [3] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017.
- [4] M. Afzali and S. Kumar, "Text document clustering: Issues and challenges," in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput. (COMITCon)*, Feb. 2019, pp. 263–268.
- [5] N. Kumar, S. K. Yadav, and D. S. Yadav, "An approach for documents clustering using K-means algorithm," in *Innovations in Information and Communication Technologies*. Cham, Switzerland: Springer, 2021, pp. 453–460.
- [6] I. Aljarah, M. Habib, H. Faris, and S. Mirjalili, "Introduction to evolutionary data clustering and its applications," in *Evolutionary Data Clustering: Algorithms and Applications*. Singapore: Springer, 2021, pp. 1–21.
- [7] P. P. Mohanty, S. K. Nayak, U. M. Mohapatra, and D. Mishra, "A survey on partitional clustering using single-objective metaheuristic approach," *Int. J. Innov. Comput. Appl.*, vol. 10, nos. 3–4, pp. 207–226, 2019.
- [8] M. Kalra, N. Lal, and S. Qamar, "K-mean clustering algorithm approach for data mining of heterogeneous data," in *Information and Communication Technology for Sustainable Development*. Singapore: Springer, 2018, pp. 61–70.
- [9] A. Rizwan, N. Iqbal, A. N. Khan, R. Ahmad, and D. H. Kim, "Toward effective pattern recognition based on enhanced weighted K-mean clustering algorithm for groundwater resource planning in point cloud," *IEEE Access*, vol. 9, pp. 130154–130169, 2021.
- [10] V. Kumar, J. K. Chhabra, and D. Kumar, "Performance evaluation of distance metrics in the clustering algorithms," *INFOCOMP J. Comput. Sci.*, vol. 13, no. 1, pp. 38–52, Jun. 2014.
- [11] A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," *Inf. Sci.*, vol. 222, pp. 175–184, Feb. 2012.
- [12] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford InfoLab, Stanford, CA, USA, Tech. Rep., Jun. 2006. [Online]. Available: <http://ilpubs.stanford.edu:8090/778/>
- [13] S. Mayani and S. Swarndeep, "A survey of text document clustering by using clustering techniques," *Int. Res. J. Eng. Technol.*, vol. 6, no. 12, pp. 1–5, Dec. 2019.
- [14] R. Lakshmi and S. Baskar, "DIC-DOC-K-means: Dissimilarity-based initial centroid selection for document clustering using K-means for improving the effectiveness of text document clustering," *J. Inf. Sci.*, vol. 45, no. 6, pp. 818–832, Dec. 2019.
- [15] K. M. Kumar and A. R. M. Reddy, "An efficient k-means clustering filtering algorithm using density based initial cluster centers," *Inf. Sci.*, vols. 418–419, pp. 286–301, Dec. 2017.
- [16] K. Singh, D. Malik, and N. Sharma, "Evolving limitations in K-means algorithm in data mining and their removal," *Int. J. Comput. Eng. Manage.*, vol. 12, no. 1, pp. 105–109, 2011.
- [17] C. Xiong, Z. Hua, K. Lv, and X. Li, "An improved K-means text clustering algorithm by optimizing initial cluster centers," in *Proc. 7th Int. Conf. Cloud Comput. Big Data (CCBD)*, Nov. 2016, pp. 265–268.
- [18] L. Abualigah, A. H. Gandomi, M. A. Elaziz, A. G. Hussien, A. M. Khasawneh, M. Alshinwan, and E. H. Houssein, "Nature-inspired optimization algorithms for text document clustering—A comprehensive analysis," *Algorithms*, vol. 13, no. 12, p. 345, Dec. 2020.
- [19] M. Rafi, B. Aamer, M. Naseem, and M. Osama, "Solving document clustering problem through meta heuristic algorithm: Black hole," in *Proc. 2nd Int. Conf. Mach. Learn. Soft Comput. (ICMLSC)*, 2018, pp. 77–81.
- [20] R. Chouhan and A. Purohit, "An approach for document clustering using PSO and K-means algorithm," in *Proc. 2nd Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2018, pp. 1380–1384.
- [21] K. Lakshmi, N. K. Visalakshi, and S. Shanthi, "Data clustering using K-Means based on crow search algorithm," *Sādhanā*, vol. 43, no. 11, pp. 1–12, Nov. 2018.
- [22] Catherine Blake. (1998). *UCI Repository of Machine Learning Databases*. [Online]. Available: <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [23] M. Eskandarzadehalamdary, B. Masoumi, and O. Sojodishijani, "A new hybrid algorithm based on black hole optimization and bisecting k-means for cluster analysis," in *Proc. 22nd Iranian Conf. Electr. Eng. (ICEE)*, May 2014, pp. 1075–1079.
- [24] Y. Gupta and A. Saini, "A new swarm-based efficient data clustering approach using KHM and fuzzy logic," *Soft Comput.*, vol. 23, no. 1, pp. 145–162, Jan. 2019.
- [25] H. A. Abdulwahab, A. Noraziah, A. A. Alsewari, and S. Q. Salih, "An enhanced version of black hole algorithm via levy flight for optimization and data clustering problems," *IEEE Access*, vol. 7, pp. 142085–142096, 2019.
- [26] A. F. Jahwar and A. M. Abdulazeze, "Meta-heuristic algorithms for K-means clustering: A review," *PalArch's J. Archaeol. Egypt/Egyptol.*, vol. 17, no. 7, pp. 12002–12020, 2020.
- [27] D. S. Rajput, "Review on recent developments in frequent itemset based document clustering, its research trends and applications," *Int. J. Data Anal. Techn. Strategies*, vol. 11, no. 2, pp. 176–195, 2019.
- [28] M. Črepinšek, S.-H. Liu, and M. Mernik, "Exploration and exploitation in evolutionary algorithms: A survey," *ACM Comput. Surv.*, vol. 45, no. 3, pp. 1–33, Jun. 2013.
- [29] R. B. Wahyu and A. Vito, "Documents clustering using K-means algorithm," *IT Soc.*, vol. 3, no. 2, pp. 1–5, Feb. 2018.
- [30] C. Mageshkumar, S. Karthik, and V. P. Arunachalam, "Hybrid metaheuristic algorithm for improving the efficiency of data clustering," *Cluster Comput.*, vol. 22, no. S1, pp. 435–442, Jan. 2019.
- [31] P. Chandrasekar and M. Krishnamoorthi, "BHOHS: A two stage novel algorithm for data clustering," in *Proc. Int. Conf. Intell. Comput. Appl.*, Mar. 2014, pp. 138–142.
- [32] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *Proc. Int. Conf. Electr., Electron., Optim. Techn. (ICEEOT)*, Mar. 2016, pp. 61–66.

- [33] L. Ma and Y. Zhang, "Using word2 Vec to process big text data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 2895–2897.
- [34] D. Dua and A. Asuncion, "UCI machine learning repository," Tech. Rep., 2017.
- [35] G. Guo, L. Chen, Y. Ye, and Q. Jiang, "Cluster validation method for determining the number of clusters in categorical sequences," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 2936–2948, Sep. 2016.
- [36] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2020, pp. 747–748.
- [37] Y. Lei, J. C. Bezdek, S. Romano, N. X. Vinh, J. Chan, and J. Bailey, "Ground truth bias in external cluster validity indices," *Pattern Recognit.*, vol. 65, pp. 58–70, May 2017.
- [38] L. Bungum, "Unsupervised clustering of structured and unstructured text collections," Institutt for Datateknologi og Informatikk, Tech. Rep., 2021. [Online]. Available: <https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2759145>
- [39] A. Vysala and D. J. Gomes, "Evaluating and validating cluster results," 2020, *arXiv:2007.08034*.
- [40] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm," *Journal*, vol. 2, no. 2, pp. 226–235, Jun. 2019.
- [41] S. J. Nanda and G. Panda, "A survey on nature inspired metaheuristic algorithms for partitional clustering," *Swarm Evol. Comput.*, vol. 16, pp. 1–18, Jun. 2014.
- [42] S. Naeem and A. Wumaier, "Study and implementing K-mean clustering algorithm on English text and techniques to find the optimal value of K," *Int. J. Comput. Appl.*, vol. 182, no. 31, pp. 7–14, Dec. 2018.
- [43] A. Statman, L. Rozenberg, and D. Feldman, "K-means++: Outliers-resistant clustering," *Algorithms*, vol. 13, no. 12, p. 311, Nov. 2020.
- [44] S. Kumar, D. Datta, and S. K. Singh, "Black hole algorithm and its applications," in *Computational Intelligence Applications in Modeling and Control*. Cham, Switzerland: Springer, 2015, pp. 147–170.
- [45] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in K-means clustering," *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.

FAZILA MALIK received the M.C.S. and M.S. degrees in computer science from COMSATS University Islamabad at Attock Campus, Pakistan, in 2019 and 2022, respectively. Her research interests include natural language processing, data mining, and machine learning. Her majors are in evolutionary computing and optimizing algorithm applications.



SALABAT KHAN received the Ph.D. degree from FAST, Islamabad. He is a currently working as a Brain-Pool Overseas Researcher at the Big Data Laboratory, Jeju National University, South Korea. He is also an Associate Professor at COMSATS University Islamabad, Attock Campus, Pakistan. His research interests include evolutionary computation, medical image processing, machine learning, federated learning, big data analytics, recommender systems, and the IOT.



ATIF RIZWAN received the Bachelor of Science (B.Sc.) degree from the University of the Punjab, Lahore, Punjab, Pakistan, in 2015, and the M.C.S. (2 years) and M.S. degrees in computer science from COMSATS University Islamabad, Attock Campus, Punjab, in 2018 and 2020, respectively. He is currently working as a Ph.D. Researcher at the Department of Computer Engineering, Jeju National University, Republic of Korea. He was awarded a fully-funded scholarship for the entire duration of his Ph.D. studies. He has good industry experience in mobile and web application development and testing. His research interests include applied machine learning, data and web mining, analysis, and optimization of core algorithms and the IoT-based applications.



GHADA ATTEIA received the Ph.D. degree in electrical & computer & geomatics engineering from the Schulich School of Engineering, University of Calgary, Calgary, AB, Canada, in 2015.

From 2010 to 2015, she was a Research Assistant with the Geomatics Engineering Department, University of Calgary. Since 2017, she has been an Assistant Professor with the Information Technology Department, CCIS, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include artificial intelligence, machine and deep learning, and new and renewable energy. She was a recipient of the Queen Elizabeth II Doctoral Award, in 2012 and 2013, and the L. R. (Dick) Newby Memorial Doctoral Award from the University of Calgary, in 2014



NAGWAN ABDEL SAMEE received the B.S. degree in computer engineering from Ein Shams University, Egypt, in 2000, and the M.S. degree in computer engineering and the Ph.D. degree in systems and biomedical engineering from Cairo University, Egypt, in 2008 and 2012, respectively. Since 2013, she has been an Assistant Professor with the Information Technology Department, CCIS, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include data science, machine learning, bioinformatics, and parallel computing. Her awards and honors include the Takafull Prize (Innovation Project Track), Princess Nourah Award in Innovation, the Mastery Award in predictive analytics (IBM), the Mastery Award in Big Data (IBM), and the Mastery Award in Cloud Computing (IBM).

...