Received January 6, 2022, accepted February 11, 2022, date of publication February 16, 2022, date of current version March 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3152159

Feature Discrimination of News Based on Canopy and KMGC-Search Clustering

MOBEEN SHAHROZ[1], MUHAMMAD FAHEEM MUSHTAQ[1], RIZWAN MAJEED[2], ALI SAMAD[3], ZAIGHAM MUSHTAQ[4], AND UROOJ AKRAM[1]

[1]Department of Artificial Intelligence, The Islamia University of Bahawalpur, Bahawalpur, Punjab 63100, Pakistan
[2]Directorate of Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Punjab 63100, Pakistan
[3]Department of Data Science, The Islamia University of Bahawalpur, Punjab 63100, Pakistan
[4]Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, Punjab 63100, Pakistan

Corresponding authors: Muhammad Faheem Mushtaq (faheem.mushtaq@iub.edu.pk) and Zaigham Mushtaq (zmqazi@gmail.com)

This work was supported by the Sir Sadiq Research and Computing Center, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan.

ABSTRACT The internet provides a very vast amount of sources of news and the user has to search for desirable news by spending a lot of time because the user always prefers their related interest, desirable and informative news. The clustering of the news article having a great impact on the preferences of the user. The unsupervised learning techniques such that K-means Clustering and Spectral Clustering are proposed to categorize the news articles by extracting discriminant features that help the user to search and get informative news without wasting time. The BBC news articles dataset is used to perform experiments that consist of 2225 news articles. The TF-IDF feature extraction technique is used with K-means clustering and Spectral clustering to get the most similar clusters to categorize the news articles in respective domains. Those domains are sports, tech, entertainment, politics, and business. The clustering algorithms are evaluated using adjusted rand index, V-measure, homogeneity score, completeness score, and Fowlkes mallows score. The experimental results illustrated that K-means clustering performs better than spectral clustering using the TF-IDF feature extraction approach. But to improve the results the canopy centroid selection is used with the grid search optimization technique to optimize the results of the Kmeans and named its as a K-Means using Grid Search based on Canopy (KMGC-Search). The experimental results shows the proposed approach can be used as a viable method for the categorization of news articles.

INDEX TERMS News categorization, K-means, clustering, canopy, machine learning, TF-IDF, grid search.

I. INTRODUCTION

The news leaves a great impact [1] on the thoughts of the people because news presents those things of the world that are hidden from the local people [2]. It has the power to change the perspective and preferences of the people [3] that are very knowledgeable to gain information and to get updates about the changes in the world. The circle of the local user is not broad enough in the old days [4] to get important updates about the world. In this age, a local user is surrounded by mobile electronic devices [5] and digital media like the internet that has a vast amount of data in different domains [6] about events, politics, supports, business, technology, etc. Now, the user can interact with the whole world and get updates in seconds about every event that occurs in this world [7]. The combination of mobile devices and the internet provide the news information direct to the user because of this youngster as well as adult lost interest in reading newspapers [8]. The news organizations feel the impact of the internet news on the perspective of the user and it also attracts the user more than the televisions broadcast or the newspaper.

A survey performed by the Pew research center that defines the growing number of news losing the interest of the users [9]. The internet provides news highlights and quick updates [10] that attracts people. Today almost every person has a mobile device that is almost connected to the internet. The people can easily interact with the news updates and gain knowledge about the world at any time. The frequent users and researchers need to get a useful source of information

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

without visiting several links and websites [11]. The usage of information resources is different in different countries and differ geographically [12]. The understanding of these differences in the popularity of news articles among the audience is important when targeting a specific audience [13]. The audience preferences among selected news and extracting records would be useful for auditors and newscasters [14]. The rapid growth of data and news needs to be categorized that helps the user to directly interact with the specific desired news [7]. There document clustering [15] is the most popular approach that is adopted by the companies to search more quickly in search engines, to extract information from in a blink of an eye and to maintain the data and information in a structured form. The most beneficial advantage to categorize the unlabeled data is to identify the domain of the data. The extraction of useful and valuable information from the raw data known as data mining and clustering is one of the data mining methods of extracting useful information.

Data mining is the extraction of knowledgeable or desired information from the raw data [16]. There are different fundamental methods like arranging the raw data into the form of rows and columns or entities and features. The different machine learning [17] techniques are adopted to extract those features and entities from the raw data. The clustering is the concept of unsupervised machine learning used to create homogeneous groups by categorizing the data, according to the similarity measure that is based on different parameters like Manhattan distance [18], Euclidean distance [19], cosine similarity and probability-based similarity. There is some application of clustering mentioned here like segmentation [20], discretization [21], and aggregation [22] of the data. It is difficult to extract the desirable and knowledgeable information from the huge amount of data in a blink of an eye [23]. The search engine optimizations [24] are also based on the clustering that extracts the relevant results from the server according to the query of the user. The servers contain billions or trillions of terabytes of data [25]. If a user searching manually in servers to extract desirable information then it takes days or months that is not a good approach.

There are two document clustering algorithms K-means clustering [26] and spectral clustering are proposed in this study to achieve the aims of manage and label the upcoming news according to the domains. The BBC news dataset is used that contains 5 classes such that politics, sport, tech (technology), entertainment, business and 2225 number of the news article. The news articles are in raw form that needs to filter unnecessary data from it for further processing. The preprocessing methods are applied such as the conversion of letter from capital case to small letter, removal of punctuation, word tokenization, stop word removal and stemming. After preprocessing the filtered news articles, the train test split is applied on the news article dataset to divide it into two portions one for the training and another for the testing. The Term Frequency-Inverse Document Frequency (TF-IDF) [27] is adopted for the feature extraction to training the data and transform those features into the vector form for further processing. The K-means and Spectral clustering algorithms are applied to these feature vectors to train and create clusters of the relevant news articles without labels by learning the pattern of the news articles. The performance is evaluated by different evaluation parameters such that Adjusted rand index (ARI), V measure (VM), Homogeneity score (HS), Completeness score (CS) and Fowlkes mallows (FM) score. To improve the results of the proposed approach the K-Means model implemented using Grid Search based on Canopy centeroid selection techniques (KMGC-Search) that shows promising results. The categorization of the news articles into clusters are created according to the domains of the news that helps the user to search the desirable news efficiently as well as it also optimizes the search engine to show relevant result according to the user. The contributions of this paper are following below.

- The paper aims to categorize the news articles to optimize the search engines and browsers to develop the news interest in the people that are achieved successfully by proposed clustering methods.
- The preprocessing methods like converting all capital case letters into small, remove the punctuation, word tokenization, stop word removal and stemming is applied to extract the useful data from the BBC news article dataset that is in raw form.
- The feature extraction method TF-IDF is used to extract the weighted feature from the data.
- The K-means Clustering and Spectral Clustering algorithms are used for the unknown news to categorize them into clusters by label the unlabeled news articles.
- Further, KMGC-Search technique is proposed to improve the results based on Canopy centeroid selection technique and Grid Search hyper parameter tuning methods are used.
- The analyses of the performance evaluations are executed with Kmeans clustering, spectral clustering and the proposed KMGC-Search methodology to validate the results using ARI, VM, HS, CS and FM Score evaluation parameters.

The structure of this paper is organized as follows: Section 2 presents the efforts of the researchers in the previous related studies. Section 3 discusses the material and methods that are used in the proposed research. Section 4 illustrates the functionality of the proposed methodology. Section 5 explains the results and discussion based on the proposed method and Section 6 presents the conclusion of this research.

II. RELATED WORK

The news categorization is attracted to the concentration of the researchers from past decades. There were different text mining approaches adopted such as information retrieval [23], natural language processing [28], information extraction from the text, text summarization [29], supervised learning methods [30], unsupervised learning methods [31], probabilistic methods for text mining [32], text streams [33] and social media mining [34], opinion mining, sentiment analysis [35] and biomedical text mining [36] for the purpose categorization of text data. The clustering is one of the most popular data mining technique that is mostly used for unsupervised machine learning. The unlabeled data that has no labels, need to categorize the data based on a similarity measure [37]. The news is the biggest source of information on the internet that provides so many text mining research issues [38]. Several clustering algorithms were used by previous studies for the news categorization such that K-means variations [26], spectral algorithms [39], hierarchical clustering [40], vector space models, methods of dimensionality reduction [41], generative algorithms [42] and methods of phrasing [43]. A classical approach vector space modeling is used in the field of homogeneity topics, K-means clustering is the partitioning and hierarchical clustering technique and its extensions are also popular as unsupervised learning. These methods need to specify the number of clusters and random initializations make them not suitable in dealing with noise and outlier values [44]. This is the reason those algorithms are less effective on heterogeneous data where spectral clustering shows outstanding results because it extracts the number of clusters according to the data spatiality. The document clustering system proposed the hierarchical clustering based on the occurrences of the words presentation. The Jaccard similarity [45] was used there to calculate the similarity measure between the clustering documents.

The reviews, feedback or comments are presenting the emotions and thoughts of the people [46]. The multinational or local companies are interested to analyze the opinion of the customers [47] through sentiment analyses of the reviews. The researchers previously put their efforts to categorize the news based on the sentiments [48] that also helps the customer to get knowledge about the company and also helps the company to grow according to customer preferences. The clustering techniques are well-known techniques to group the data into different clusters according to the sentiment analyses and also create sub clusters. There are K-means clustering [15] and decision tree algorithm with TF-IDF feature extraction techniques [27] used and the performance is evaluated by the precision, recall, f1 score and accuracy. The text document clustering is used to group the records based on similarity to provide the clients to extract information rapidly. The document clustering is consistently used as a device in information retrieval systems to improve retrieval and navigation of enormous information [23]. Mostly document clustering is used in the search engines for browsing the collection of documents and respond to the user query. The contrast of K-means clustering with Euclidean and Manhattan distance [18] and K Medoids clustering [19] proposed by using WEKA and java programming. The evaluation results illustrate that the K Medoids perform better than the K-means Clustering [49].

A large number of news articles are published on various web sources every year but in the past few years, it has a dramatic increase that contains multimodal information [50]. The article writers, readers and media organizations face the issues of categorizing the large amounts of text data to distinguish the important or desired topics and events all around the world. The classification of the news article is referred to as a document clustering problem [51]. Machine learning is learning from a given number of training samples that are used for the classification by performing prediction. The document clustering is applied in numerous classification and clustering tasks like indexing of document [50], information extraction, document filtering, and disambiguation [52] of word sense. The classification is performed on Reuters, BBC [53], and The Guardian news article datasets [50] that contain multimodel including textual information as well as visual like video and images, which is already categorized into four classes Lifestyle Leisure, Business Finance, Sports and Science, and Technology. The several classification experiments have been performed by using supervised machine learning models like Logistic Regression, Naive Bayes [54], Rocchio, Log-Linear models, SVM [55] and Random Forest model with N-gram feature extraction technique [27] as well as using late fusion strategies to categorize the news and topics.

III. MATERIAL AND METHODS

The categorization of the news articles based on KMGC-Search Clustering system is proposed in this paper. The BBC news dataset is used that consists of news articles in the form of raw text. To enhance the training of the models, the dataset needs to filter unnecessary data such as the conversion of letters from capital case to lower case, stopwords removal, punctuation removal, word tokenization and stemming. After preprocessing the TF-IDF is applied to the filtered data to extract the most discriminant features in the form of vectors. The K-means clustering and spectral clustering are applied to these vectors to train and learn to create the most efficient news clusters. The Canopy centeroid selection and Grid search hyper parameter tuning is further added to improve the results of the clustering to categorize the news much more efficiently. In the last, the performance evaluation of the clustering model is measured in terms of adjusted rand index, V measure, homogeneity score, completeness score, and Fowlkes mallows score. These are the methods used in this research discussed in detail below.

A. BBC NEWS DATASET

The BBC News dataset was provided by the BBC organization as a benchmark dataset for research purposes. This dataset consists of 2225 number of news articles corresponding to different stories [53] that belong to five categories such that tech, sport, entertainment, business, and politics as shown in Table 1. There are five randomly selected samples of the news article that presents each category of news.

Figure 1 illustrates the category-wise distribution of the BBC news dataset. The maximum number of documents belong to sports and business category that contain 511 and

TABLE 1. BBC news dataset samples.

Category	News data
Tech	The trend of freeing up multimedia so that people can watch what they want when they want.
Business	Japanese mogul arrested for fraud one of japan s best-known businessmen was arrested on thursday on charges...
Sports	Newry to fight cup exit in courts newry city are expected to discuss legal avenues on friday regarding overturning their ejection from the nationwide irish cup...
politics	Campbell returns to election team ex-downing street media chief alastair campbell will return to the fold to strengthen labour s general election campaign the party has confirmed...
Entertainment	Dance music not dead says fatboy dj norman cook - aka fatboy slim - has said that dance music is not dead but...

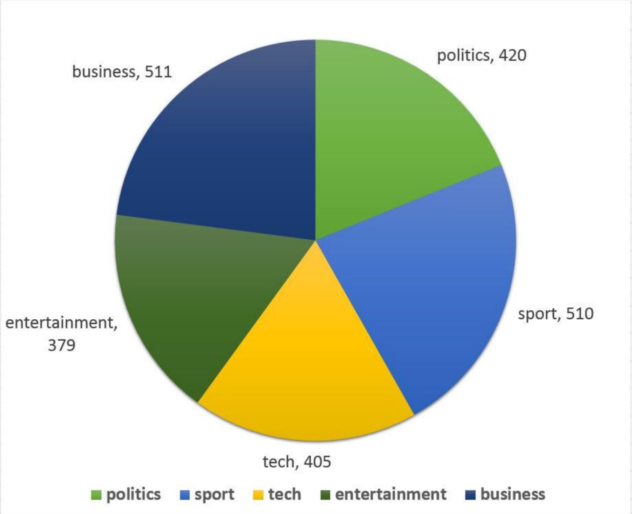

FIGURE 1. BBC news dataset distribution based on classes.

510 number of news articles out of 2225, which is 46% of the total dataset. Besides, other news categories are entertainment, politics and tech categories that contain 379, 420 and 405 documents that are about 54% of the dataset.

B. FEATURE ENGINEERING

Feature engineering is a process that makes the model computational processing works by extracting features from the data in vector form. The features are attributes or properties shared by all of the independent units. The TF-IDF is used in this research [27] that consists of Term Frequency (TF) times Inverse Document Frequency (IDF) of documents. It is the cross product of the TF and IDF. The term-document matrix is built with the table of frequencies that are the occurrences of the terms in each document. It extracts the feature by applying the weighting functions to estimate the relative importance of a term within the document and set of documents. The TF-IDF illustrates the importance of the word in the document. Since each document contains different words, this table is a high dimensional sparse matrix.

$$TF(t) = \frac{(Number\ of\ times\ term\ t\ appear\ in\ \ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)} \tag{1}$$

$$IDF(t) = log\Big(\frac{(Total\ number\ of\ \ documents)}{(Number\ of\ documents\ with\ term\ t\ in\ it)}\Big) \tag{2}$$

$$\tilde{}IDF(t) = log\Big(\frac{(1+n)}{(1+df(t))}+1\Big) \tag{3}$$

where n is the total number of documents in the document set, and $df(t)$ is the number of documents in the document set that contains the term. The $TF-IDF$ is the cross product of the TF and IDF as shown in equation 4.

$$TFIDF(t,d) = TF(t,d) * IDF(t) \tag{4}$$

C. UNSUPERVISED LEARNING

Data mining is the process of extracting information from the raw data that needs to arrange the data according to labels or features. The most common technique that is used to label the data without any supervision that means this data does not have any true labels. The unsupervised machine learning [31] technique is used to perform training on the data without true labels that needs to learn from the features of the data based on similarity and dissimilarity measures to create clusters.

1) K-MEANS CLUSTERING

K-means clustering is the one of the most popular techniques of unsupervised machine learning [19] that used to label the data without defined labels. The main aim is to find clusters of the relevant or similar objects by measuring the similarity or dissimilarity based on Euclidean distance [18] of the features. The number of clusters specified by the variable K. The algorithm processes the data iteratively to assign the object to the clusters that specified in initial steps on the bases of those extracted features. Those objects are assigned to the specific cluster by comparing their feature that how they are similar to each other. Each cluster contains similar objects, which have a similar type of properties or features.

$$\sum_{i=0}^{n} min_{\mu\in c}\Big(||x_i-\mu_j||^2\Big) \tag{5}$$

The K-means model contains three major steps. The first step is to choose the centroids that are the centers of the clusters. It distributes the given number of samples from the dataset into a specified K number of clusters. After initializing the K number of clusters, K-means consists of looping between the two other steps. The second step contains the process of assigning the objects to the nearest centroid. After complete the first step, in which every object assigns to its specific centroid. It became in the form of clusters. The third step is to calculate the location of the new centroid of each cluster on behalf of those assigned objects to a cluster. It repeats those last two steps until it reaches the threshold or until centroid does not move significantly. The difference between the centroid and the objects based on relevance concept and the Euclidean distance is used to measure the similarity in two objects [18] by measure the distance of the straight line between two points p and q as shown in equation 6.

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{6}$$

2) SPECTRAL CLUSTERING

The spectral clustering [56] performs the low-dimensional embedding of the affinity matrix between samples that are followed by a K-Means in the low dimensional space. It is especially efficient if the affinity matrix is sparse. The Spectral clustering requires the number of clusters to specify that extracts the spectrum of the similarity matrix from the data of clustering or multivariate statics to perform operations of dimensionality reduction. The similarity matrix created based on quantitative assessments of each pair of the objects that are given as input to produces better results with a small number of clusters.

D. CANOPY CENTEROID SELECTION

The Canopy is the pre-clustering technique used for initial initialization of centroid points to maximize the speed and efficiency of the clustering [61], [62]. The canopy clustering technique is used in this study it initiate the kmeans clustering technique to effectively create clusters. The canopy is dependent on the $T1$ and $T2$ distance points. $T1$ considers as a loose distance point and $T2$ considers as a tight distance point. $T1$ has to be greater than the $T2$ that will allocate the data points to the canopies. If the remaining points distance to the first point is less than the distance of $T1$ then assign it to the new canopy. Remove points from them as well if the distance of these point is also less than the $T2$. Repeat these step again again until single data point left to create cluster. The canopy algorithm generate the group of centroid for the kmeans clustering algorithm to create clusters efficiently based on these centroids.

E. GRID SEARCH HYPER-PARAMETER TUNING

The grid search hyper parameter tuning [63], [64] is used to autonomously optimize the hyper parameters for computational processing of the kmeans algorithm to create more effective clusters. The kmeans clustering based on following parameters n_clusters = 8, init = 'k-means++', n_init = 10, max_iter = 500, tol = 0.001, verbose = 0, random_state = int, copy_x = True, algorithm = 'elken'. There are multiple values of these parameters but those are selected after multiple number of iteration performed in the grid search technique to find the best parameters to imporve the clustering results.

F. PERFORMANCE EVALUATION

The performance evaluation is the process in which measures the model that how much or how accurately model works. In terms of machine learning, the most important measure is how efficiently machine-learning models perform prediction based on the relevance of true labels and predicted labels. But in case of unsupervised learning that means to create groups of related objects based on their features is called clusters, without any true labels. Those clusters have no true labels that make it a challenging part to evaluate those clusters. The different clustering models create clusters based on similarity or dissimilarity measures in between features that are evaluated based on internal and external evaluation parameters. The internal evaluations are those in which no true labels are available. The processing of some clustering models based on the principle of similarity that illustrates all the objects in a cluster are similar to each other or not and how many objects are not similar to the rest of the objects in the cluster. The external evaluations are those in which true labels are available that measure the data from a single class must belong to a single cluster. Based on the five classes that are present in this dataset, the number of clusters must be five and each cluster contains only single class data.

The ARI [57] measures a similarity score between two clusters based on the similarity in between two data points and count the pairs of these data points that belong to relevant or non-relevant clusters by comparing the predicted and true clusters. The raw RI score is then "adjusted for the chance" into the ARI score using equation 7.

$$ARI = \left(\frac{RI - RI_{expected}}{max(RI) - RI_{expected}}\right) \tag{7}$$

The Homogeneity score [58] of a cluster is to measure objects of each cluster that are all objects from each cluster belongs to a single class based on the ground truth. A permutation of a cluster labels not change the score value or results in any case that makes the metric independent of the absolute values of the labels.

$$h = 1 - \left(\frac{H(C|K)}{H(C)}\right) \tag{8}$$

The ground truth is given by cluster labeling is known as the Completeness score [58]. If all the objects of the same cluster belong to a single class that satisfies the completeness score. Homogeneity and Completeness score formally have given by:

$$c = 1 - \frac{H(K|C)}{H(K)} \tag{9}$$

$$H(C|K) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \left(\frac{n_{c,k}}{n}\right).log\left(\frac{n_{c,k}}{n}\right) \tag{10}$$

$$H(C) = -\sum_{c=1}^{|C|} \left(\frac{n_c}{n}\right).log\left(\frac{n_c}{n}\right) \tag{11}$$

where $H(K|C)$ is the conditional entropy of the classes. $H(C)$ is the entropy of the classes. n is the total number of samples, n_c and n_k the number of samples respectively belonging to class c and cluster k, and finally $n_{c,k}$ the number of samples from class c assigned to cluster k.

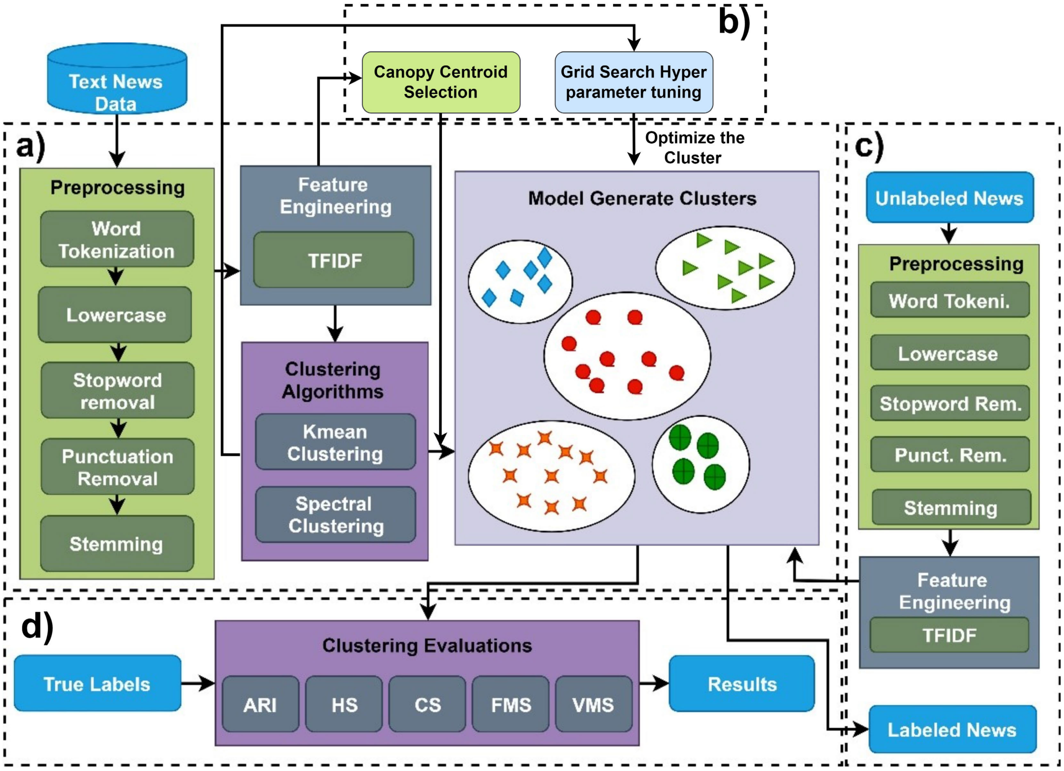

FIGURE 2. KMGC-search clustering based news text data classification architecture is presented. a) presents the training phase of the model after preprocessing and feature engineering, b) presents the canopy centroid selection and grid search hyper parameter tuning techniques to improve the clustering calculation. c) News categorization phase in which a trained model assigns the label to the unlabeled news. d) presents the evaluation phase in which evaluation parameters measure the similarity of clusters created by model.

The V-measure [59] is the harmonic mean between homogeneity score and completeness score just as precision and recall are calculated as harmonic mean to measure F-measure. The Fowlkes-mallows score and V-measure score are the weighted measures with the contributions of homogeneity score or completeness score. It is an entropy-based measure that explicitly measures how successfully the criteria of homogeneity score and completeness score have been satisfied.

$$v = \beta.\left(\frac{h * c}{h + c}\right) \tag{12}$$

where is 2, h is homogeneity and c is completeness To measure the similarity of two clusters of a set of points. The Fowlkes-Mallows index (FMI) [60] defined as the geometric mean between the precision and recall:

$$FMI = \frac{TP}{\sqrt{(TP + FP) * (TP + FN)}} \tag{13}$$

where TP is the number of True Positive, FP is the number of False Positive and FN is the number of False Negatives.

IV. METHODOLOGY

The main aim of this research is to categorized news articles in the world of the internet that facilitates the user to interact with informative, desired and knowledgeable. The news articles are one of the major sources of the data that grow on the internet on a daily or hourly basis. This research proposed the approach using the BBC news dataset and perform clustering by using proposed K-Means using Grid Search based on Canopy (KMGC-Search) clustering to categorize the news. The experiments are conducted with Kmeans, spectral and proposed KMGC-Search clustering techniques. First, apply the preprocessing on the dataset that consists of word tokenization, lower case conversion, stop word removal, punctuation removal and stemming methods. The preprocessing filter unnecessary data from the raw news article text. The proposed machine learning model K-means clustering and Spectral clustering need to train on the better terms of data to create efficient clusters and accurately perform predictions. There preprocessing eliminates unnecessary terms and data to increase the model performance. After preprocessing the most important part is to extract valuable features to enhance the model performance.

The TF-IDF is performed much better than other techniques because it extracts the feature from the filtered data based on weighted scores that transform the preprocessed data into vector space in the form of extracted features. These scores are based on the concept of the relevance of data and it helps to handle outliers that also increase the performance

TABLE 2. Impact of preprocessing on text data.

Sent. ID	Before Preprocessing	After Preprocessing
Sent. 1	tv future in the hands of viewers with home theatre systems plasma high-definition tvs and digital video recorders moving into the living room the way people watch tv will be radically different in five years time.	trend of freeing up multimedia so that people can watch what they want when they want.
Sent. 2	howard hits back at mongrel jibe michael howard has said a claim by peter hain that the tory leader is acting like an attack mongrel shows labour is rattled by the opposition.	howard hit back mongrel jibe michael howard said claim peter hain tory leader act like attack mongrel show labour rattle opposite
Sent. 3	worldcom boss left books alone former worldcom boss bernie ebbers who is accused of overseeing an $11bn (5.8bn) fraud never made accounting decisions a witness has told jurors.	worldcom boss left book alone former worldcom boss bernie ebber accuse oversee 11bn £58bn fraud never made account decision with told juror david
Sent. 4	henman hopes ended in dubai third seed tim henman slumped to a straight sets defeat in his rain-interrupted dubai open quarter-final against ivan ljubicic.	henman hope end dubai third seed tim henman slump straight set defeat raininterrupt dubai open quarterfinal ivan ljubic

of the proposed model. Then proposed KMGC-Search, K-means and spectral clustering are applied to conduct the experiments to categorize the news article into different clusters. These clustering models are based on the manually specified number of clusters that give the feasibility to analyze the performance of the kmeans and spectral clustering model in contrast with true labels. The canopy algorithm is used for the proposed approach for the selection of the group of centroids and hyper parameter tuning for the kmeans clustering to create better clusters based on canopy centroids. The analyses of the performance are evaluated by using several evaluation parameters like Adjusted Rand Index, homogeneity score, completeness score, Fowlkes-mallows score, and v measure that evaluates the clusters based on the true labels and cluster similarity measures as shown in Figure 2.

A. PREPROCESSING

The news article dataset is originated from BBC news that consists of 2225 document that is categorized in 5 different classes. The text news is in the raw form that needs to be preprocessed. The machine learning models can learn from this preprocessed data more efficiently. The real-world data is regularly incomplete, inconsistent, or missing in certain behaviors or trends, and is in all likelihood to incorporate many errors. Data preprocessing is an established approach to resolving such problems. In the world, the incompleteness of data is a general thing, lacking attribute values, errors, and outliers or containing only aggregate data. In the preprocessing, First, word tokenization has applied that acts as a breaking of a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements are known as tokens. These tokens can be individual words, phrases or even a whole sentence. The word tokenization is applying to split the documents or sentences into individual terms. That helpful in filtering non-important words and punctuation.

Second, in the case of sensitive system, capital case letter and lower-case letter consider as different terms because the conversion of capital case letters into lower case reduce the unique terms in the documents that increase the efficiency of the feature extraction process. Third, the process of changing statistics to something that the system can understand referred to as pre-processing. One of the primary forms of pre-processing is to filter useless data. In natural language processing, useless words (information), called stopwords. The stopwords are commonly used words in daily life (like ''the'', ''a'', ''an'', ''in''). The stop word removal reduces the sparsity of the data and increases the computational power of the model. Those words taking up precious processing time or affect the evaluation results. Fourth, the process of converting word into its root form known as stemming. A stemmer or stemming algorithm reduces the words ''chocolates'', ''chocolatey'', ''choco'' to root word, ''chocolate'' and ''retrieval'', ''retrieved'', ''retrieves'' reduce to the ''retrieve''. Table 2 shows the after the change of the preprocessing as compared to the data before preprocessing

B. FEATURE ENGINEERING

The TF-IDF feature extraction technique is applied after the preprocessing of the data that converts the text data into the data vectors as shown in Table 3. There are 10 features presents as a sample that is extracted from the four preprocessed sentences taken from the Table 2. The first column (Sent ID) shows the sentences id and the first row of the table shows the names of terms. The first term boss show 0 value in sent 1 because there is no boss term in sentence one but it shows a weighted value of 0.6666666 is sent 3 because this sentence contains a boss term as shown in Table 2 in after preprocessed (sent 3) sentence. Similarly, the whole TF-IDF matrix is created with the data vectors. These data vector facilitates efficient and clustering models in the computational process.

C. PROPOSED CLUSTERING MODELS

After the preprocessing and feature extraction, the data become ready as a feature vector for the further training process of the clustering model. The proposed study conduct experiments with proposed KMGC-Search, K-means and Spectral clustering models using the TF-IDF feature extraction technique. The Figure 3 illustrates the original dataset and clustered dataset. The original dataset figure shows the

TABLE 3. TF-IDF weighted matrix with a sample of 10 most frequent features.

Sent. ID	boss	dubai	henman	howard	opposite	oversee	people	they	want	worldcom
Sent. 1	0	0	0	0	0	0	0.333	0.666	0.666	0
Sent. 2	0	0	0	0.894	0.447	0	0	0	0	0
Sent. 3	0.666	0	0	0	0	0.333	0	0	0	0.666
Sent. 4	0	0.707	0.707	0	0	0	0	0	0	0

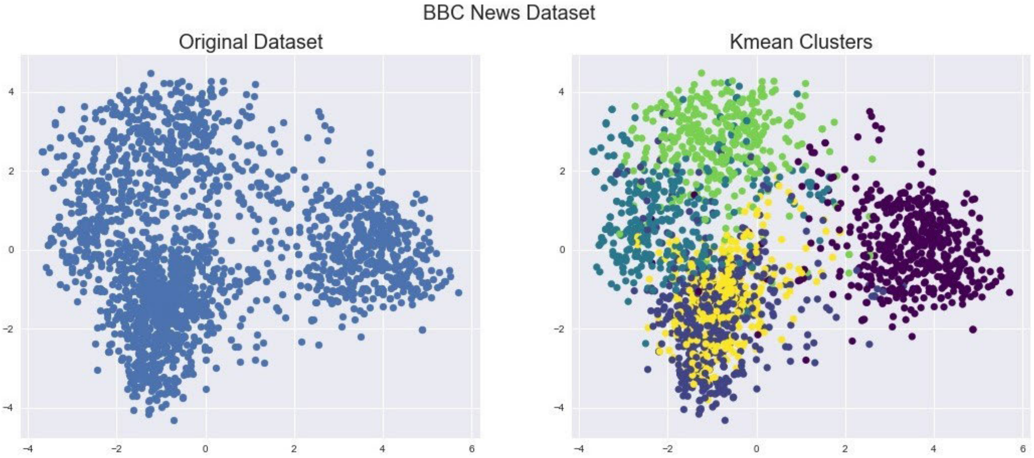

FIGURE 3. The difference of the dataset after and before proposed KMG-Search clustering and canopy based clustering technique.

data point that is created based on TF-IDF vectors. The train test split applied on this dataset to divide the data into two portions one is for training purpose and one for testing purpose.

The ratio is 80% of training data and 20% of testing data. The training data are given to the clustering algorithms for the training process. The K-means clustering is implemented to get the best results with the canopy centeroid selection and grid search hyper-parameter tuning.

One of the important hyper-parameter to specified the number of clusters is "n_cluster = 5" because the dataset is divided into five classes. For the simple Kmeans clustering model implementation on the dataset, number of cluster 5. The K-means++ parameter is used for the initialization of the center of the cluster that speeds up the learning process of the model. Another hyper-parameter are "n_init is 10", "random state is 40" and "max_iter is 300" which means K-means algorithms perform 300 iterations in each pass with different centroids to extract the best number of clusters as an output. In last "n_jobs with -1" is used to get the parallel processing and speed up the model training process.

Similarly, another experiment conducted with the spectral clustering algorithm with hyperparameters of the number of specified clusters n_clusters is 5, random_state is 40, n_init is 10, gamma is 1.0, affinity is 'rbf', n_neighbors is 10, assign_labels is 'kmeans' and n_jobs is -1. These clustering algorithms create 5 different clusters of data and evaluate the proposed approach by contrasting the actual labels with the number of clusters that how much accurately clusters data are similar. The results of the experimental phase are discussed in the upcoming section Results and Discussions that shows the kmeans clustering shows better results then the spectral clustering.

The proposed KMGC-Search clustering technique is proposed in this paper based on main three components Kmeans clustering, canopy centeroid selection and grid search based hyper parameter tuning methods. The canopy is used to initialize the number of centeroid groups for the selection of centers by kmeans clustering algorithms then the next step is applied as a grid search hyper parameter tuning approach to optimize the computational processing of the kmeans to create better number of clusters. The comparative results of the proposed Kmeans, Spectral and KMGC-Search clustering algorithms are discussed in details in next section V.

V. RESULTS AND DISCUSSIONS

The results of the proposed study discussed in this section in which several numbers of experiments are performed to categorize the news articles in different clusters by using K-means, Spectral and KMGC-Search clustering techniques. To evaluate those experimental results following evaluation parameters are used such as ARI, HS, CS, VM and FM scores.

There are two types of clustering models are available, first in which we define the number of clusters and the second select number of clusters on the bases of its clustering terminology. The implementation of the proposed KMGC-Search,

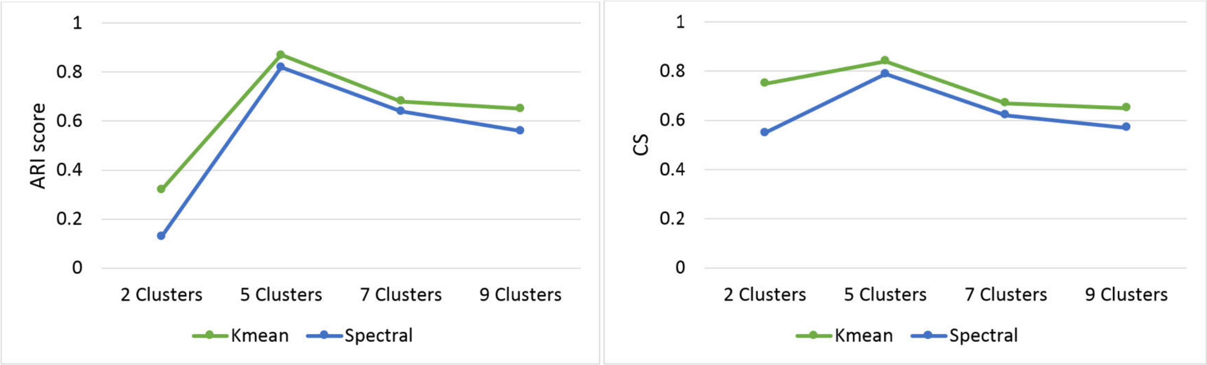

FIGURE 4. ARI and CS score using K-means and spectral clustering models.

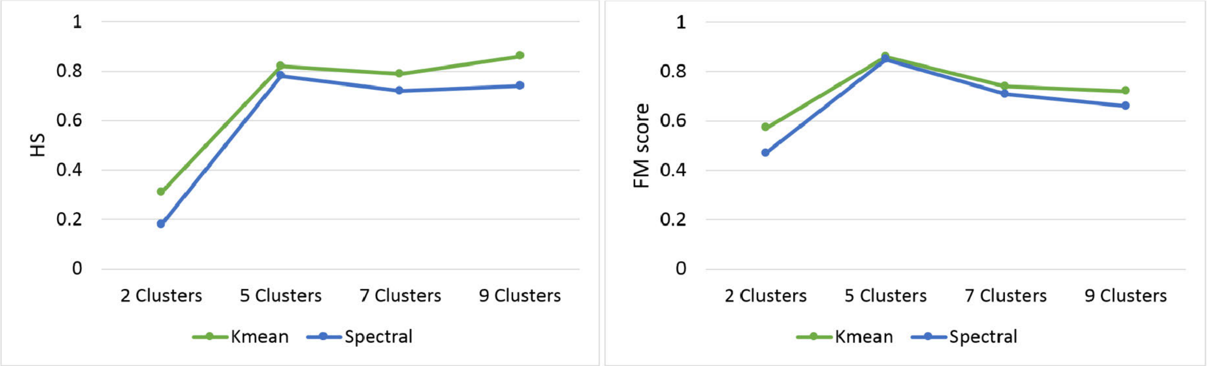

FIGURE 5. HS and FM score using K-means and spectral clustering models.

TABLE 4. K-means clustering results.

No. of clusters	ARI Score	HS Score	CS Score	VM Score	FM Score
2	0.32	0.31	0.75	0.44	0.57
5	0.87	0.84	0.84	0.84	0.89
7	0.68	0.79	0.67	0.73	0.74
9	0.65	0.86	0.65	0.74	0.72

TABLE 5. Spectral clustering results.

No. of clusters	ARI Score	HS Score	CS Score	VM Score	FM Score
2	0.13	0.18	0.55	0.27	0.47
5	0.82	0.78	0.79	0.78	0.85
7	0.64	0.72	0.62	0.67	0.71
9	0.56	0.74	0.57	0.64	0.66

TABLE 6. Proposed KMGC-Search clustering technique results are evaluated based on no. of clusters specified by canopy.

No. of clusters	ARI Score	HS Score	CS Score	VM Score	FM Score
2	0.46	0.44	0.77	0.45	0.58
5	0.96	0.92	0.94	0.94	0.95
11	0.78	0.77	0.79	0.75	0.78
17	0.74	0.75	0.74	0.76	0.71
19	0.68	0.66	0.71	0.70	0.69

TABLE 7. Comparative analyse of the clustering proposed KMGC-Search with other clustering algorithms using 5 clusters.

Clustering Models	ARI Score	HS Score	CS Score	VM Score	FM Score
Kmeans	0.87	0.84	0.84	0.84	0.89
Spectral	0.82	0.78	0.79	0.78	0.85
KMGC-Search	**0.96**	**0.92**	**0.94**	**0.94**	**0.95**

K-means and spectral clustering belongs to the second type in which the number of clusters is needed to be specified first. The experiments performed with the different number of clusters with both K-means and Spectral clustering model. The canopy and Grid search techniques are used with the proposed KMGC-Search clustering to specified the number of clusters and for better performance. The first experimental results of K-means and Spectral clustering are evaluated using the ARI score as shown in Figure 4. According to the ARI score, results with K-means clustering and spectral clustering create a set of 2, 5, 7, and 9 clusters. The ARI score measured the similarity between clusters as discussed in section III-F.

TABLE 8. Comparative literature analyse with proposed approach.

Literature	Year	Technology	Dataset	Results
[65]	2019	Density-Canopy-Kmeans Algorithm and MDS Embedding	gene expression data of cells based news	The highest ARI score is achieved with kmeans is 95.0%, F-measure is 96.4%
[66]	2019	JS (Jensen-Shannon) divergence, LDA (latent Dirichlet allocation) model and Gibbs Sampling method is used with K-means	Selected five news data sets are: D1, D2, D3, D4,and D5 based on keywords tourism, car, education, culture, military, finance, sports, and IT	The highest results are obtained by D5 data, Fmeasure is 82.3%, recall 81.3%, and precision 83.3%
[68]	2020	Bag of near synonyms (BoNS), Hash Set Frequency-Inverse Document Frequency (hSF-IDF) used to cluster the web news	a real-world dataset by harvesting news documents from web portals	Highest f measure is89.71% with TFIDF
[69]	2020	Hybrid Personalised NEws Recommendation (HYPNER) based on Collaborative Filtering (CF) and Content-based technique	Real time dataset is crawled using Twitter information streams includes 1,619 users have 20 tweets, 2,316,204 news records dataset	HYPNER achieved 81.56% improvement in F1 -score and 5.33% in diversity
[61]	2021	Canopy with k-means clustering algorithm for big data analytics was used	Dental healthcare insurance news dataset based on Hadoop Distributed File System (HDFS)	The execution time of the kmeans with canopy is reduces from 68.567 to 60.75 seconds
[67]	2021	Semantic Clustering TextRank (SCTR) and Bidirectional Encoder Representation from Transformers (BERT) model are used	dataset news based keywords are extracted from the web links	The highest precision is gain 71%, Fmeasure 75% and recall 92%
Proposed KMGC-Search	2022	News text dataset consist of 5 different categories of news	Kmeans clustering based on Canopy centeroid selection and grid search based hyper parameter tuning based KMGC-Search approach	The highest score is achieved with 0.96% of ARI, 0.92% of HS, 0.94% of CS, 0.94% of VM and 0.95% of FM score.

The highest ARI score is achieved as 87% with 5 clusters using the K-means clustering algorithm and 82% is achieved using spectral clustering. The completeness is the measure of all members of a given class assigned to the same cluster. It describes the clusters data that all the class data assigned to every single cluster or multiple clusters. The highest CS score of 84% is gain with five clusters using TF-IDF features with the K-means clustering model that presents in Figure 4 and by using spectral clustering 79% CS score is gained. The Homogeneity score illustrates that each cluster contains only members of a single class. It describes that the how perfectly clustering model creates clusters of data by evaluating its data that it belongs to single class data or not.

The highest HS score of 86% is gain with 9 clusters using K-means clustering and 78% score gain with 5 clusters using spectral clustering as shown in Figure 5. The FM score measures the similarity in between clusters that shows how much data of the same cluster are similar to each other and how much dissimilar to the data of the other clusters. The highest FM score of 86% is gain with 5 clusters of K-means algorithm and 85% FM score gain with 5 clusters using spectra clustering shown in Figure 5. The VM score is the weighted score of the HS and CS score same as the harmonic mean of HS and CS that describes the similarity of data in a cluster and by summing up this score of each cluster calculate weighted average score all of the clusters. The highest VM score of 82% achieved with 5 clusters that are created by the K-means clustering model using the TF-IDF feature extraction technique as shown in Figure 6. The clustering has been done by using unsupervised machine learning models in the proposed study. In this section, the results of K-means clustering and spectral clustering models are comparatively discussed with several performance evaluations such as ARI, HS, CS, VM and FM score. The results of K-means clustering shown in Figure 4. The clusters of 2, 5, 7 and 9 are created by using K-means clustering with TF-IDF features. When create 5 clusters then K-means show the highest results.

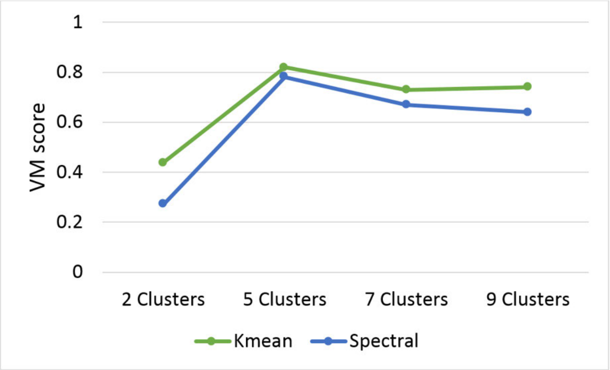

FIGURE 6. VM score of K-means and spectral clustering models.

Table 6 presents the spectral clustering results. The highest results achieved in 5 clusters that are 87% ARI, 84% of HS, 84% of VM, and 89% of FM score. In the comparison of K-means and spectral clustering, both show the highest results with 5 number of clusters using the TF-IDF feature. However, in both models the K-means clustering outperforms then spectral clustering because of its similarity calculation of the news articles based on its centroids that is 5 and K-means, perform several numbers of iteration until the entire news article categorized to its similar cluster.

Table 7 presents the comparative results of used clustering models such as Kmeans, Spectral and KMGC-Search. The impact of the clustering algorithms are displayed and highest results with KMGC-Search model is 0.96 of ARI, 0.92 of HS, 0.94 of CS, 0.94 of VM, and 0.95 of FM score that outperforms the kmeans and spectral clustering. The comparative analyse of the proposed approach with previously used technologies in literature shown in Table 8.

VI. CONCLUSION

The categorization of the news articles based on the clustering algorithm with the TF-IDF feature extraction technique and hyperparameter tuning makes the user get the desirable, valuable, and informative news. Now, the user does not feel the hassle to spend a lot of time in search of specific news. In this paper, the categorization of the news articles is proposed based on K-means, Spectral and proposed KMGC-Search clustering. To enhance the computational power and speed of the training process, first, clean up the unnecessary data from the news article text dataset using preprocessing. After that, the TF-IDF feature extraction method is applied to get more discernment features. Then these feature vector pass to the K-means clustering as input features with the selective seeding method to create clusters of news and spectral clustering by defining the specific number of clusters. In this experiment, the news are categorize and classify efficiently and to get better results with 2, 5, 7, and 9 number of clusters. The K-means and spectral clustering both show the highest results with 5 clusters that show the classification process gives better results with the five categories. K-means shows better results than spectral clustering and achieved ARI score of 82%, HS score of 78%, CS score of 79%, VM score of 78%, and FM score of 85%. The canopy centeroid selection and grid search hyperparameter optimization is used to proposed the KMGC-Search clustering model by achieving the highest results of 0.96% of ARI, 0.92% of HS, 0.94% of CS, 0.94% of VM and 0.95% of FM score. The deep learning models can be used as a future directions.

REFERENCES

[1] J. Petit, C. Li, and K. Ali, "Fewer people, more flames: How pre-existing beliefs and volume of negative comments impact online news readers' verbal aggression," *Telematics Inform.*, vol. 56, pp. 101471–101509, Jan. 2020.

[2] J. Strömbäck, Y. Tsfati, H. Boomgaarden, A. Damstra, E. Lindgren, R. Vliegenthart, and T. Lindholm, "News media trust and its impact on media use: Toward a framework for future research," *Ann. Int. Commun. Assoc.*, vol. 44, no. 2, pp. 139–156, Apr. 2020.

[3] D. Zhu and S. Lee, "Autonomous readers: The impact of news customisation on audiences' psychological and behavioural outcomes," *Commun. Res. Pract.*, vol. 6, no. 2, pp. 125–142, Apr. 2020.

[4] J. P. Ferré, "A short history of media ethics in the United States," in *The Handbook of Mass Media Ethics*. Evanston, IL, USA: Routledge, 2020, pp. 16–27.

[5] M. Shahroz, M. Mushtaq, M. Ahmad, S. Ullah, A. Mehmood, and G. Choi, "IoT-based smart shopping cart using radio frequency identification," *IEEE Access*, vol. 8, pp. 68426–68438, 2020.

[6] P. Jiao, A. Veiga, and A. Walther, "Social media, news media and the stock market," *J. Econ. Behav. Org.*, vol. 176, pp. 63–90, Aug. 2020.

[7] J. L. Nelson, "The persistence of the popular in mobile news consumption," *Digit. J.*, vol. 8, no. 1, pp. 87–102, Jan. 2020.

[8] S. C. Lewis, "Lack of trust in the news media, institutional weakness, and relational journalism as a potential way forward," *Journalism*, vol. 21, no. 3, pp. 345–348, Mar. 2020.

[9] M. Anderson and J. Jiang, "Teens, social media & technology 2018," *Pew Res. Center*, vol. 31, pp. 1673–1689, May 2018. [Online]. Available: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/

[10] N. A. Swidan, S. K. Guirguis, and O. G. Abood, "Text document clustering using hashing deep learning method," *Int. J. Soft Comput.*, vol. 14, no. 2, pp. 44–52, Apr. 2020.

[11] C. Chang, S.-Y.-J. Chao, and B. Chiang, "INTERNET RESOURCES: East Asian studies: Sites to help meet the growing demand for information," *College Res. Libraries News*, vol. 59, no. 7, pp. 514–521, Feb. 2020.

[12] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, pp. 102025–102051, 2019.

[13] A. Talmor and T. M. Brown, "Automatic clustering by topic and prioritizing online feed items," U.S. Patent 10 592 841, Mar. 17, 2020.

[14] I. Park, H. Shim, J. H. Kim, C. Lee, and D. Lee, "The effects of popularity metrics in news comments on the formation of public opinion: Evidence from an internet portal site," *Social Sci. J.*, pp. 1–16, Jun. 2020.

[15] M. S. Rathore, P. Saurabh, R. Prasad, and P. Mewada, "Text classification with K-nearest neighbors algorithm using gain ratio," in *Progress in Computing, Analytics and Networking*. Singapore: Springer, 2020, pp. 23–31.

[16] T. Jenson and A. S. Girsang, "Performance of news clustering using ant colony optimization," *J. Phys., Conf. Ser.*, vol. 1566, no. 1, pp. 12101–12108, 2020.

[17] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA, USA: MIT Press, 2020.

[18] M. E. Faisal and M. Zamzami, "Comparative analysis of inter-centroid K-means performance using Euclidean distance, Canberra distance and Manhattan distance," *J. Phys., Conf. Ser.*, vol. 1566, no. 1, pp. 12112–12119, 2020.

[19] E. L. Lydia, P. Govindaswamy, S. Lakshmanaprabu, and D. Ramya, "Document clustering based on text mining K-means algorithm using Euclidean distance similarity," *J. Adv. Res. Dyn. Control Syst.*, vol. 10, no. 2, pp. 208–214, 2018.

[20] P. Haloi, P. Gadde, and M. K. Bhuyan, "News video indexing and story unit segmentation using text cue," in *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, Mar. 2019, pp. 501–507.

[21] Y. Bar-Sinai, S. Hoyer, J. Hickey, and M. P. Brenner, "Learning data-driven discretizations for partial differential equations," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 31, pp. 15344–15349, 2019.

[22] M. De Grandis, G. Pasi, and M. Viviani, "Fake news detection in microblogging through quantifier-guided aggregation," in *Proc. Int. Conf. Modeling Decis. Artif. Intell.*, 2019, pp. 64–76.

[23] M. Pourvali, S. Orlando, and H. Omidvarborna, "Topic models and fusion methods: A union to improve text clustering and cluster labeling," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 5, no. 4, pp. 28–34, 2019.

[24] C. Dimoulas and A. Veglis, "Factors and models contributing to the optimization of search engine results credibility and application on news content: The cross-credibility engine optimization (CCEO) model," in *Proc. 23rd Pan-Hellenic Conf. Informat.*, Nov. 2019, pp. 144–147.

[25] I. C. Drivas, D. P. Sakas, G. A. Giannakopoulos, and D. Kyriaki-Manessi, "Big data analytics for search engine optimization," *Big Data Cogn. Comput.*, vol. 4, no. 2, pp. 5–27, 2020.

[26] U. Buatoom, W. Kongprawechnon, and T. Theeramunkong, "Document clustering using K-means with term weighting as similarity-based constraints," *Symmetry*, vol. 12, no. 6, p. 967, Jun. 2020.

[27] X. Ao, X. Yu, D. Liu, and H. Tian, "News keywords extraction algorithm based on TextRank and classified TF-IDF," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2020, pp. 1364–1369.

[28] Z. Mahmood, "Deep sentiments in Roman Urdu text using recurrent convolutional neural network model," *Inf. Process. Manage.*, vol. 57, no. 4, pp. 102233–102247, 2020.

[29] J. Xu and G. Durrett, "Neural extractive text summarization with syntactic compression," 2019, *arXiv:1902.00863*.

[30] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[31] M. Afreen and S. Badugu, "Document clustering using different unsupervised learning approaches: A survey," in *Advances in Decision Sciences, Image Processing, Security and Computer Vision*. Cham, Switzerland: Springer, 2020, pp. 619–629.

[32] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019.

[33] J. Yin, D. Chao, Z. Liu, W. Zhang, X. Yu, and J. Wang, "Model-based clustering of short text streams," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2634–2642.

[34] A. Tommasel and D. Godoy, "A social-aware online short-text feature selection technique for social media," *Inf. Fusion*, vol. 40, pp. 1–17, Mar. 2018.

[35] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1495–1545, 2019.

[36] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[37] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10394–10403.

[38] L. Tay, S. E. Woo, L. Hickman, and R. M. Saef, "Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining," *Eur. J. Pers.*, vol. 34, no. 5, pp. 826–844, Sep. 2020.

[39] G. Sun, Y. Cong, Q. Wang, J. Li, and Y. Fu, "Lifelong spectral clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5867–5874.

[40] S. Lv, L. Huang, L. Zang, W. Zhou, J. Han, and S. Hu, "Yet another approach to understanding news event evolution," *World Wide Web*, vol. 23, no. 4, pp. 2449–2470, Jul. 2020.

[41] S. Tasoulis, N. G. Pavlidis, and T. Roos, "Nonlinear dimensionality reduction for clustering," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107508.

[42] Y. F. Huang and P. H. Chen, "Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms," *Expert Syst. Appl.*, vol. 159, pp. 113584–113620, Nov. 2020.

[43] J. Sheela and B. Janet, "An abstractive summary generation system for customer reviews and news article using deep learning," *J. Ambient Intell. Hum. Comput.*, vol. 12, pp. 7363–7373, Aug. 2020.

[44] S. Im, M. M. Qaem, B. Moseley, X. Sun, and R. Zhou, "Fast noise removal for k-means clustering," 2020, *arXiv:2003.02433*.

[45] M. Tang, Y. Kaymaz, B. L. Logeman, S. Eichhorn, Z. S. Liang, C. Dulac, and T. B. Sackton, "Evaluating single-cell cluster stability using the Jaccard similarity index," *Bioinformatics*, vol. 37, no. 15, pp. 2212–2214, 2021.

[46] D. Veeraiah and J. N. Rao, "Use of clustering sentiments for opinion mining: An experimental analysis," in *Information and Communication Technology for Sustainable Development*. Singapore: Springer, 2020, pp. 625–632.

[47] K. Rangasamy, P. Anjana, S. Bavatarani, and D. Kumar, "A study on human behavior based color psychology using K-means clustering," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Feb. 2020, pp. 608–612.

[48] A. Kabra and S. Shrawne, "Location-wise news headlines classification and sentiment analysis: A deep learning approach," in *Proc. Int. Conf. Intell. Comput. Smart Commun.*, 2020, pp. 383–391.

[49] A. S. Ramkumar and B. Poorna, "Text document clustering using dimension reduction technique," *Int. J. Appl. Eng. Res.*, vol. 11, no. 7, pp. 4770–4774, 2016.

[50] D. Liparas, Y. H. Kerner, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris, "News articles classification using random forests and weighted multimodal features," in *Proc. Inf. Retr. Facility Conf.*, 2014, pp. 63–64.

[51] I. Blokh and V. Alexandrov, "News clustering based on similarity analysis," *Proc. Comput. Sci.*, vol. 122, pp. 715–719, Jan. 2017.

[52] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," 2017, *arXiv:1707.02919*.

[53] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 377–384.

[54] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *Telkomnika*, vol. 18, no. 2, pp. 799–806, 2020.

[55] P. Saigal and V. Khanna, "Multi-category news classification using support vector machine based classifiers," *Social Netw. Appl. Sci.*, vol. 2, no. 3, pp. 1–12, Mar. 2020.

[56] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[57] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Inf. Process. Manage.*, vol. 57, no. 2, pp. 102034–102084, 2020.

[58] R.-G. Radu, I.-M. Radulescu, C.-O. Truica, E.-S. Apostol, and M. Mocanu, "Clustering documents using the document to vector model for dimensionality reduction," in *Proc. IEEE Int. Conf. Autom., Quality Test., Robot. (AQTR)*, May 2020, pp. 1–6.

[59] O. Sánchez and G. Sierra, "Joint sentiment topic model for objective text clustering," *J. Intell. Fuzzy Syst.*, vol. 36, no. 4, pp. 3119–3128, Apr. 2019.

[60] D. N. Campo, G. Stegmayer, and D. H. Milone, "A new index for clustering validation with overlapped clusters," *Expert Syst. Appl.*, vol. 64, pp. 549–556, Dec. 2016.

[61] N. S. Sagheer and S. A. Yousif, "Canopy with k-means clustering algorithm for big data analytics," *AIP Conf. Proc.*, vol. 2334, no. 1, 2021, Art. no. 070006.

[62] G. Zhang, C. Zhang, and H. Zhang, "Improved K-means algorithm based on density Canopy," *Knowl.-Based Syst.*, vol. 145, pp. 289–297, Apr. 2018.

[63] G. Li, W. Wang, W. Zhang, Z. Wang, H. Tu, and W. You, "Grid search based multi-population particle swarm optimization algorithm for multimodal multi-objective optimization," *Swarm Evol. Comput.*, vol. 62, Apr. 2021, Art. no. 100843.

[64] M. Bhagat, D. Kumar, I. Haque, H. S. Munda, and R. Bhagat, "Plant leaf disease classification using grid search based SVM," in *Proc. 2nd Int. Conf. Data, Eng. Appl. (IDEA)*, Feb. 2020, pp. 1–6.

[65] M. Li, H. Wang, H. Long, J. Xiang, B. Wang, J. Xu, and J. Yang, "Community detection and visualization in complex network by the density-canopy-Kmeans algorithm and MDS embedding," *IEEE Access*, vol. 7, pp. 120616–120625, 2019, doi: 10.1109/ACCESS.2019.2936248.

[66] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, "Research on topic detection and tracking for online news texts," *IEEE Access*, vol. 7, pp. 58407–58418, 2019, doi: 10.1109/ACCESS.2019.2914097.

[67] A. Xiong, D. Liu, H. Tian, Z. Liu, P. Yu, and M. Kadoch, "News keyword extraction algorithm based on semantic clustering and word graph model," *Tsinghua Sci. Technol.*, vol. 26, no. 6, pp. 886–893, Dec. 2021, doi: 10.26599/TST.2020.9010051.

[68] Z. Zhang, L. Chen, F. Yin, X. Zhang, and L. Guo, "Improving online clustering of Chinese technology web news with bag-of-near-synonyms," *IEEE Access*, vol. 8, pp. 94245–94257, 2020, doi: 10.1109/ACCESS.2020.2995516.

[69] A. Darvishy, H. Ibrahim, F. Sidi, and A. Mustapha, "HYPNER: A hybrid approach for personalized news recommendation," *IEEE Access*, vol. 8, pp. 46877–46894, 2020, doi: 10.1109/ACCESS.2020.2978505.

MOBEEN SHAHROZ received the M.C.S. degree from the Department of Computer Science, Khawaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2018, and the M.S. degree in computer science, in 2020. He is currently pursuing the Ph.D. degree in computer science with the Department of Artificial Intelligence, The Islamia University of Bahawalpur. His current research interests include the Internet of Things (IoT), artificial intelligence, data mining, natural language processing, machine learning, deep learning, and image classification.

MUHAMMAD FAHEEM MUSHTAQ received the B.S. degree in IT and the M.S. degree in CS from The Islamia University of Bahawalpur, Punjab, Pakistan, in 2011 and 2013, respectively, and the Ph.D. degree from the Department of Information Security, Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia (UTHM), Malaysia, in 2018.

He has made several contributions through research publications and book chapters toward Information Security. He received Microsoft certifications of internet security and acceleration (ISA) server, Microsoft certified professional (MCP), Microsoft certified technology professional (MCTS), in 2010. He is currently working as the Head of the Department of Artificial Intelligence, The Islamia University Bahawalpur. Previously, he was working as the Head/an Assistant Professor with the Department of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan. He was worked as a Research Assistant during his Ph.D. degree, from March 2016 to August 2018. His research interests include information security, artificial intelligence, and cognitive systems and applications.

RIZWAN MAJEED received the B.S. degree in IT from the University of Punjab, in 2008, and the M.Phil. degree in computer sciences from NCBA, Pakistan, in 2018. He is currently pursuing the Ph.D. degree with UTHM, Malaysia. He joined The Islamia University of Bahawalpur as the Director of IT. His current research interests include IoT-based smart systems, artificial intelligence, data mining, machine learning, image processing, and deep learning.

ALI SAMAD received the Bachelor of Science degree from the University of Punjab, in 2004, the master's degree in computer sciences from NCBA University, Pakistan, in 2008, and the M.S. degree, in 2013. He is currently pursuing the Ph.D. degree with UTHM, Johar, Malaysia. He joined the Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan. His current research interests include IoT-based smart systems, artificial intelligence, data mining, machine learning, image processing, and deep learning.

ZAIGHAM MUSHTAQ received the M.S. degree in computer science from the COMSATS Institute of Information Technology, Lahore, in 2010, where he also received the Ph.D. degree in computer science. He is currently working as an Assistant Professor with the Department of Computer Science, The Islamia university of Bahawalpur, Bahawalpur. He was involved in software process improvement and semantic web-based SQL statements. He is involved in design recovery of multilingual applications through recognition of J2EE Pattern. His research interests include source code analysis especially cross language dependence analysis, program comprehension, and source code documentation.

UROOJ AKRAM receive the B.S. degree in computer science from The Islamia University of Bahawalpur, Punjab, Pakistan, in 2013, and the M.S. degree in information technology from the Faculty of Computer Science and Information Technology, UTHM, Malaysia, in 2018. She is currently pursuing the Ph.D. degree in computer science with the Department of Artificial Intelligence, The Islamia University of Bahawalpur. She is currently working as an Associate Lecturer with the Department of Artificial Intelligence, The Islamia University of Bahawalpur. She has one year of experience as a Lecturer with the Department of Information Technology, KFUEIT, Rahim Yar Khan. Her research work is published in well-reputed high-impact journals. Her research interests include machine learning, deep learning, natural language processing, and information security.