IEEE Access
Multidisciplinary Rapid Review Open Access Journal

Received August 7, 2019, accepted September 26, 2019, date of publication October 1, 2019, date of current version October 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944973

Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering

JUNAID RASHID[1], SYED MUHAMMAD ADNAN SHAH[1], AUN IRTAZA[1], TOQEER MAHMOOD[1], MUHAMMAD WASIF NISAR[2], MUHAMMAD SHAFIQ[3], AND AKBER GARDEZI[4]

[1]Department of Computer Science, University of Engineering and Technology Taxila, Punjab 47050, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah Cantt 47040, Pakistan
[3]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea
[4]Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan

Corresponding author: Muhammad Shafiq (shafiq.pu@gmail.com)

This work was supported by the Brain Korea 21 Plus Program funded by the National Research Foundation of Korea (NRF) under Grant 22A20130012814.

ABSTRACT Text data plays an imperative role in the biomedical domain. As patient's data comprises of a huge amount of text documents in a non-standardized format. In order to obtain the relevant data, the text documents pose a lot of challenging issues for data processing. Topic modeling is one of the popular techniques for information retrieval based on themes from the biomedical documents. In topic modeling discovering the precise topics from the biomedical documents is a challenging task. Furthermore, in biomedical text documents, the redundancy puts a negative impact on the quality of text mining as well. Therefore, the rapid growth of unstructured documents entails machine learning techniques for topic modeling capable of discovering precise topics. In this paper, we proposed a topic modeling technique for text mining through hybrid inverse document frequency and machine learning fuzzy k-means clustering algorithm. The proposed technique ameliorates the redundancy issue and discovers precise topics from the biomedical text documents. The proposed technique generates local and global term frequencies through the bag-of-words (BOW) model. The global term weighting is calculated through the proposed hybrid inverse documents frequency and Local term weighting is computed with term frequency. The robust principal component analysis is used to remove the negative impact of higher dimensionality on the global term weights. Afterward, the classification and clustering for text mining are performed with a probability of topics in the documents. The classification is performed through discriminant analysis classifier whereas the clustering is done through the k-means clustering. The performance of clustering is evaluated with Calinsiki-Har-abasz (CH) index internal validation method. The proposed toping modeling technique is evaluated on six standard datasets namely Ohsumed, MuchMore Springer Corpus, GENIA corpus, Bioxtext, tweets and WSJ redundant corpus for experimentation. The proposed topic modeling technique exhibits high performance on classification and clustering in text mining compared to baseline topic models like FLSA, LDA, and LSA. Moreover, the execution time of the proposed topic modeling technique remains stable for different numbers of topics.

INDEX TERMS Topic modeling, biomedical, text mining, bag-of-words, classification, clustering, LDA, LSA.

I. INTRODUCTION

The tremendous amount of biomedical text documents is a valuable sourced of information in the biomedical field. Biomedical documents are categorized by the extensive amount of disorganized and infrequent information during a vast variety of forms like medical documents, scientific papers, electronic health records, a case summary of reports and so forth. In biomedical domain numbers of papers and

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

articles, publications are rising rapidly on the web due to the expanding of online publishing. The overall various articles indexed in MEDLINE increased to twenty-three million [1] and the number of citations attained eight hundred six thousand. The clinical text [2] like patient medical histories, findings notes during examinations from electronic health records and pathology reports of patients are needs to analyze for discovering the hidden information. Therefore, extraction of knowledgeable information in the form of themes from this immense collection of documents is a challenging and time-consuming task. Topic modeling techniques help in the extraction of unknown topics from a huge collection of documents [3], available articles and discover the topic distributions for every document. Topic models discover the topics from documents which are represented by the distribution of words. Topic modeling is a concept in which documents are the mixture of topics that are converted into probability over words distribution. Biomedical documents contain high dimension unstructured nature, so there are various methods exist for handling the biomedical text documents [4] such as FLSA, LDA, and LSA. Latent Dirichlet Allocation(LDA) finds the probabilities, which predict a posterior distribution of various words and topics from the input collection of text corpus [5]. LDA extract topics distribution by using Gibbs sampling which is an iterative method. Furthermore, this method selects some parameters like numbers of topics, iterations and Dirichlet priors. Therefore, the LDA topic model requires extensive empirical analysis with several configurations for finding the optimal configuration. The latent semantic analysis (LSA) method extracts topics and shows the semantic meaning of words with statistical computation on a huge collection of documents [6]. LSA discover the latent classes while minimizing the dimensionality for the vector space model [7]. Moreover, LSA has no powerful statistical approach and cause from mathematical complexity issue [8]. The proposed topic model is a part of the probabilistic modeling in which words are distributed over topics and topics are distributed over documents. We used the generative process that defined the joint probability distribution for observed and hidden random variables. The data analysis is performed by the joint distribution to compute the conditional distribution of the hidden variables given the observed variables. We compared the proposed topic model with probabilistic topic models like LDA, LSA, and PLSA. However, other topic models like STC and GSTC depend on the sparseness of topics in documents. Topic modeling is widely used for text mining to reveal the hidden structures from biomedical documents. Topic modeling categorized the documents into topics and represents these topics into the distribution of words. Topic modeling extracts the needed information very effectively from biomedical text documents. Topic modeling is an efficient technique for biomedical text mining but needs some improvement because biomedical text documents are words redundant [9] and redundancy is a negative impact on topic modeling and text mining [10]. The biomedical text documents consist of hundreds to thousands of medical features which cause the high-dimensionality problem [11]. Dealing with the high-dimensionality data is a challenging task. Handling very high dimensional data is difficult because it introduces many parameters into the model and makes the model much more complicated. Also, higher-order dimensional data is likely related to noise and sparsity problems.

A. TOPIC MODELING FOR BIOMEDICAL DOCUMENTS

Topic modeling techniques are utilized for the summarization of a large collection of text documents. Probabilistic topic modeling techniques are used to identify the core topics from the biomedical text collection of documents. In addition, the topic modeling techniques, are used in various tasks like computational linguistics, overview of source programs documents [12], summary opinion of the product review [13], description of the topic revolution [14], aspect discovery of documents analysis [15], analyzed of Twitter texts messages [16], and sentiments analytics [17]. The probabilistic of topic modeling caught the thought of researchers in the biomedical discipline. The most striking collections of biomedical texts endure from high dimensional issue and methods of topic modeling are effective for managing large-scale document collections. Thus, topic modeling can yield promising results in the extraction of biological and biomedical texts [18]. For example, in [19] discussed a probabilistic topic modeling technique to discover protein with protein interactions from the biologic research. In this method, the association among various approaches and correlated words are probabilistically modeled to obtain the discovery approaches. In [20] used the latent Dirichlet allocation technique to determine clinically appropriate topics and structured medical text finding reports. In [21] utilized the latent Dirichlet allocation technique to manipulate statistical investigation on bioinformatics from the huge text compilation of PubMed Central papers. In [22] topic modeling used to show clinical reports in a concise manner and these reports are processed effectively. In another study [23] topic modeling applies to drug labeling, which is an intensive human task with numerous unstructured meaningful explanations. In this way, the challenges of manual annotations are reduced. In [24] LDA used for discovering the relationship from protein-protein. One of its characteristics for candidate gene-drug pairs is ranking which is thematic distance derived from the LDA. Similarly, in [25] introduced a technique based on topic modeling to discover annotations from gene sets. In [26] used topic modeling based on the latent Dirichlet allocation reveal the interrelates between the cell endpoints. Experimental results analyzed that LDA can significantly improve the perspective of systems biology. The probabilistic topic modeling of subjects is used to identify drug repositioning strategies [27]. In [28] subject models are used to examine 17,723 abstracts of PubMed articles related to substances and depressive disorder in adolescents. In [29] introduced a topic modeling based system for biological mining of literature. Topic modeling used for the identification of unreliable nutritional supplemental from documents [30]. In a study [31], the hospital admission process in a precise way is represented using the probabilistic topic modeling. In the biomedical field, RedLDA topic modeling technique is utilized for identification of redundancy in the patient records [32]. Latent semantic analysis (LSA) utilized to instantly machine grading for clinical case summaries [33]. In [34] LSA discovers the clinical reports from the psychiatric narrative. Its produce the semantic space from psychiatric terms. LSA also used for discovering the semantic concepts and ontology domain which construct the model for speech act for spoken utterance [35]. In clinical filed LSA also gives better results in topic labeling and segmentation [36]. In [37] fuzzy clustering is used for diagnosis of medical imaging.

B. BIOMEDICAL APPLICATION OF FUZZY CLUSTERING

Clustering helps to achieve the objectives of biomedical research to retrieve knowledge from an immense amount of data that is further applied for treatment. In biomedical research clustering applications are pervasive [38]. Fuzzy clustering combines features objects with their relationship which relied on co-occurrences information [39]. Fuzzy clustering is utilized for medical diagnosis [40], medical image segmentation [41], biomedical signal classification [42] and diabetic neuropathy [43]. Fuzzy clustering also enhances decision making radiation therapy [44]. Fuzzy clustering also used for microarray data analysis [45]and breast elastography [46].In [47] proposed an FLSA technique that discovers the topics from medical documents using fuzzy c-means clustering. Fuzzy clustering has many applications for the biomedical domain mainly in image processing filed but very less consider for topic modeling. Therefore, in this research, we proposed a topic modeling technique for biomedical documents through fuzzy perspective. To ameliorate problems including sparsity, redundancy and high-dimensionality issue for biomedical text documents, in this research, a topic technique with fuzzy perspective is proposed. The proposed technique shows higher performance for both redundant and non-redundant biomedical text documents. The proposed technique performance is evaluated on six different real-world datasets. The main contributions of this research are listed below.

- A new hybrid inverse document frequency is proposed for global term weighting. The proposed topic modeling technique generates local and global term weighting through the bag-of-words model. Global term weighting process help in filtering the high-frequency common words in a probabilistic way.
- The high dimensionality negative effect on global term weighting is eliminated with the robust principal component analysis dimension reduction technique. After that, a fuzzy k-means clustering algorithm is utilized.
- The proposed topic modeling technique discovers the more precise and correct topics from biomedical text documents as compared to state-of-the-art topic modeling techniques. Proposed topic modeling technique performance is better on classification and clustering tasks in text mining than baseline topic models such as FLSA, LDA, and LSA.
- Topic modeling techniques LDA and LSA time consumption is increasing with various numbers of topics, but proposed topic modeling technique time consumption is stable with various numbers of topics.
- The proposed topic modeling technique remove the redundancy issue in biomedical text documents and its performance in terms of log-likelihood is higher than LDA and RedLDA.

The remainder of the paper is arranged as follows. Section 2 discusses the proposed topic modeling technique and section 3 presents the experimental results. In section 4 discussion of results of the experiment and section 5 concludes the overall paper.

II. PROPOSED TOPIC MODELING TECHNIQUE

In this section, we discuss our proposed topic modeling technique. The proposed topic modeling technique discover the more precise topics form biomedical text documents and remove the redundancy problem from these documents. Furthermore, it can be utilized for biomedical documents classification and clustering tasks in text mining. It also minimized the time consumption cost while discovering topics from big health news twitter dataset. The proposed topic modeling technique find the five matrices which are the probability of documents, the probability of words, the probability of words in documents, the probability of topics in documents and probability of words in topics. The proposed technique has the following steps.

A. STEP 1

Text documents preprocessing is an important task for text mining. Biomedical text data contain noise such as punctuations, special characters, words variations, numbers and stop words. Therefore, text data is preprocessed through following steps.

1) CONVERT TEXT DATA INTO LOWER CASE

Text datasets are converted into the lower case for preventing the various words differences.

2) TOKENIZATION

Text datasets are converted into tokens (words). The tokenization identifies the meaningful keywords from the text data. The tokens are the input for further step.

3) PUNCTUATION

Punctuation such as (“.”, “,”, “-”, “!” etc.) are eliminated from the text datasets.

4) STOP WORDS

The stops words are removed from the text datasets. Several words in a document occur very often, but they are essentially insignificant as they are utilized to join words collectively in a sentence. It is generally implicit that stop words do not cooperate to the context of text documents. The stops words high frequency of occurrence in documents, the existence of text mining becomes a hurdle in understanding the content of the document. Stop words are widely used words such as "and", "are", "this" etc. They are not helpful for document classification and clustering. So, they must get rid of documents. Therefore, all the stop words removed from the biomedical text documents.

5) SHORT AND LONG WORDS

The short words with 2 characters and long words with greater than 15 characters are eliminated from the text datasets.

6) NORMALIZATION OF WORDS

Words are normalized and inflexional ending of words are removed using the Porter stemmer.

B. STEP 2

Bag-of-word (BOW) model representation is used in information retrieval and natural language processing. BOW is utilized for representing the features in the text documents [48]. BOW model represents the occurrences of various words in a variety of documents while neglecting the words order and grammar. The BOW model makes a list of words with words count per document; this process gives words, words count, numbers of words and numbers of documents. Words count is the frequencies of occurrence of words in the documents.

C. STEP 3

Local term weight (LTW) is computed in this step. There are many LTW methods, among all one of the most popular method term frequency (TF) [49] is used in this step. Term frequency measure how frequently a term appears in a document. Documents length is different so longer documents may display the terms much longer than short documents. Therefore, term frequency is subdivided by the document length (term sum in documents) which is the normalization method. The t_f is normalized term frequency, $t_i$is terms importance in the document d_j as shown in equation 1.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_m n_{m,j}} \tag{1}$$

where, $n_{i,j}$is several term occurrences for $t_j$term in the d_j document. The denominator is the sum of the terms occurrences m in d_j documents.

D. STEP 4

In this step, the global term weighting (GTW) is calculated. First, we find $a(tf_{i,j})$ and $b_{i,j}$for global term weighting through equation 2 and 3.

$$a(tf_{i,j}) = \left\{ \begin{matrix} 1 & tf_{i,j} > 0 \\ 0 & tf_{i,j} = 0 \end{matrix} \right\} \tag{2}$$

$$b_{i,j} = \frac{tf_{i,j}}{\sum_j tf_{i,j}} \tag{3}$$

The $a(tf_{i,j})$ and $b_{i,j}$ values are find for calculating the hybrid inverse documents frequency. After that, we proposed a hybrid inverse documents frequency method (Hybrid IDF) for global term weighting. The method is proposed through the combination of inverse document frequency max $(IDF_M)$and inverse document frequency smooth (IDF_S). The hybrid inverse document frequency method has the following steps.

$$IDF_M = log\left(\frac{max\{t' \in d\}n_{t'}}{1+n_t}\right) \tag{4}$$

$$IDF_s = log\left(1+\frac{N}{n^t}\right) \tag{5}$$

$$IDF_M + IDF_S = log\left(\frac{max\{t' \in d\}n_{t'}}{1+nt}\right) + log\left(1+\frac{N}{n^t}\right) \tag{6}$$

By applying the product rule ($log_bxy = \log_b x + \log_b y$) to equation 6.

$$IDF_M + IDF_S = log\left(\frac{max\{t' \in d\}n_{t'}}{1+nt} log\left(1+\frac{N}{nt}\right)\right) \tag{7}$$

$$IDF_M + IDF_S = log\left(\frac{max\{t' \in d\}n_{t'}}{1+nt}\right) log\left(1+\frac{N}{nt}\right) \tag{8}$$

By apply quotient rule ($log_b(x/y) = \log_b x - \log_b y$) on equation 8.

$$IDF_M + IDF_S = \left(log\left(\max\left\{t' \in d\right\}n_{t'} - \log\left(1+nt\right)\right)\right) \\ \left(\log\left(nt+N\right) - log((nt))\right) \tag{9}$$

By applying the product rule ($log_bxy = log_bx + log_by$) to equation 9.

$$IDF_M + IDF_S = \left(log\left(\max\left\{t' \in d\right\}n_{t'} - \log\left(1.nt\right)\right)\right) \\ \left(\log\left(nt.N\right) - log(nt)\right) \tag{10}$$

By apply quotient rule $\left(log_b\left(\frac{x}{y}\right)\right) = log_bx - log_by$ on equation 10.

$$IDF_M + IDF_S = \log\left(\frac{\max\left\{t'\varepsilon d\right\}n_{t'}}{nt}\right)\left(\log\left(\frac{nt.N}{nt}\right)\right) \tag{11}$$

Simplify the equation 11.

$$IDF_M + IDF_S = \log\left(\left(\frac{\max\left\{t'\varepsilon d\right\}n_{t'}}{nt}\right)(N)\right) \tag{12}$$

Now simplifying the equation 12 we proposed a hybrid inverse document frequency (hybrid IDF) in equation 13.

$$Hybrid\ IDF = \log\left(\max\left\{t'\varepsilon d\right\}n_{t'}\left(\frac{N}{nt}\right)\right) \tag{13}$$

This step output documents term matrix which is TF-Hybrid IDF. After that, to eliminate the high dimensionally effect on documents term matrix and before fuzzy k-means clustering robust principal component analysis (RPCA) [49] method is utilized.

E. STEP 5

In this step, the Fuzzy k-means (FKM) clustering method is used to fuzzy clustered the documents, which represented in ten GTW methods.

The FKM algorithm partitions the data point into k clusters where $S_l, (l = 1, 2, 3, \ldots k)$are associated with the clusters centeredC_l. The data point and clusters relationship are fuzzy.

The membership $u_{i,j} \in [0, 1]$represents belonging of clusters centers C_j and data point X_i. The set of data point is $S = \{X_i\}$. The fuzzy k means algorithm based on minimizing distortion as shown in equation 14. Where $u_{i,j}$is membership and $C_j$represents clusters.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{i,j}^{q} d_{i,j} \tag{14}$$

N represents the number of data points, fuzzifier parameter is q; numbers of clusters are represented by k and $d_{i,j}$ is the squared Euclidean distance between cluster representative C_j and data point X_i. The fuzzification parameters q significantly influence the resultant clusters.

The fuzzy k-means clustering mapping the representative vectors and improving through the partition of the data point. Its start with initial clusters centers and repeat the process until stop criteria satisfied. It is supposed that no two clusters are the same representative. If $d_{i,j} < n$ then, $u_{i,j} = 1$ and $u_{i,j} = 0$ for $l \neq j$ where n is a positive number. Now the fuzzy k-means algorithm performs the following steps.

Set the initial clusters SC_o equal to$(C_j(0))$ and ε value $p = 1$.

Set of clusters SC_p is given and compute $d_{i,j}$for $i = 1$ to$N$, $j = 1$ to k and update the memberships $u_{i,j}$using equation 15. The $u_{i,j}$membership degree for each document (D) with clusters (topics). Value of $u_{i,j}$ is the probability of topics i with documents j as $P(T_i|D_j)$. The m is a fuzzification coefficient.

$$u_{(i,j)} = ((d_{(i,j)}{}^{(1/m-1)} \sum_{l=1}^{k} (\frac{1}{d_{il}})^{\frac{1}{m}-1}))^{-1} \tag{15}$$

where, $d_{i,j} < n$and value of n is very small and $u_{i,j} = 1$. $P(T_i|D_j)$ can be used to find (words × topics) matrix.

Using equation 15 centers for every cluster are computed to get new clusters which are represented by $SC_p + 1$.

$$C_j(p) = \frac{\sum_{j=1}^{N} u_{i,j}{}^{m} X_i}{\sum_{j=1}^{N} u_{i,j}{}^{m}} \tag{16}$$

If the $\left(\left\| C_j(p) - C_j(p-1) \right\|\right) <\in$ and $j = 1$ to k, then stop where $\in>$ *zero* is the small positive number. Otherwise set $p{+}1 \to p$ and move to step 2 of the fuzzy k-means algorithm.

In terms of numbers of calculation, the computational complexity of fuzzy-k means is $O(N_{kt})$and t represents several iterations.

F. STEP 6

The probability of documents $P(D_j)$ is found through documents term matrix and hybrid IDF (words × documents matrix) as shown in equation 17. Where, i and j represent the words and documents.

$$P(D_j) = \frac{\sum_{i=1}^{m} (W_i, D_j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} (W_i, D_j)} \tag{17}$$

G. STEP 7

The probability of documents over topics $P(D_j|T_k)$ is found through the probability of documents $P(D_j)$ and probability of topics over documents $P(T_k|D_j)$ as shown in equation 18. Where j are documents in k topics.

$$P(D_j, T_k) = P(T_k|D_j) \times P(D_j) \tag{18}$$

After finding $P(D_j, T_k)$, normalize $P(D, T)$ using equation 19 for each topic.

$$P(D_j|T_k) = \frac{P(D_j, T_k)}{\sum_{j=1}^{n} P(D_j, T_k)} \tag{19}$$

H. STEP 8

In this step, the probability of words i in given documents j and $P(W_i|D_j)$is derived using the document term matrix and global term weight methods as shown in equation 20.

The probability of words given documents $P(W_i|D_j)$ is find using the document-term matrix and global term weighting as shown in equation 20. Where i words in j documents.

$$P(W_i|D_j) = \frac{P(W_i, D_j)}{\sum_{i=1}^{m} P(W_i, D_j)} \tag{20}$$

I. STEP 9

The probability of words in topics $P(W_i|T_k)$is found through the probability of documents in topics $P(D_j|T_k)$ and probability of words in documents $P(W_i|D_j)$ as shown in equation 21. Where i words in k topics.

$$P(W_i|T_k) = \sum_{j=1}^{n} P(W_i, D_j) \times P(D_j|T_k) \tag{21}$$

III. EXPERIMENTAL RESULTS

In this section performance of proposed topic modeling technique in terms of classification, clustering, redundancy and execution time is evaluated.

A. DATASETS

Six publicly available datasets are used in this research.

- The first dataset is Ohsumed Collection labeled dataset which is a medical abstract of MeSH categories. In this research categories including bacterial infections, mycoses and virus diseases are selected. Weblink:(http://disi.unitn.it/moschitti/corpora/ohsumed-first-20000-docs.tar. gz.)
- The second dataset is MuchMore Springer Bilingual Corpus labeled dataset which is English scientific corpus abstracts. Two categories of federal health standard sheet and arthroscopy are selected for experiments. Weblink:(http:// http://muchmore.dfki.de/resources1.htm)

- The third dataset is GENIA corpus which is a collection of Medline articles abstracts and representing the molecular biology literature [51]. Weblink:(http://neuro.imm.dtu.dk/wiki/GENIA)
- The fourth dataset is Biotext which contain abstracts of disease and treatments and collected from Medline [52]. Weblink:(http://biotext.berkeley.edu/data/dis treat data.html)
- The fifth dataset is tweets dataset which is a big dataset in this research. Weblink:(https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter)
- Sixth datasets are wall street journal (WSJ) redundant corpus and broadly utilized in natural language processing [53, 54].

B. DATASETS STATISTICS

Table 1 show the statistics of six datasets, which are used in this research.

TABLE 1. Basic statistics of datasets.

Dataset Names	Documents Preprocess	Terms	Unique Terms
Ohsumed	2092	22669	13238
MuchMore			
Springer	1527	19835	5008
GENIA Corpus	2000	21560	17834
Biotext	40	25921	10267
Twitter	58,927	395,635	25,309
WSJ	1300	680K	36K

C. DOCUMENTS CLASSIFICATION

Documents classification is performed on Ohsumed and MuchMore springer labeled datasets with Bayesian optimization. Documents classification assign document to one or more classes and features are extracted from these documents. BOW create high dimensionality effect for documents classification. Therefore, the proposed topic modeling technique reduce numbers of features and find meaningful words in topics. Optimization is the process of determining points that reduce the actual value function, called the objective function. Bayesian optimization is a Gaussian process model of the objective function and uses the objective function evaluation to train the model. Bayesian optimization is used to minimize errors in cross-evaluation reactions using Bayesian optimization. The appropriate function Fit in MATLAB is utilized for Bayesian optimization. Documents classification is performed on P(T|D) using the discriminant analysis classifier with Bayesian optimization. The performance of the proposed topic modeling technique is measured with FLSA, LDA, and LSA through tenfold cross-validation technique. In this technique, data is divided into ten subsets for the ten iterations. Proposed topic modeling technique classification performance measured with 50, 100, 150 and 200 topics with input of the features for documents. The classification results are evaluated in terms of precision, recall, accuracy, and F1-measure. The results of discriminant analysis classifier are measured using the confusion matrix as shown in Table 2. Where, true negative(TN), false positive(FP), false negative (FN) and true positive (TP) are correct predictions with negative instance, incorrect predictions with positive instance, incorrect predictions with negative instance and correct prediction with the positive instance. The performance measurement precision, recall, accuracy, and F1-measure formulas are described in equation 22, 23, 24 and 25.

TABLE 2. Confusion matrix.

Actual	Predicated	-
-	Negative	Positive
Negative	TN	FP
Positive	FN	TP

TABLE 3. Ohsumed datasets classification results on 50-200 topics.

Method	Accuracy (%)	Precision	Recall	F1 Score	topics
LSA	48.36	0.4146	0.4224	0.4185	50
LDA	54.10	0.4798	0.5155	0.4970	50
FLSA(Entropy)	75.21	0.720	0.722	0.746	50
FLSA(IDF)	75.90	0.722	0.723	0.746	50
FLSA(Normal)	71.25	0.6551	0.654	0.677	50
FLSA(ProbIDF)	74.87	0.715	0.714	0.735	50
Proposed	**92.35**	**0.9236**	**0.9006**	**0.9119**	**50**
LSA	51.37	0.4430	0.4099	0.4258	100
LDA	54.92	0.4873	0.4783	0.4828	100
FLSA(Entropy)	76.24	0.727	0.726	0.747	100
FLSA(IDF)	74.35	0.701	0.703	0.726	100
FLSA(Normal)	71.08	0.670	0.674	0.694	100
FLSA(ProbIDF)	74.52	0.702	0.704	0.724	100
Proposed	**87.70**	**0.8867**	**0.8261**	**0.8553**	**100**
LSA	52.73	0.4651	0.4969	0.4805	150
LDA	57.10	0.5123	0.5155	0.5139	150
FLSA(Entropy)	74.87	0.715	0.714	0.735	150
FLSA(IDF)	76.59	0.732	0.731	0.752	150
FLSA(Normal)	72.46	0.671	0.673	0.691	150
FLSA(ProbIDF)	75.04	0.715	0.712	0.735	150
Proposed	**90.16**	**0.8788**	**0.9006**	**0.8896**	**150**
LSA	49.73	0.4303	0.4410	0.4356	200
LDA	54.37	0.4819	0.4969	0.4893	200
FLSA(Entropy)	75.21	0.720	0.721	0.74	200
FLSA(IDF)	74.18	0.705	0.704	0.725	200
FLSA(Normal)	71.94	0.671	0.673	0.683	200
FLSA(ProbIDF)	74.87	0.701	0.702	0.729	200
Proposed	**88.25**	**0.8986**	**0.8261**	**0.8608**	**200**

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

$$F1\_Score = 2 \times \frac{Precision * Recall}{(Precision + Recall)} \tag{25}$$

The experimental results of the classification are shown in table 3 and 4 on Ohsumed and MuchMore Springer datasets. The proposed topic modeling technique is compared with the baseline topic models FLSA [47], LDA [5] and LSA [7].

D. DOCUMENTS CLUSTERING

The performance of documents clustering is measured on Genia corpus and Biotext dataset. The documents clustering

TABLE 4. MuchMore springer datasets classification results on 50-200 topics.

Method	Accuracy (%)	Precision	Recall	F1 Score	topics
LSA	57.52	0.6667	0.7221	0.6933	50
LDA	60.95	0.6938	0.7356	0.7141	50
FLSA(Entropy)	97.66	0.955	0.9554	0.977	50
FLSA-(IDF)	95.90	0.937	0.935	0.959	50
FLSA(Normal)	91.22	0.890	0.894	0.912	50
FLSA(ProbIDF)	97.66	0.954	0.953	0.977	50
Proposed	**98.29**	**0.9880**	**0.9883**	**0.9880**	**50**
LSA	56.19	0.6676	0.6791	0.6733	100
LDA	58.85	0.6854	0.7011	0.6932	100
FLSA(Entropy)	96.49	0.943	0.942	0.965	100
FLSA(IDF)	98.24	0.961	0.960	0.982	100
FLSA(Normal)	92.39	0.902	0.900	0.924	100
FLSA(ProbIDF)	97.66	0.955	0.952	0.977	100
Proposed	**98.87**	**0.9879**	**0.9841**	**0.9844**	**100**
LSA	62.67	0.7051	0.7536	0.7285	150
LDA	59.23	0.6991	0.6791	0.6890	150
FLSA(Entropy)	95.90	0.937	0.935	0.959	150
FLSA(IDF)	97.66	0.955	0.952	0.977	150
FLSA-(Normal)	95.32	0.932	0.931	0.953	150
FLSA(ProbIDF)	97.07	0.950	0.952	0.971	150
Proposed	**98.97**	**0.9822**	**0.9882**	**0.9886**	**150**
LSA	60.00	0.6980	0.7020	0.7000	200
LDA	63.42	0.7039	0.7765	0.7384	200
FLSA(Entropy)	97.076	0.950	0.9501	0.971	200
FLSA-(IDF)	97.66	0.955	0.9553	0.977	200
FLSA(Normal)	92.39	0.901	0.902	0.924	200
FLSA(ProbIDF)	97.66	0.955	0.950	0.977	200
Proposed	**98.86**	**0.9883**	**0.9870**	**0.9859**	**200**

is conducted on P(T|D). The internal validation method is used for clustering measurement which is a better choice as compared to external validation [55]. The different number of clusters and topics are measured through Calinsiki-Har-abasz (CH) index [56] internal validation method with k-means clustering. In CH index cohesion is computed which is the distance between clusters to centroids as described in equation 26. The higher value of CH index indicates the better performance of clustering. The CH index evaluated the validation of clusters on the average sum of squared error cluster between and within the clusters.

$$CH(C) = \frac{(N-K)}{(K-1)} \frac{\sum_{ck} \in C |C_k| d_e|(\bar{C}_k, \bar{X})}{\sum_{ck} \in C \sum_{xi} C_k d_e(x^i, \bar{C}_k)} \tag{26}$$

Fig.1,2, 3, 4, 5, 6, 7 and 8 shows the clustering results of proposed topic modeling technique as compared to FLSA, LDA, and LSA using CH index on Genia corpus and Biotext datasets.

E. REDUNDANCY ISSUE

Log-likelihood for Synthetic WSJ redundant corpora with different numbers of topics from 50 to 450 is shown in Fig.9. The proposed topic modeling technique is compared with RedLDA and LDA for redundant corpora.

F. EXECUTION TIME

Fig.10 shows the execution time of the proposed topic modeling technique with a comparison to LDA and LSA.

IV. DISCUSSION

The experimental results of a classification in table 3 and 4 shows that proposed topic modeling technique performance is better and higher against FLSA, LDA, and LSA with several numbers of topics. The proposed topic technique precision, recall, accuracy, and F1 score values are greater than FLSA, LDA, and LSA. The classification accuracy of the proposed topic modeling technique is 92.35, 87.70, 90.16, 88.25 and 98.29, 98.87, 98.97, 98.86 percent on Ohsumed and Much-More Springer Corpus datasets. The classification accuracy of the proposed topic modeling technique is higher than the baseline topic models FLSA, LDA, and LSA. Proposed topic modeling technique performance of clustering is evaluated

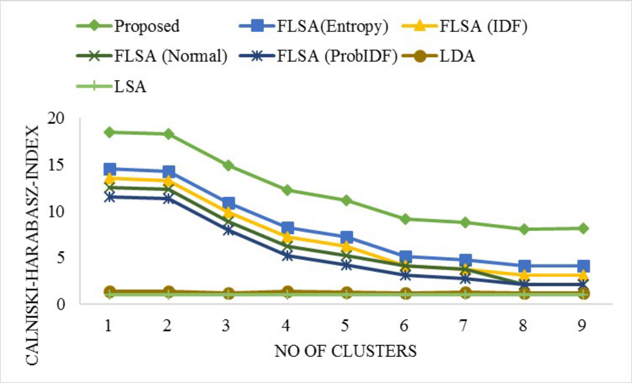


FIGURE 1. Calinski-Harabasz(CH) results for 50 topics of Genia corpus.

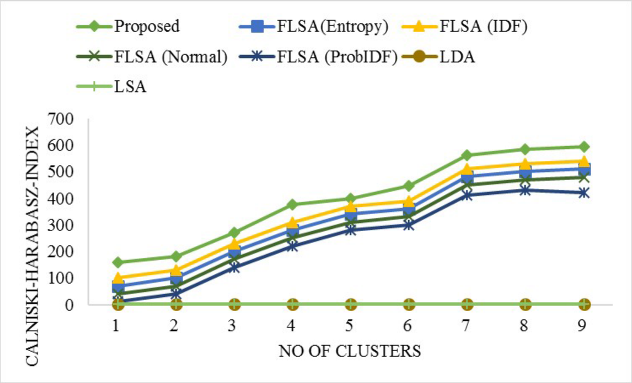


FIGURE 2. Calinski-Harabasz(CH) results for 100 topics of Genia corpus.

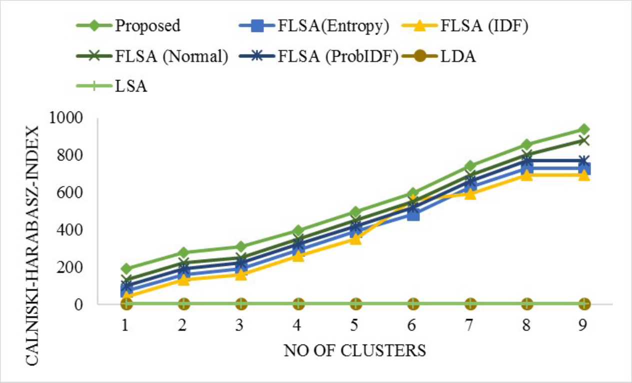


FIGURE 3. Calinski-Harabasz(CH) results for 150 topics of Genia corpus.

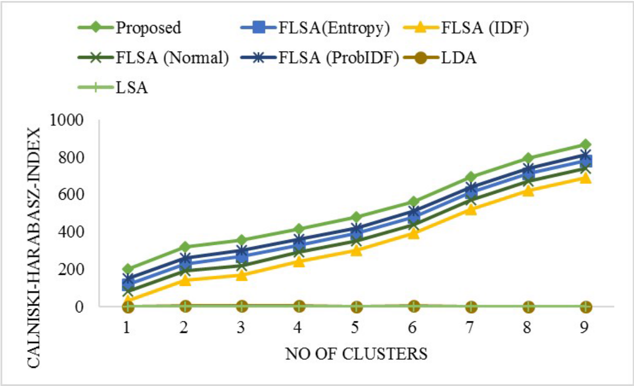


FIGURE 4. Calinski-Harabasz(CH) results for 200 topics of Genia corpus.

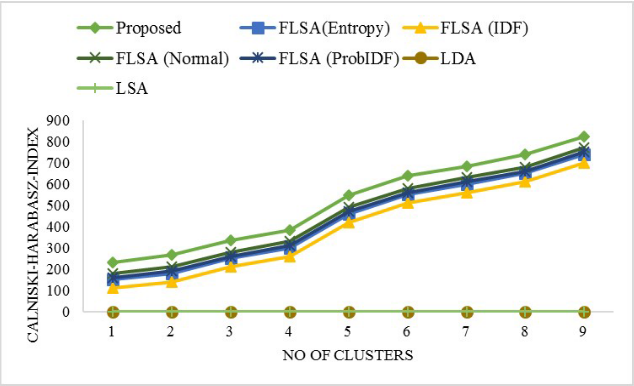


FIGURE 7. Calinski-Harabasz(CH) results for 150 topics of Biotext.

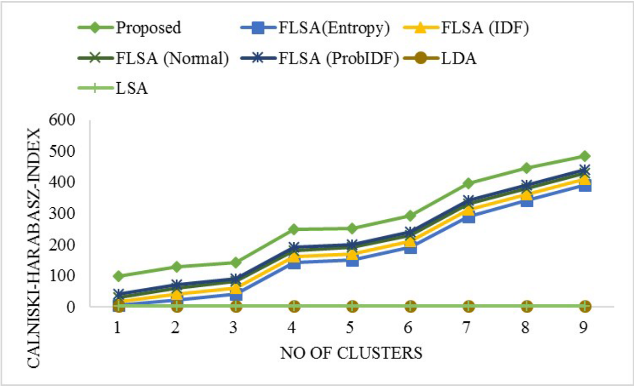


FIGURE 5. Calinski-Harabasz(CH) results for 50 topics of Biotext.

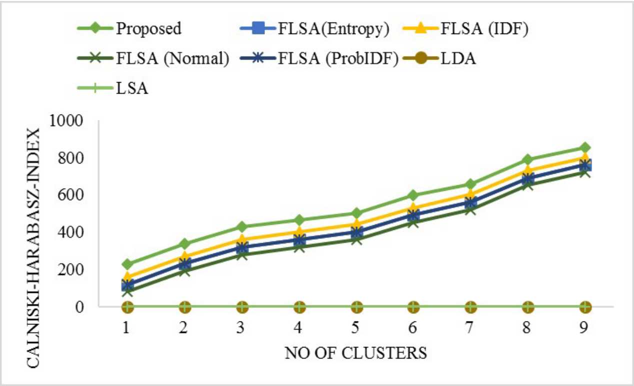


FIGURE 8. Calinski-Harabasz(CH) results for 200 topics of Biotext.

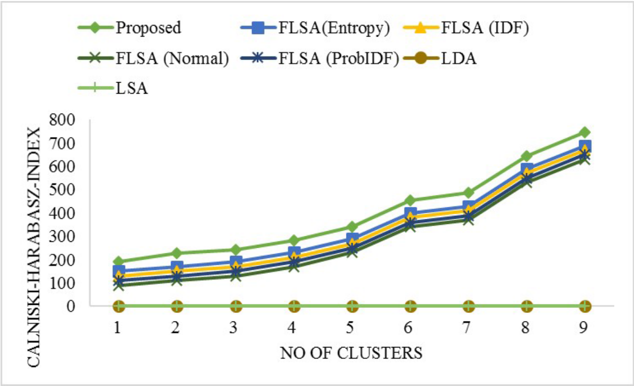


FIGURE 6. Calinski-Harabasz(CH) results for 100 topics of Biotext.

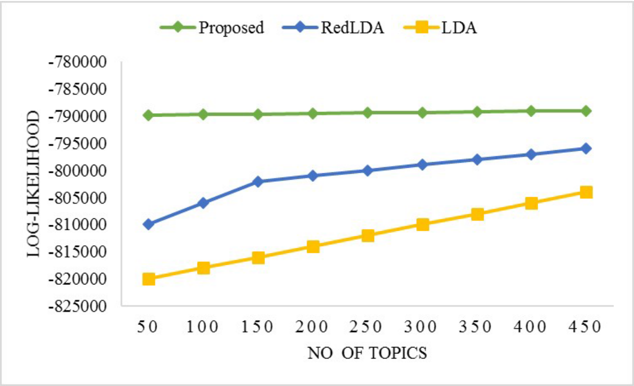


FIGURE 9. Loglikelihood comparison for WSJ corpora.

and compared to FLSA, LDA, and LSA with various number of clusters range from 0 to 8 on different numbers of topics such as 50, 100,150 and 200. CH index showing that proposed topic modeling technique performance is greater and better than FLSA, LDA, and LSA with various numbers of topics as shown in the Fig.1, 2, 3, 4, 5, 6, 7 and 8. Redundancy effect is examined with Synthetic WSJ redundant corpora. The proposed topic modeling technique is compared with LDA and RedLDA which used for removing the redundancy issue in biomedical text corpora [32]. Topic modeling techniques proposed, RedLDA and LDA are trained on the same Synthetic WSJ redundant corpora for performance comparison through log-likelihood. A log-likelihood greater score indicating the better performance of generalization and topic modeling technique is more successful in modeling the document's structure for text corpora. Log-likelihood for Synthetic WSJ redundant corpora with different numbers of topics from 50 to 450 is shown in Fig.9. The experiments show that the proposed topic modeling technique performance is better than LDA and RedLDA for redundant text

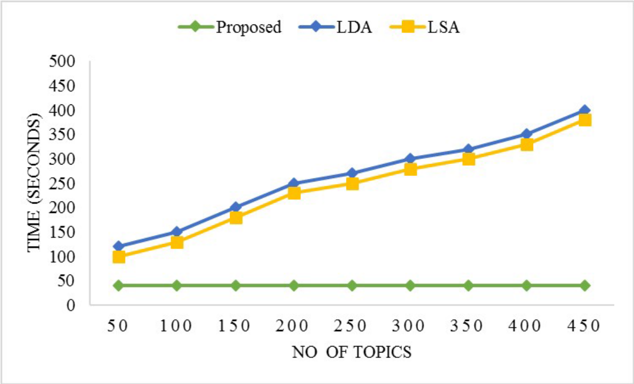


FIGURE 10. Execution time comparison with health new tweets dataset.

corpora. The proposed topic modeling technique execution time is compared with LDA and LSA using tweets of health news big dataset. Topic modeling process based on the probability distribution of topics and words and documents with greater probability in every topic with utilization of posterior distribution. In this comparison Gibbs sampling most, prominent method is used for LDA. The algorithm enhanced the computational cost for documents, topics, and words due to several numbers of iterations. Fig.10 shows that proposed technique execution time is stable with increased numbers of topics and improved than its competitors LDA and LSA.

V. CONCLUSION

Biomedical text documents are continuously increasing nowadays while analyzing these documents is very important for discovering the valuable resource of information. Archives of biomedical text documents like PubMed is providing valuable services for the Scientific Community. Topic Modeling is a popular method that discovers the hidden theme and structure in unorganized biomedical text documents. These documents structure is used for searching, indexing and summarizing of documents. In machine learning, fuzzy techniques are widely used for biomedical image processing and text processing. Existing topic modeling techniques are the focus on linear algebra and statistical distribution approaches. In this paper, we proposed a topic modeling technique that discovers the latent semantic topics from biomedical documents. The proposed topic modeling technique obviates the negative effect of word redundancy in biomedical documents and its performance is enhanced and greater than RedLDA and LDA for redundant corpora. Proposed topic modeling technique improves the classification accuracy for biomedical datasets and provides a new technique for text mining over biomedical datasets. The clustering results of proposed topic modeling technique are better with various numbers of topics. Furthermore, experimental results show that the time performance of the proposed topic modeling technique is stable with increased in numbers of topics. The proposed topic modeling technique has the flexibleness to work with an extensive variety of fuzzy clustering and dimension reduction techniques. Additionally, the proposed topic modeling technique works with discrete and continuous data and estimates the number of topics in biomedical documents. Quantitative evaluation of six datasets shows that the proposed technique outperforms progressive baselines with vital enhancements. The experimental results indicate that the proposed topic modeling technique is a strong method that identifies the hidden structure in the biomedical dataset. Experiments results also show that proposed topic modeling technique classification and clustering performance is higher than state-of-the-art baselines topic models such as FLSA, LDA, and LSA.

REFERENCES

[1] V. Renganathan, “Text mining in biomedical domain with emphasis on document clustering,” *Healthcare Inform. Res.*, vol. 23, no. 3, pp. 141–146, 2017.

[2] D. Zhou and Y. He, “Extracting interactions between proteins from the literature,” *J. Biomed. Inform.*, vol. 41, no. 2, pp. 393–407, Apr. 2008.

[3] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.

[4] A. Holzinger, J. Schantl, M. Schroettner, C. Seifert, and K. Verspoor, “Biomedical text mining: State-of-the-art, open problems and future challenges,” in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Berlin, Germany: Springer, 2014, pp. 271–300.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[6] T. K. Landauer, D. Laham, B. Rehder, and M. E. Schreiner, “How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans,” in *Proc. 19th Annu. Meeting Cogn. Sci. Soc.*, 1997, pp. 412–417.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[8] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proc. ACM SIGIR Forum*, 2017, pp. 211–218.

[9] J. O. Wrenn, D. M. Stein, S. Bakken, and P. D. Stetson, “Quantifying clinical narrative redundancy in an electronic health record,” *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 1, pp. 49–53, 2010.

[10] R. Cohen, M. Elhadad, and N. Elhadad, “Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies,” *BMC Bioinf.*, vol. 14, p. 10, Jan. 2013.

[11] C. Lee, Z. Luo, K. Y. Ngiam, M. Zhang, K. Zheng, G. Chen, B. C. Ooi, and W. L. J. Yip, “Big healthcare data analytics: Challenges and applications,” in *Handbook of Large-Scale Distributed Computing in Smart Healthcare*. Cham, Switzerland: Springer, 2017, pp. 11–41.

[12] K. Tian, M. Revelle, and D. Poshyvanyk, “Using latent Dirichlet allocation for automatic categorization of software,” in *Proc. 6th IEEE Int. Work. Conf. Mining Softw. Repositories*, May 2009, pp. 163–166.

[13] Z. Zhai, B. Liu, H. Xu, and P. Jia, “Constrained LDA for grouping product features in opinion mining,” in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2011, pp. 448–459.

[14] Q. Wu, C. Zhang, Q. Hong, and L. Chen, “Topic evolution based on LDA and HMM and its application in stem cell research,” *J. Inf. Sci.*, vol. 40, no. 5, pp. 611–620, 2014.

[15] A. Bagheri, M. Saraee, and F. de Jong, “ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences,” *J. Inf. Sci.*, vol. 40, no. 5, pp. 621–636, 2014.

[16] L. Hong and B. D. Davison, “Empirical study of topic modeling in Twitter,” in *Proc. 1st Workshop Social Media Anal.*, Jul. 2010, pp. 80–88.

[17] Z. Chen, Y. Huang, J. Tian, X. Liu, K. Fu, and T. Huang, “Joint model for subsentence-level sentiment analysis with Markov logic,” *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 9, pp. 1913–1922, Sep. 2015.

[18] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, “An overview of topic modeling and its current applications in bioinformatics,” *SpringerPlus*, vol. 5, p. 1608, Dec. 2016.

[19] H. Wang, M. Huang, and X. Zhu, “Extract interaction detection methods from the biological literature,” *BMC Bioinf.*, vol. 10, no. 1, p. S55, 2009.

[20] C. W. Arnold, S. M. El-Saden, A. A. Bui, and R. Taira, "Clinical case-based retrieval using latent topic analysis," in *Proc. AMIA Annu. Symp.*, 2010, p. 26.

[21] M. Song and S. Y. Kim, "Detecting the knowledge structure of bioinformatics by mining full-text collections," *Scientometrics*, vol. 96, no. 1, pp. 183–201, Jul. 2013.

[22] E. Sarioglu, K. Yadav, and H.-A. Choi, "Topic modeling based classification of clinical reports," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics Student Res. Workshop*, 2013, pp. 67–73.

[23] H. Bisgin, Z. Liu, H. Fang, X. Xu, and W. Tong, "Mining FDA drug labels using an unsupervised learning technique-topic modeling," *BMC Bioinf.*, vol. 12, no. 10, p. S11, 2011.

[24] T. Asou and K. Eguchi, "Predicting protein-protein relationships from literature using collapsed variational latent Dirichlet allocation," in *Proc. 2nd Int. Workshop Data Text Mining Bioinf.*, Oct. 2008, pp. 77–80.

[25] V. Wang, L. Xi, A. Enayetallah, E. Fauman, and D. Ziemek, "GeneTopics-interpretation of gene sets via literature-driven topic models," *BMC Syst. Biol.*, vol. 7, no. 5, p. S10, Dec. 2013.

[26] H. Bisgin, M. Chen, Y. Wang, R. Kelly, H. Fang, X. Xu, and W. Tong, "A systems approach for analysis of high content screening assay data with topic modeling," *BMC Bioinf.*, vol. 14, no. 14, p. S11, 2013.

[27] H. Bisgin, Z. Liu, R. Kelly, H. Fang, X. Xu, and W. Tong, "Investigating drug repositioning opportunities in FDA drug labels through topic modeling," *BMC Bioinf.*, vol. 13, no. 15, p. S6, 2012.

[28] S.-H. Wang, Y. Ding, W. Zhao, Y.-H. Huang, R. Perkins, W. Zou, and J. J. Chen, "Text mining for identifying topics in the literatures about adolescent substance use and depression," *BMC Public Health*, vol. 16, p. 279, Mar. 2016.

[29] X. Wang, P. Zhu, T. Liu, and K. Xu, "BioTopic: A topic-driven biological literature mining system," *Int. J. Data Mining Bioinf.*, vol. 14, no. 4, pp. 373–386, 2016.

[30] R. Sullivan, A. Sarker, and K. O'Connor, A. Goodin, M. Karlsrud, and G. Gonzalez, "Finding potentially unsafe nutritional supplements from user reviews with topic modeling," in *Proc. Biocomput. Pacific Symp.*, 2016, pp. 528–539.

[31] J. H. Chen, M. K. Goldstein, S. M. Asch, L. Mackey, and R. B. Altman, "Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 3, pp. 472–480, 2017.

[32] R. Cohen, I. Aviram, M. Elhadad, and N. Elhadad, "Redundancy-aware topic modeling for patient record notes," *PloS ONE*, vol. 9, Feb. 2014, Art. no. e87555.

[33] W. Kintsch, "The potential of latent semantic analysis for machine grading of clinical case summaries," *J. Biomed. Inform.*, vol. 35, no. 1, pp. 3–7, Feb. 2002.

[34] T. Cohen, B. Blatter, and V. Patel, "Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative," *J. Biomed. Inform.*, vol. 41, no. 6, pp. 1070–1087, Dec. 2008.

[35] J.-F. Yeh, C.-H. Wu, and M.-J. Chen, "Ontology-based speech act identification in a bilingual dialog system using partial pattern trees," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 59, no. 5, pp. 684–694, Mar. 2008.

[36] F. Ginter, H. Suominen, S. Pyysalo, and T. Salakoski, "Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application," *Int. J. Med. Inform.*, vol. 78, no. 12, pp. e1–e6, Dec. 2009.

[37] F. Masulli and A. Schenone, "A fuzzy clustering based segmentation system as support to diagnosis in medical imaging," *Artif. Intell. Med.*, vol. 16, no. 2, pp. 129–147, Jun. 1999.

[38] B. Andreopoulos, A. An, X. Wang, and M. Schroeder, "A roadmap of clustering algorithms: Finding a match for a biomedical application," *Brief Bioinf.*, vol. 10, pp. 297–314, May 2009.

[39] C.-H. Oh, K. Honda, and H. Ichihashi, "Fuzzy clustering for categorical multivariate data," in *Proc. Joint 9th IFSA World Congr. 20th NAFIPS Int. Conf.*, Jul. 2001, pp. 2154–2159.

[40] Y. Wu, H. Duan, and S. Du, "Multiple fuzzy c-means clustering algorithm in medical diagnosis," *Technol. Health Care*, vol. 23, no. S2, pp. S519–S527, 2015.

[41] D. Aneja and T. K. Rawat, "Fuzzy clustering algorithms for effective medical image segmentation," *Int. J. Intell. Syst. Appl.*, vol. 5, no. 11, p. 55, 2013.

[42] D. Li, W. Pedrycz, and N. J. Pizzi, "Fuzzy wavelet packet based feature extraction method and its application to biomedical signal classification," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 6, pp. 1132–1139, Jun. 2005.

[43] L. Di Lascio, A. Gisolfi, A. Albunia, G. Galardi, and F. Meschi, "A fuzzy-based methodology for the analysis of diabetic neuropathy," *Fuzzy Sets Syst.*, vol. 129, no. 2, pp. 203–228, Jul. 2002.

[44] E. I. Papageorgiou, C. D. Stylios, and P. P. Groumpos, "An integrated two-level hierarchical system for decision making in radiation therapy based on fuzzy cognitive maps," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 12, pp. 1326–1339, Dec. 2003.

[45] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biol.*, vol. 3, no. 11, 2002, Art. no. research0059-1.

[46] W. K. Moon, S.-C. Chang, C.-S. Huang, and R.-F. Chang, "Breast tumor classification using fuzzy clustering for breast elastography," *Ultrasound Med. Biol.*, vol. 37, no. 5, pp. 700–708, May 2011.

[47] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, "Fuzzy approach topic discovery in health and medical corpora," *Int. J. Fuzzy Syst.*, vol. 20, no. 4, pp. 1334–1345, 2018.

[48] Y. Zhang, R. Jin, and Z. Zhou, "Understanding bag-of-words model: A statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, nos. 1–4, pp. 43–52, Dec. 2010.

[49] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[50] A. Podosinnikova, S. Setzer, and M. Hein, "Robust PCA: Optimization of the robust reconstruction error over the Stiefel manifold," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 121–131.

[51] J.-D. Kim, T. Ohta, Y. Tateisi, and J. I. Tsujii, "GENIA corpus—A semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, pp. i180–i182, Jul. 2003.

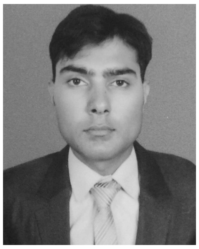
[52] B. Rosario and M. A. Hearst, "Classifying semantic relations in bioscience texts," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2004, p. 430.

[53] D. Gildea, "Corpus variation and parser performance," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2001, pp. 1–6.

[54] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "Developing a robust part-of-speech tagger for biomedical text," in *Proc. Panhellenic Conf. Inform.*, 2005, pp. 382–392.

[55] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27–34, 2011.

[56] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat., Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.



JUNAID RASHID received the B.S. (CS) degree from the COMSATS Institute of Information Technology, Wah, Pakistan, in 2014, and the M.S. (CS) degree from the Department of Computer Science, COMSATS Institute of Information Technology, Wah, in 2016. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Engineering and Technology Taxila, Pakistan.

He has published many research articles in international journals and conferences. He has been interested in research domains like topic modeling, text mining, information retrieval, machine learning, fuzzy systems, software engineering, data science, and so on. He is a member of the IEICE. He gets the fully funded scholarship for his Ph.D. degree from the University of Engineering and Technology. He has served/been serving as a Reviewer for IEEE Access, the *Journal of Intelligent and Fuzzy Systems*, the *Journal of Software: Evolution and Process*, the *Technical Journal UET Taxila*, and the IEEE International Conference on Biomedical and Health Informatics (the IEEE BHI 2019), USA.

SYED MUHAMMAD ADNAN SHAH received the M.S. degree in computer engineering from the CASE (Center for Advanced Studies in Engineering) Islamabad, Pakistan, in 2010, and the Ph.D. degree in computer engineering from the University of Engineering and Technology Taxila (UET Taxila), Pakistan, in 2014, where he is currently an Assistant Professor with the Department of Computer Science. His research interests include acoustic scene analysis, topic modeling, multimedia signal processing, and machine learning.

AUN IRTAZA received the Ph.D. degree from FAST-nu, Islamabad, Pakistan, in 2016. During his Ph.D., he remained working as a Research Scientist with the Gwangju Institute of Science and Technology (GIST), South Korea. In 2017, he became an Associate Professor and the Department of Computer Science Chair with the University of Engineering and Technology (UET) Taxila, Pakistan, in 2018. He is currently working as a Visiting Associate Professor with the University of Michigan–Dearborn. His research areas include computer vision, multimedia forensics, audio–signal processing, medical image processing, and big data analytics. He has more than 40 publications in the IEEE, Springer, and Elsevier journals.

TOQEER MAHMOOD received the M.S. degree from the Center for Advanced Studies in Engineering (CASE) Islamabad, Pakistan, in 2010, and the Ph.D. degree from the University of Engineering and Technology Taxila, Pakistan, in 2017, both in computer engineering. He has authored or coauthored many scientific articles in conferences and journals of international repute. He has been serving as a Reviewer for the *Technical Journal UET Taxila*, the *Journal of King Saud University—Computer and Information Sciences*, the *ETRI Journal*, the *Journal of Information Security and Applications*, the *Australian Journal of Forensic Sciences*, *Forensic Science International*, and many more. His research interests include image processing, image retrieval, steganography, and numerical techniques with particular attention to multimedia forensics.

MUHAMMAD WASIF NISAR received the B.Sc. and M.Sc. degrees from the University of Peshawar, Pakistan, in 1998 and 2000, respectively, and the Ph.D. degree from the Institute of Software, GUCAS China, in 2008, all in computer science. He is currently serving as the Head and also an Associate Professor with the Department of Computer Science, COMSATS University Islamabad, Wah Campus, Pakistan. His research interests include software engineering, data mining, distributed systems, semantic search engines, and machine learning.

MUHAMMAD SHAFIQ received the M.S. degree in computer science from the University Institute of Information Technology, Arid Agriculture University, Rawalpindi, Pakistan, the master's degree in information technology from the University of the Punjab, Gujranwala, Pakistan, and the Ph.D. degree in information and communication engineering from Yeungnam University, South Korea, in 2018. He was a Lecturer with the Department of Computer Science, Federal Urdu University, Islamabad, Pakistan. From 2010 to 2018, he was a Lecturer with the Department of Information Technology, University of Gujrat. He is currently a Postdoctoral Fellow with Yeungnam University, and is also an Assistant Professor with the Faculty of Computer Science, GC Women University, Sialkot, Pakistan. His research interests include the data mining, machine learning, design of spectrum management, routing, and medium access control protocols for mobile ad hoc networks, the Internet of Things, and cognitive radio networks.

AKBER GARDEZI received the master's and Ph.D. degrees from the University of Sussex, U.K. After completion of the Ph.D. degree, he was awarded a Postdoctoral Tutorial Fellowship with the Department of Engineering and Design, University of Sussex, where he was closely involved with projects on embedded systems for a period of two years. At CIIT, apart from being an Assistant Professor, he also holds the additional charge of managing the Inter Islamic Network on Information Technology (INIT). He is currently an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad. From the platform of INIT, he is involved in collaboration with the UNESCO Chair in ICTs for Development. The initiative is a five year programme focusing on a three tiered approach of advocacy, policy making, and development of novel technologies for the marginalized. Considering his background, he is closely involved with projects relating toward ICTs for development; hence utilizing his network of peers in terms of launching advocacy and awareness initiatives for the marginalized communities within the OIC member states. Subsequently, his involvement with INIT has steered his research interests toward ICTs for development, reinforcing the belief that the technology is indeed a vehicle for betterment for the most marginalized.