Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 10(106)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10101071

A hybrid approach for generative process model with topic modelling towards efficient and dynamic document clustering

Gugulothu Venkanna[1*] and K.F Bharati[2]
Research Scholar, Department of Computer Science & Engineering, Jawaharlal Nehru Technological University, Anantapuramu, India[1]
Associate Professor, Department of Computer Science & Engineering, Jawaharlal Nehru Technological University, Ananthapuramu, India[2]

Received: 25-January-2023; Revised: 21-September-2023; Accepted: 23-September-2023

Abstract

Clustering text documents has a wide range of applications across various domains. However, due to the diversity and rapid growth of textual data, performing clustering on a given text corpus has become increasingly challenging. Several existing approaches for text document clustering rely on natural language processing (NLP) and text similarity measures. However, there is a pressing need for a generative process model to systematically and progressively handle text corpora. Furthermore, a hybrid approach that enhances clustering performance is essential. Therefore, developing a model for a given text corpus and dynamically updating it as new documents arrive, rather than starting clustering from scratch, is of paramount importance. In this paper, a framework known as the hybrid approach for dynamic document clustering (HADDC) was proposed. This framework is realized through the definition of two algorithms that collaborate to achieve dynamic document clustering. The first algorithm, called similar document identification (SDI), leverages a lexical dictionary, WordNet, and similarity measures to effectively identify similar documents. The second algorithm, topic modelling for efficient and dynamic document clustering (TM-EDDC), is designed as a dynamic process model based on latent Dirichlet allocation (LDA). It has the capability to cluster documents incrementally as new ones become available. Experimental results demonstrate that the proposed methods outperform existing ones, as evidenced by a lower mean absolute error (MAE). The proposed framework and underlying algorithms were evaluated using the news groups dataset. The empirical study showcases the enhanced utility and efficiency of the proposed framework, making it a valuable tool for organizations to integrate into their existing applications.

Keywords

Document clustering, Natural language processing, Generative process model, Document similarity, Dynamic document clustering.

1.Introduction

In the area of machine learning (ML), data mining and natural language processing (NLP), there is a standard problem known as document clustering. Since clustering can discover knowledge from the document corpora it is given high importance. Moreover, in many applications, clustering could be used as a pre-processing phase for optimization of further ML processes. Thus document clustering assumes greater significance in different knowledge discovery applications.

Clustering has potential to generate and organize documents into some meaningful clusters helping in discovering latent structure in presences of unlabelled document corpora [1]. Latent Dirichlet allocation (LDA) is a widely used generative process model to deal with textual corpora in a systematic fashion. It is the model that helps in processing documents and underlying text for solving different problems [2]. Often LDA is used along with clustering techniques like K-means clustering. Another important aspect associated with processing documents is known as dimensionality reduction which improves processing performance.

There are many existing methods that focused on document clustering problem. Most of the solutions,

*Author for correspondence

as explored in [3–6] to mention few, are based on LDA model. LDA with topic modelling is also widely used in the existing research as found in [7−10]. LDA based document clustering is used for solving problems in different domains.

Raghuveer [4] investigated on LDA based methodology for document clustering. They considered corpora as legal documents as the clustering of legal documents has significant important in judiciary domain. Their observations reveal that LDA is the model that has potential to process documents efficiently. Ning et al. [11] opined that document clustering has potential to discover knowledge that could solve specific problems. Then explored an improved form of LDA besides clustering process that could lead to solving on-site assembly problems.

Raja and Pushpa [12] proposed a methodology based on LDA and clustering. Their method was designed to deal with mobile multimedia application data. It was meant for discovering useful recommendations with optimized clustering process based on the given historical data. From the literature, many insights are witnessed. First, LDA is the process model suitable for document clustering. Second, LDA along with topic modelling has its utility in improving performance. Third, it is inferred that there is need for novel approaches including hybrid models towards optimizing document clustering process. Challenges in the previous works found in the literature are as described here. In [1], the LDA based approach lacks procedure to identify similar document identification (SDI). Moreover, without advanced neural network usage, performance is deteriorated. The LDA models used in [2−4] could improve clustering performance. However, their modus operandi need more intelligent approach in leaning based scalable means of document clustering. The proposed methodology addresses these issues by considering a hybrid approach that considers SDI prior to learning based clustering of documents.

Our contributions in this paper are as follows.

1. A framework known as hybrid approach for dynamic document clustering (HADDC) was proposed for improving efficiency in document clustering.
2. Our framework is realized by defining two algorithms that work in tandem with each other towards dynamic document clustering. Similar SDI is an algorithm which exploits lexical dictionary WordNet and similarity measure to find similar documents effectively is proposed. Another algorithm known as topic modelling for efficient and dynamic document clustering (TM-EDDC) is proposed for document clustering process model. It is a dynamic process model based on LDA which is capable of clustering documents incrementally when new documents arrive.
3. Empirical study is made with News Groups dataset is used to evaluate the proposed framework and underlying algorithms. The experimental results revealed the enhanced utility and efficiency of the proposed framework.

The remainder of the paper is organized as follows: Section 2 provides a review of the literature concerning various approaches involving LDA and other methods for document clustering. In section 3, the methodology used was presented to enhance the efficiency of document clustering. Section 4 presents the results of the empirical study. Section 5 discusses the results in detail. Finally, in section 6, conclusions have been drawn.

2.Related work

This section reviews literature on various methods pertaining to document clustering. Syed et al. [13] opined that deep learning models do have their ability to solve many real world problems. Their empirical study made with reinforcement learning could prove the utility of such models. Bui et al. [1] explored LDA model along with K-means algorithm that could improve document clustering performance. Their method involves a distance measure that is probability based. Their results showed the importance of a generative process model for systematic analysis and clustering of document corpora. Han [2] investigated on LDA topic model for analysing research documents from 1996 to 2019. Their topic model is found to have improved efficiency towards clustering when compared with its traditional LDA counterpart. Montenegro et al. [3] exploited LDA model and used it for topic modelling. Their research and analysis focused on document clustering with refined modelling. Their empirical study showed the significance of using topic modelling. Raghuveer [4] investigated on LDA based methodology for document clustering. They considered corpora as legal documents as the clustering of legal documents has significant important in judiciary domain. Their observations reveal that LDA is the model that has potential to process documents efficiently.

Tresnasari et al. [5] combined LDA model and topic modelling process to process documents linked to social child case. Their research found that LDA for topic modelling has its utility in making document clusters. Duan et al. [6] proposed a topic modelling method based on Bayesian progressive approach. Their methodology is meant for processing large textual data to arrive at actionable knowledge. Crain et al. [7] recognized the significance of dimensionality reduction in processing document corpora. Towards this end, they proposed a methodology for achieving topic modelling based on LDA and dimensionality reduction. Their research encompasses different aspects of LDA and latent semantic indexing. Jui- Yeh et al. [8] investigated on the usage of LDA for specific set of documents that are related to records reflecting conversational dialog among people. Their work has merits in terms of revealing the importance of process model to deal with such documents. Sharaff and Nagwani [10] studied the process of email contents as text corpora. Their research has investigations into LDA model along with clustering that exploits "non-negative matrix factorization". Their investigation has brought significant improvement in clustering of email contents and email thread identification.

Shafiei and Milios [14] thought about importance of co-clustering when processing document corpora. They opined that co-clustering has merits in discovering more useful knowledge from documents. Their research on co-clustering was based on the proven LDA model. Ning et al. [11] opined that document clustering has potential to discover knowledge that could solve specific problems. Then explored an improved form of LDA besides clustering process that could lead to solving on-site assembly problems. Andrzejewski and Zhu [9] investigated on the significance of semi-supervised learning process. Their methodology for document clustering involves LDA along with a specially designed topic-in-set knowhow that could improve clustering performance. Curiskis et al. [15] explored Reddit and Twitter generated text corpora for document clustering. They revealed the importance of processing social media data. Their methodology has topic modelling and document clustering two deal with the data from two virtual platforms. Hong et al. [16] investigated on JCDL'11 dataset to analyse the text corpora for discovering events. They exploited spatial variant of LDA for their investigation. It is observed that spatial domain in the document processing has its role to play in discovering events. Jelodar et al. [17] exploited topic modeling and LDA to deal with textual corpora. Thier investigation has revealed potential applications and models for textual data processing. Liu et al. [18] proposed a novel approach for document clustering. Their method is interactive in nature. Therefore, it is named as interactive LDA. This model was designed by them to bring about quality in topic modelling. Wang et al. [19] did their research on document classification using semi-supervised learning phenomenon. Their usage of LDA is also based on this kind of learning based approach. Their research has resulted in efficient classification of documents through the underlying learning based method. Tang et al. [20] explored object oriented (OO) approach in clustering of image content using a variant of LDA known as multiscale LDA. Their analysis was on the satellite images pertaining to very high resolution (VHR). Saif et al. [21] investigated on the optimization of LDA based approach in the processing of documents. They proposed a method to reduce vectors reflecting semantic representation using LDA model. Thus it could optimize performance in the clustering process.

Bird et al. [22] opined that LDA has its potential to bring about systematic approach in dealing with given data corpora. They used it for analysing data related to software and found its utility. Their research has revealed the art and science of discovering knowledge from software data using LDA. Raja and Pushpa [12] proposed a methodology based on LDA and clustering. Their method was designed to deal with mobile multimedia application data. It was meant for discovering useful recommendations with optimized clustering process based on the given historical data. Abinaya and Winster [23] explored the importance of a technique known as "named entity recognition" and combined it with LDA. Their methodology is meant for discovering events from the social media text corpora. With their hybrid approach, they could achieve better performance in event detection when compared with existing methods. Lienou et al. [24] investigated on satellite imagery that plays crucial role in different real world applications. Their focus was on making annotations on such images so as to use them in different applications. In the process, their methodology is designed based on LDA model. Other research contributions include usage of "Partially Collapsed Gibbs Sampling" along with LDA [25] and multi-model classification of cardiology data using LDA [26].

Ma et al. [27] proposed a methodology by combining topic modelling and semantic analysis for improving clustering performance. Their approach is found similar to that of generative process model. Lossio-ventura et al. [28] investigated on clustering of documents pertaining to healthcare domain using topic modelling and clustering methods. The underlying documents include emails and tweets. Rani and Kumar [29] explored topic modelling and its significance in the science and engineering domain. Thirumoorthy and Muneeswaran [30] incorporated Jaya optimization method towards realizing a hybrid methodology for document clustering. Murshed et al. [31] proposed a method known as short text topic modelling which is designed to reap benefits of clustering social media data. Their work is associated with big data analytics. Pathak et al. [32] proposed a deep learning based approach for topic modelling towards efficient sentiment analysis. They considered social media data towards opinion mining. Khan et al. [33] focused on food recommendations based on an ensemble topic modelling. Their work was context-aware and personalized recommendations addressing complexity of such systems. Shaik et al. [34] explored the usage of NLP in different applications including education feedback analysis. Their research includes observations on trends and challenges involving the adoption of NLP techniques. Hindistan and Yetkin [35] studied differential privacy and generative adversarial network techniques for processing data in IoT use cases. Their research focused on processing of data with preservation of privacy. Curiac and Micea [36] investigated on social media data to identify hot topics on information security. They proposed an LDA based methodology to achieve this.

Murshed et al. [37] proposed a methodology that is designed for topic modelling on the textual documents featuring short texts. It makes use of topic modelling but applies big data concepts and distributed programming paradigm. However, their approach is theoretical in nature and there is need for empirical approaches. Vayansky and Kumar [38] discussed many topic modelling approaches as part of text mining. Their work covers traditional approaches and also LDA kind of model along with its variants. However, it is yet to be improved to achieve efficiency with clustering case study. Farkhod et al. [39] explored topic modelling with generative process models for systematic processing of textual documents. Their approach led to sentiment analysis on top of topic modelling. Alamsyah et al. [40] proposed a methodology for topic modelling and sentiment analysis considering large volumes of data. Their approach has limitations in dealing with sentence level similarities while processing Twitter data. Gurcan and Cagiltay [41] considered big data and different approaches in topic modelling that are LDA based. In the process, they employed big data and software engineering approach towards analysing knowledge domains. However, its clustering approach is yet to be improved. Sundarkumar and Ravi [42] used topic modelling and machine learning (ML) approaches to achieve detection of malware in given dataset. In the process, their approach considers text mining and clustering procedures. Shahbazi and Byun [43] exploited topic modelling and deep learning approaches with proper integration to achieve topic prediction and knowledge discovery.

Miles et al. [44] discussed various approaches in topic modelling and POS-based text document clustering methods. Their research could reveal the necessity for using ML for dealing with large volumes of data. Acharya et al. [45] focused on familial classification approach using deep learning along with transfer learning to expedite the process of detecting malware with low computational cost.

Chehal et al. [46] considered pre-processing based on making clusters of documents containing reviews of e-commerce products. In their methodology, they compared various topic modelling techniques. Pathak et al. [47] proposed a deep learning architecture supporting big data processing. It was based on temporal topic modelling which is dynamic and adaptive in nature. However, their methodology needs improvement with proper clustering to reduce error rate. Mazzei and Ramjattan [48] investigated on Industry 4.0 and the ML models along with deep learning for topic modelling. Their study focused on the learning based approaches. It could find utility of the models but suggests improvements in the document clustering techniques.

From the literature, many insights are witnessed. First, LDA is the process model suitable for document clustering. Second, LDA along with topic modelling has its utility in improving performance. Third, it is inferred that there is need for novel approaches including hybrid models towards optimizing document clustering process.

3.Materials and methods

We used 20 News Groups dataset collected from [49] in our empirical study. This dataset comprises of

18000 newsgroup documents collected from 20 News Groups. The dataset has two columns such as identification (ID) and news (string). This dataset is widely used for text clustering, classification and other ML applications. It is the dataset which is widely used by researchers to experiment with ML techniques and text-based process such as text clustering and classification.

3.1Methods

The proposed methodology for document clustering is made up of two important parts. They are known as finding similar documents and topic modelling for formation of clusters. In the latter phase, LDA is exploited for topic modelling. Finding similar document is achieved using the proposed algorithm known as SDI while topic modelling and clustering is achieved using proposed algorithm named TM-EDDC. *Figure 1* shows overview of the proposed system.

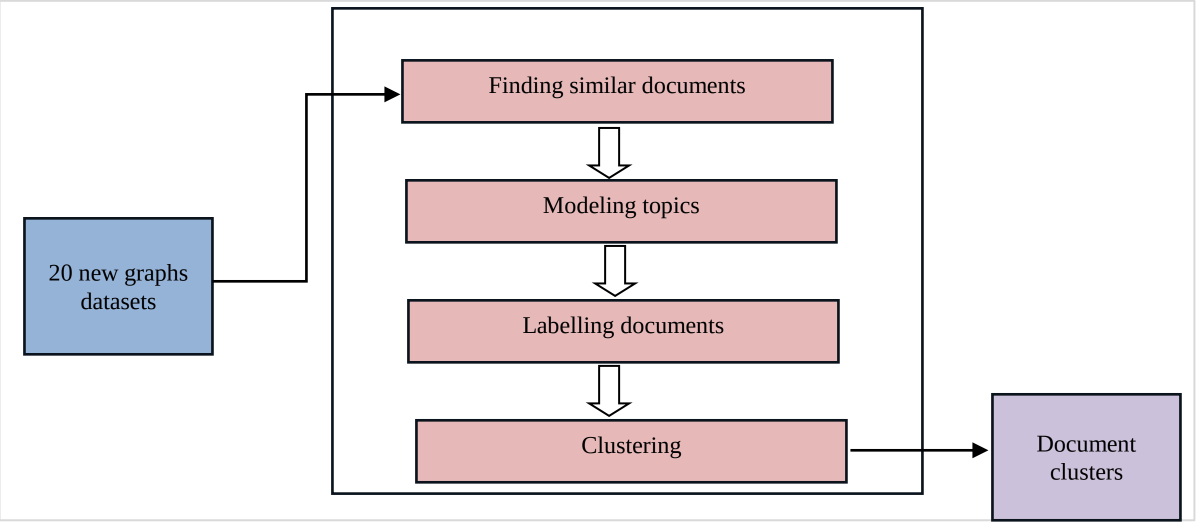

Figure 1 Shows overview of the proposed system

As presented in *Figure 1*, the proposed system takes 20 News Groups dataset as input and performs clustering of documents. In the process, it has various activities involved. They include finding similar documents, modelling topics, labelling documents and completing clustering process. More details are provided on these activities later in the paper. *Figure 2* shows the LDA model on which the proposed system is based.

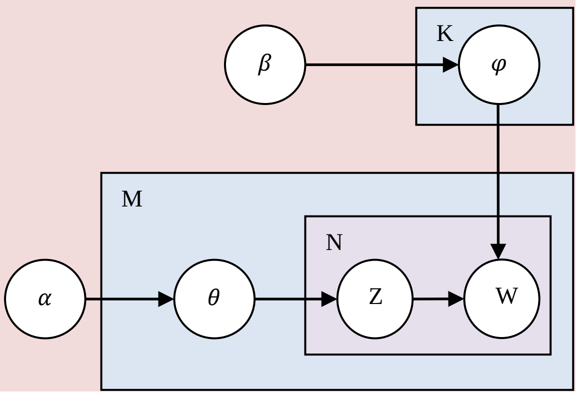

Figure 2 LDA model used for topic modelling

LDA is the generative probabilistic model that brings about systematic approach in processing documents. It has two Dirichlet prior denoted as α and β. A word in any given document is denoted as w_i while W denotes the words in entire corpus. Z refers to latent topics associated with W. The probability of z=k in given document d is expressed as in Equation 1.

$$\theta = \{\theta_{d,k}\}:\theta_{d,k} = P(z=k|d) \qquad (1)$$

In the same fashion, the probability of z=k and w=v is expressed as in Equation 2.

$$\boldsymbol{\Phi} = \{\emptyset_{k,v}\}: \emptyset_{k,v} = (\boldsymbol{w} = \boldsymbol{v}|\boldsymbol{z} = \boldsymbol{k}) \qquad (2)$$

The words present in d associated with topic k is denoted by $\Omega_{d,k}$while the word count in v in given corpus linked to k is denoted a $\psi_{k,v}$.
We can also describe the process of generating a document by the following process:

1. Choose document length N~Poisson(ξ).
2. Choose Θ from Dirichlet distribution Θ~Dir(α).
3. For each word in the document, choose w_n from a multinomial distribution p(w_n |z_n, β) where z_n represents the topic of this word.

3.2Algorithm design

The proposed framework is realized by defining two algorithms that work in tandem with each other towards dynamic document clustering. SDI is an algorithm which exploits lexical dictionary WordNet and similarity measure to find similar documents effectively is proposed. Another algorithm known as TM-EDDC is proposed for document clustering process model. It is a dynamic process model based on LDA which is capable of clustering documents incrementally when new documents arrive.

In the similar document identification (SDI) algorithm (Algorithm 1), the input comprises the 20 News Groups dataset, denoted as D, and it generates groups of similar documents, denoted as D'. The algorithm utilizes path similarity, synsets derived from lexical dictionaries such as WordNet, and computes similarity scores.

Algorithm 1: Similar document identification (SDI)
Input: 20 News Groups dataset D
Output: Groups of similar documents D'

```
Begin
Initialize similarity score map M
Finding Similar Documents
For each doc d in D
Find path similarity of d with rest of the documents in D
    Generate synsets of d using lexical dictionary
    Compute similarity score s
    Add d and s to M
End For
Finding Most Similar Documents
For each entry in M
Compare s for documents
    Group most similar documents and add to D'
End For
Return D'
End
```

As outlined in Algorithm 2, it takes the 20 News Groups dataset, denoted as D, as input and generates optimally formed clusters of documents. Initially, it performs SDI, followed by topic modeling and document labeling. Subsequently, there is an iterative process to efficiently complete the clustering procedure. To support the clustering process, a 1D Convolutional Neural Network (CNN) is employed for document labeling, and the structure of this CNN is detailed in *Table 1*. The process employed in the proposed methodology is visualized in *Figure 3.*

Algorithm 2: Topic Modelling for Efficient and Dynamic Document Clustering (TM-EDDC)
Input: 20 News Groups dataset D
Output: Final document clusters C

```
Begin
Initialize labels vector L
Running SDI Algorithm
D'←RunSDIAlgorithm(D)
Create an LDA model (random state 34, passes 25 and topics 10)
Modeling Topics
topics←FindTopics(LDA, D')
Word2Vec model for document representation
Labelling Documents
Train a CNN model
For each d' in D'
    Add label to d' using CNN knowledge model
    Add label to L
End For
Final Clustering Process
For each label x in L
Create a cluster c
    For each d' in D'
IF d' has label x Then
Add d' to c
        End If
    End For
    Add c to C
End For
Return C
```

Table 1 Structure and parameters of CNN model

Layer type	Output shape	Param#
embedding_2(Ebedding)	(None, 1582,100)	879700
conv1d_2(Conv1D)	(None, 1567, 16)	25616
max_pooling1D_2(MaxPooling1)	(None, 783,16)	θ
flatten_2(Flatten)	(None, 12528)	θ
dropout_2(Dropout)	(None, 12528)	θ
dense_2(Dense)	(None, 5)	62645

Layer type	**Output shape**	**Param#**
Total Params:967,961		
Trainable params:967,961		
Non-trainable params: θ		

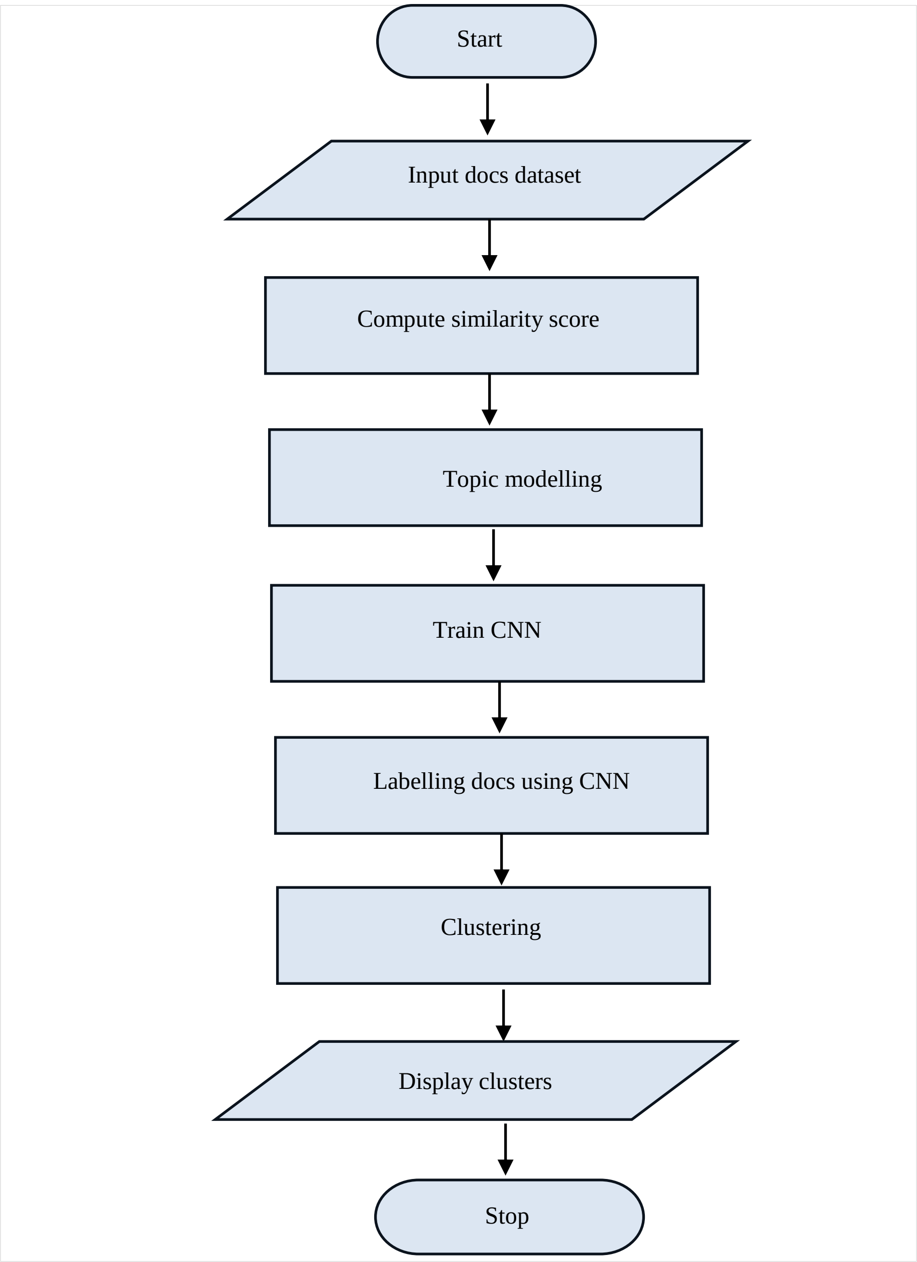

Figure 3 Flow of the proposed methodology

The flow of the methodology is summarized. It takes document dataset as input and computes similarity score with the help of synset generation and lexical dictionary. This will enable the system to find similar documents. Finding similar documents makes the further process easier. The similar documents are used to have a topic modelling which is suitable for training CNN model. Once CNN model is trained, it is capable of labelling documents so as to complete clustering process. Chi-Square method is used for feature selection and the data used for training is 75% and testing is 25%. Chi-Square method is very useful in ML applications. It finds whether a variable in independent of response variable or it has dependency. Based on this information, it determines the features that could contribute to the decision making in the given ML task.

4. Experimental results

This section present results of experiments meant for document clustering using proposed algorithms that make use of LDA and CNN for improving clustering process. CNN model is used in the process of labelling to support efficient clustering. Experimental setup includes Processor Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz, 2401 Mhz, 2 Core(s), 4 Logical Processor(s) and 8 GB RAM running Windows 11 OS. Jupytor notebook is used as IDE for empirical study. *Figure 4* shows loss dynamics of CNN model. The training loss and validation loss values are provided for different number of epochs. As the number of epochs is increasing, there is gradual increase in the validation loss but the training loss is minimal. *Figure 5* presents accuracy of CNN model.

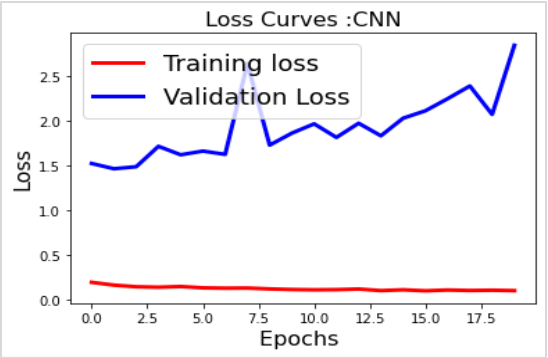

Figure 4 Loss curves of CNN model

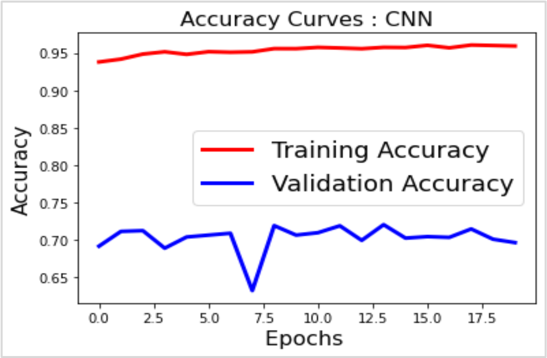

Figure 5 Accuracy curves of CNN model

The training accuracy and validation accuracy values are provided for different number of epochs. As the number of epochs is increased, there is gradual increase in training accuracy while validation accuracy is less affected. *Figure 6* shows class distribution dynamics for different topics. There are many topics reflecting class distribution and its visualization. Word2Vec model has generated the document representation model as follows.

array ([-0.16455922, -0.11885681, -0.00563807, 0.09740357, -0.04749494, 0.074497, -0.16606624, 0.05180889, -0.15255453, 0.10945424, 0.08661386, -0.07608347, 0.14448506, -0.04327876, -0.02835492, 0.04918522, -0.06202773, 0.13642298, -0.02533604, -0.07803591, -0.17755291, 0.03582903, -0.09580665, -0.05667198, -0.01266002, -0.01265419, -0.078663 , -0.03218543, 0.04942513, -0.06423364, -0.16591193, 0.06270715, -0.04431333, -0.09292921, -0.05958976, 0.08091794, 0.13686615, 0.09417998, 0.0523802 , 0.05229336, -0.0519472 , -0.13058558, 0.03128384, 0.13987564, 0.10631746, 0.11836006, -0.22727226, 0.03469679, -0.01036359, -0.02907789, -0.04447671, 0.03792319, 0.00247446, -0.02025136, 0.15072758, -0.07777885, -0.05931862, -0.03052842, -0.06955599, 0.0176603 , 0.11573175, -0.2625492, -0.09589634, -0.01159286, -0.0717172, 0.17287774, -0.06778304, 0.0628913, -0.08182127, -0.12239533, 0.01862247, -0.07850103, 0.05114951, -0.01600525, -0.0346313, 0.2860824, -0.23223224, 0.04210597, 0.10645259, -0.09770118, 0.04773833, -0.03881938, -0.10600597, 0.01120001, 0.12594979, -0.0944834 , 0.14193651, -0.0378931 , -0.0512672 , 0.19445121, 0.01936254, 0.04909763, 0.08363069, -0.16288522, -0.01282108, 0.00544338, 0.08739529, -0.04079762, -0.05747539, 0.15601933], dtype=float32)

Word cloud representation for the dataset is provided in *Figure 7*. Based on the text corpus available in the given dataset, there is generation of word cloud reflecting important words in the given corpus. *Figure 8* shows K-means clustering results with best normalized mutual information (NMI).

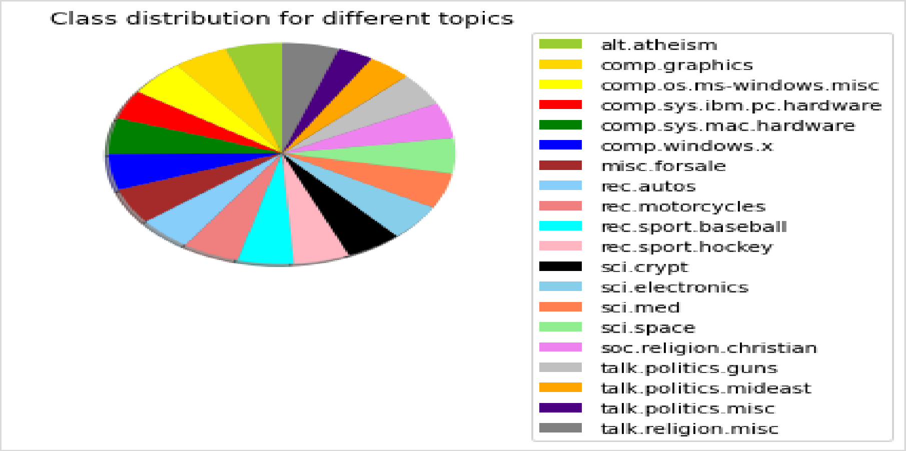

Figure 6 Shows class distribution for various topics

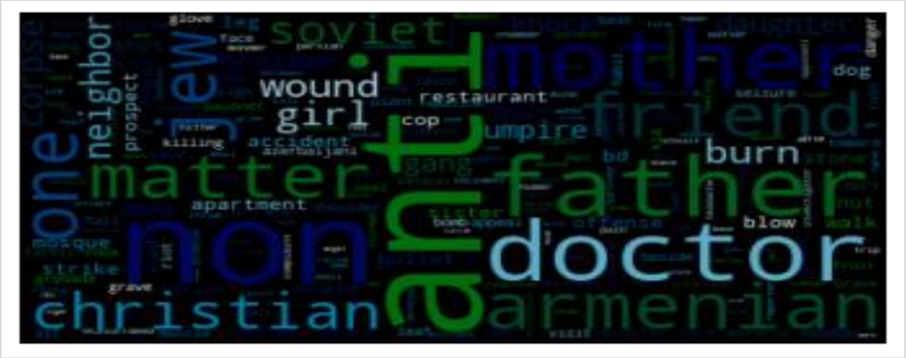

Figure 7 Word cloud representation

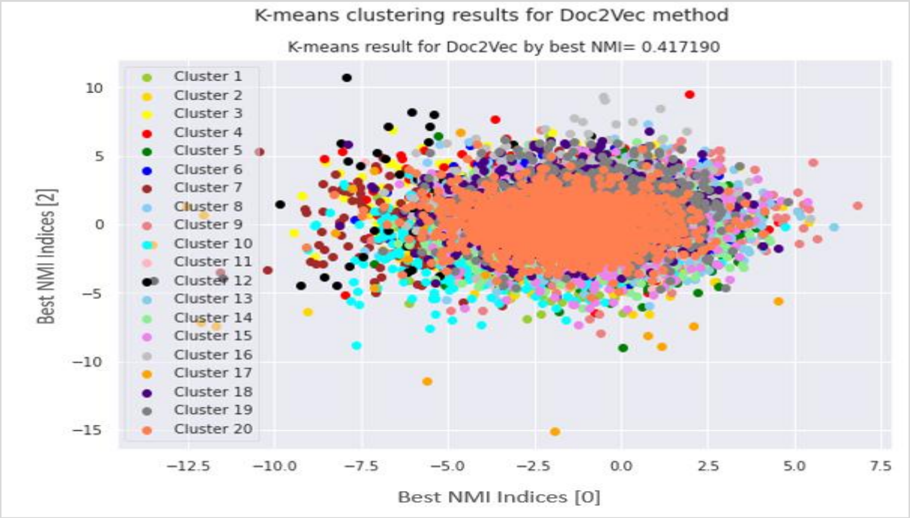

Figure 8 K-means clustering results with best NMI

The results of clustering are provided with the help of K-means and Doc2Vec method with best NMI shown as 0.417190. *Figure 9* shows 3-dimensional K-means clustering results. Different clusters are formed and the 3D K-means clustering is visualized for Doc2Vec method. *Figure 10* presents word count distribution dynamics. The document word count is provided for different number of documents along with summary statistics such as mean, median and standard deviation. *Figure 11* shows topic coherence against topic numbers. Topic coherence dynamics is provided for finding optimal number of topics.

Figure 12 shows the mean square error (MSE) dynamics reflecting performance.

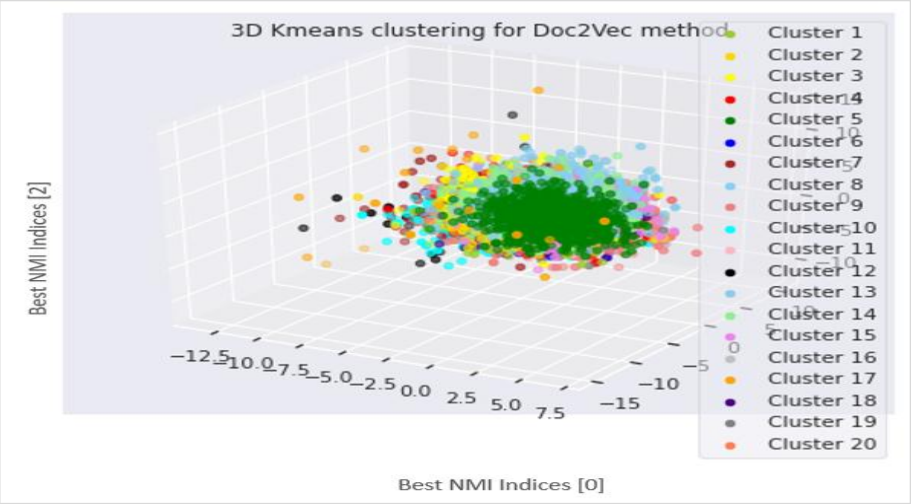

Figure 9 Shows 3D K-means clustering results

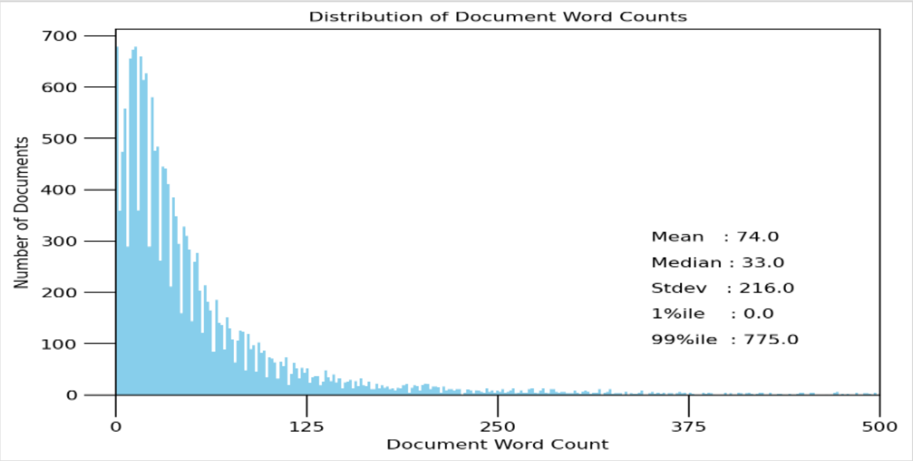

Figure 10 Word count distribution dynamics

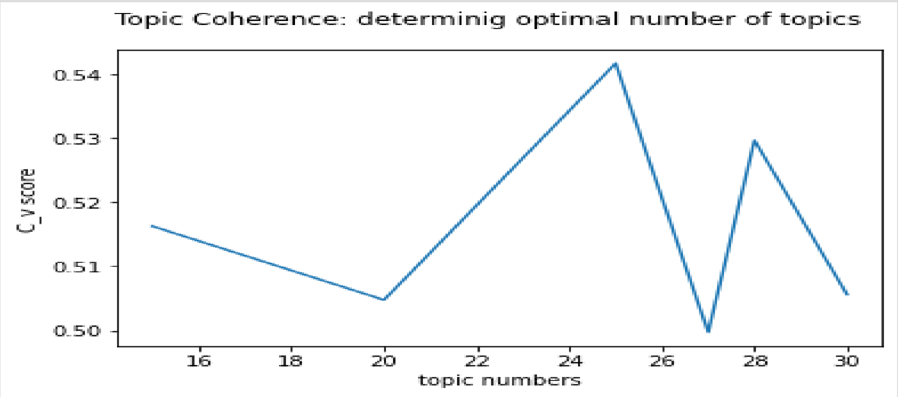

Figure 11 Shows topic coherence against topic numbers

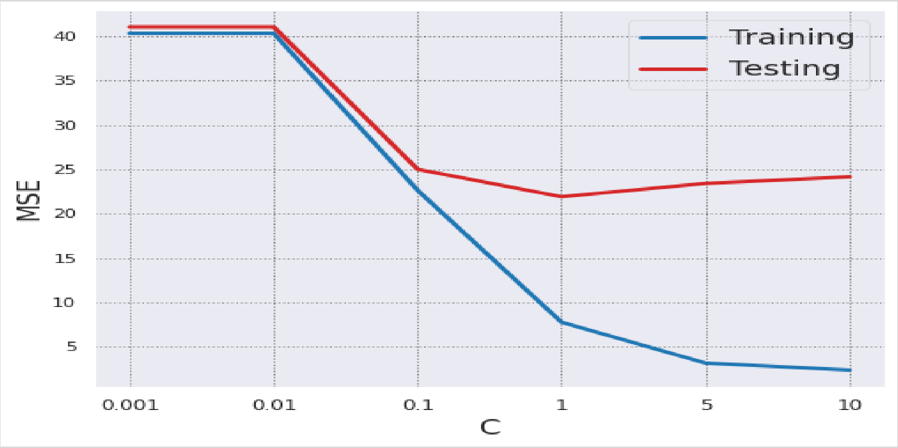

Figure 12 Shows the MSE dynamics reflecting performance

The training and testing processes are evaluated against different C values used in the empirical study. As discussed in this section, News Groups dataset is used to evaluate the proposed framework and underlying algorithms. The empirical study revealed the enhanced utility and efficiency of the proposed framework. The proposed methodology is found to be efficient for clustering of documents. As presented in *Figure 13*, performance of two ML baseline models such as support vector machine (SVM) and random forest (RF) the CNN which is used in the proposed methodology are compared in terms of accuracy. Higher accuracy refers to better performance. Experimental results revealed that each model has different level of performance. The rationale behind this is that the modus operandi used in functionality of the models is not the same. SVM model showed 83.45% accuracy while RF exhibited better performance over SVM with 85.76% accuracy. The deep learning model used in the proposed methodology is the CNN which could achieve highest performance with 96.10% accuracy.

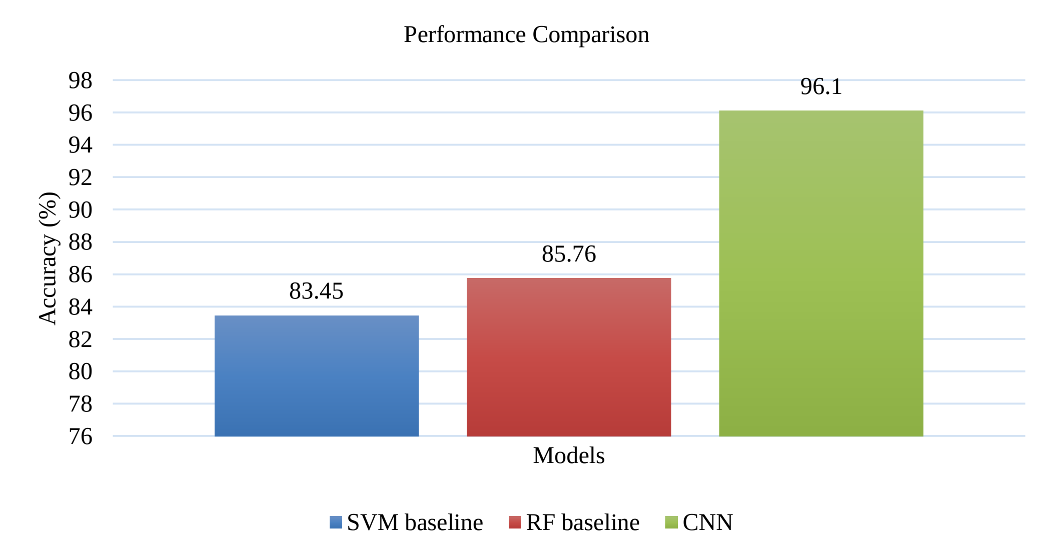

Figure 13 Performance comparison

5. Discussion

HADDC is the framework proposed in this paper. It is designed to cluster documents. The framework is based on LDA and also topic modelling. Besides it exploits CNN model for labelling documents prior to clustering them. LDA is a widely used generative process model to deal with textual corpora in a systematic fashion. It is the model that helps in processing documents and underlying text for solving different problems. By reusing LDA based approach, we could process documents systematically using generative process model approach. SDI is an algorithm which exploits lexical dictionary WordNet and similarity measure to find similar documents effectively is proposed. Another algorithm known as TM-EDDC is proposed for document clustering process model. It is a dynamic process model based on LDA which is capable of clustering documents incrementally when new documents arrive. News Groups dataset is used to evaluate the proposed framework and underlying algorithms. Experiments made with a prototype application could help in evaluating the proposed framework and underlying algorithm. In terms of accuracy, the CNN model used in the proposed methodology could do well in terms of accuracy. Word2Vec model used for feature extraction could help in processing textual documents towards clustering. Doc2Vec model used in the proposed system helped in finding similarity between documents. The training accuracy and validation accuracy values are provided for different number of epochs. As the number of epochs is increased, there is gradual increase in training accuracy while validation accuracy is less affected. The empirical study revealed the enhanced utility and efficiency of the proposed framework. The proposed methodology is found to be efficient for clustering of documents.

5.1 Limitations

The research carried out in this paper has significant limitations. The proposed framework and utility cannot be generalized with one specific dataset. There is need for evaluation of the system with datasets from different domains. It is also interesting to experiment the proposed system using datasets of different size. HADDC framework is based on LDA based methodology. However, it is to be understood that there is possibility of leveraging topic modelling further towards improving efficiency of generative process model. Another important limitation is that the proposed system takes more time to deal with clustering of long documents. At present there is no provision for taking representative portion of the document instead of processing the whole document.

A complete list of abbreviations is shown in *Appendix I.*

6. Conclusion and future work

In this paper, a framework known as HADDC was proposed. This framework was realized by defining two algorithms that worked in tandem with each other towards dynamic document clustering. SDI was an algorithm that exploited the lexical dictionary WordNet and similarity measures to effectively identify similar documents. Another algorithm known as TM-EDDC was proposed for the document clustering process model. It was a dynamic process model based on LDA, capable of clustering documents incrementally as new documents arrived. The News Groups dataset was used to evaluate the proposed framework and underlying algorithms. The empirical study revealed the enhanced utility and efficiency of the proposed framework. In the future, further investigation will be conducted on different variants of generative process models to leverage performance in dynamic document clustering.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

Gugulothu Venkanna: Contributed in proposing methodology, prototype implementation and performance evaluation. **K.F Bharati**: Contributed in guiding every aspect of the research including algorithm and implementation besides dataset collection.

References

[1] Bui QV, Sayadi K, Amor SB, Bui M. Combining latent Dirichlet allocation and K-means for documents clustering: effect of probabilistic based distance measures. In intelligent information and database systems: 9th Asian conference, ACIIDS, Kanazawa, Japan, Proceedings, Part I 2017 (pp. 248-57). Springer International Publishing.

[2] Han X. Evolution of research topics in LIS between 1996 and 2019: an analysis based on latent Dirichlet allocation topic model. Scientometrics. 2020; 125(3):2561-95.

[3] Montenegro C, Ligutom III C, Orio JV, Ramacho DA. Using latent dirichlet allocation for topic modeling and document clustering of Dumaguete city twitter dataset. In proceedings of the international conference on computing and data engineering 2018 (pp. 1-5). ACM.

[4] Raghuveer K. Legal documents clustering using latent Dirichlet allocation. IAES International Journal of Artificial Intelligence 2012; 2(1):34-7.

[5] Tresnasari NA, Adji TB, Permanasari AE. Social-child-case document clustering based on topic modeling using latent Dirichlet allocation. Indonesian Journal of Computing and Cybernetics Systems (IJCCS). 2020; 14(2):179-88.

[6] Duan Z, Liu X, Su Y, Xu Y, Chen B, Zhou M. Bayesian progressive deep topic model with knowledge informed textual data coarsening process.

In international conference on machine learning 2023 (pp. 8731-46). PMLR.

[7] Crain SP, Zhou K, Yang SH, Zha H. Dimensionality reduction and topic modeling: from latent semantic indexing to latent Dirichlet allocation and beyond. Mining Text Data. 2012:129-61.

[8] Yeh JF, Lee CH, Tan YS, Yu LC. Topic model allocation of conversational dialogue records by latent Dirichlet allocation. In signal and information processing association annual summit and conference (APSIPA), Asia-Pacific 2014 (pp. 1-4). IEEE.

[9] Andrzejewski D, Zhu X. Latent Dirichlet allocation with topic-in-set knowledge. In proceedings of the NAACL HLT workshop on semi-supervised learning for natural language processing 2009 (pp. 43-8). Association for Computational Linguistics.

[10] Sharaff A, Nagwani NK. Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques. Journal of Information Science. 2016; 42(2):200-12.

[11] Ning W, Liu J, Xiong H. Knowledge discovery using an enhanced latent Dirichlet allocation-based clustering method for solving on-site assembly problems. Robotics and Computer-Integrated Manufacturing. 2022; 73:102246.

[12] Raja DK, Pushpa S. Diversifying personalized mobile multimedia application recommendations through the latent Dirichlet allocation and clustering optimization. Multimedia Tools and Applications. 2019; 78(17):24047-66.

[13] Syed AR, Yau KL, Qadir J, Mohamad H, Ramli N, Keoh SL. Route selection for multi-hop cognitive radio networks using reinforcement learning: an experimental study. IEEE Access. 2016; 4:6304-24.

[14] Shafiei MM, Milios EE. Latent Dirichlet co-clustering. In sixth international conference on data mining (ICDM'06) 2006 (pp. 542-51). IEEE.

[15] Curiskis SA, Drake B, Osborn TR, Kennedy PJ. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing & Management. 2020; 57(2):102034.

[16] Hong F, Lai C, Guo H, Shen E, Yuan X, Li S. FLDA: latent Dirichlet allocation based unsteady flow analysis. IEEE Transactions on Visualization and Computer Graphics. 2014; 20(12):2545-54.

[17] Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications. 2019; 78:15169-211.

[18] Liu Y, Du F, Sun J, Jiang Y. iLDA: an interactive latent Dirichlet allocation model to improve topic quality. Journal of Information Science. 2020; 46(1):23-40.

[19] Wang D, Thint M, Al-rubaie A. Semi-supervised latent Dirichlet allocation and its application for document classification. In IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology 2012 (pp. 306-10). IEEE.

[20] Tang H, Shen L, Qi Y, Chen Y, Shu Y, Li J, et al. A multiscale latent Dirichlet allocation model for object-oriented clustering of VHR panchromatic satellite images. IEEE Transactions on Geoscience and Remote Sensing. 2012; 51(3):1680-92.

[21] Saif A, Ab AMJ, Omar N. Reducing explicit semantic representation vectors using latent Dirichlet allocation. Knowledge-Based Systems. 2016; 100:145-59.

[22] Bird C, Menzies T, Zimmermann T. The art and science of analyzing software data. Elsevier; 2015.

[23] Abinaya G, Winster SG. Event identification in social media through latent Dirichlet allocation and named entity recognition. In proceedings of IEEE international conference on computer communication and systems ICCCS14 2014 (pp. 142-6). IEEE.

[24] Lienou M, Maitre H, Datcu M. Semantic annotation of satellite images using latent Dirichlet allocation. IEEE Geoscience and Remote Sensing Letters. 2009; 7(1):28-32.

[25] Park H, Park T, Lee YS. Partially collapsed Gibbs sampling for latent Dirichlet allocation. Expert Systems with Applications. 2019; 131:208-18.

[26] Pérez J, Pérez A, Casillas A, Gojenola K. Cardiology record multi-label classification using latent Dirichlet allocation. Computer Methods and Programs in Biomedicine. 2018; 164:111-9.

[27] Ma T, Zhou X, Liu J, Lou Z, Hua Z, Wang R. Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies. Technological Forecasting and Social Change. 2021; 173:121159.

[28] Lossio-ventura JA, Gonzales S, Morzan J, Alatrista-salas H, Hernandez-boussard T, Bian J. Evaluation of clustering and topic modeling methods over health-related tweets and emails. Artificial Intelligence in Medicine. 2021; 117:102096.

[29] Rani S, Kumar M. Topic modeling and its applications in materials science and engineering. Materials Today: Proceedings. 2021; 45:5591-6.

[30] Thirumoorthy K, Muneeswaran K. A hybrid approach for text document clustering using Jaya optimization algorithm. Expert Systems with Applications. 2021; 178:115040.

[31] Murshed BA, Abawajy J, Mallappa S, Saif MA, Al-ghuribi SM, Ghanem FA. Enhancing big social media data quality for use in short-text topic modeling. IEEE Access. 2022; 10:105328-51.

[32] Pathak AR, Pandey M, Rautaray S. Topic-level sentiment analysis of social media data using deep learning. Applied Soft Computing. 2021; 108:107440.

[33] Khan MA, Smyth B, Coyle D. Addressing the complexity of personalized, context-aware and health-aware food recommendations: an ensemble topic modelling based approach. Journal of Intelligent Information Systems. 2021; 57(2):229-69.

[34] Shaik T, Tao X, Li Y, Dann C, Mcdonald J, Redmond P, Galligan L. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. IEEE Access. 2022; 10:56720-39.

[35] Hindistan YS, Yetkin EF. A hybrid approach with GAN and DP for privacy preservation of IIoT data. IEEE Access. 2023; 11:5837-49.

[36] Curiac CD, Micea MV. Identifying hot information security topics using LDA and multivariate mann-kendall test. IEEE Access. 2023; 11:18374-84.

[37] Murshed BA, Mallappa S, Abawajy J, Saif MA, Al-ariki HD, Abdulwahab HM. Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. Artificial Intelligence Review. 2023; 56(6):5133-260.

[38] Vayansky I, Kumar SA. A review of topic modeling methods. Information Systems. 2020; 94:101582.

[39] Farkhod A, Abdusalomov A, Makhmudov F, Cho YI. LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model. Applied Sciences. 2021; 11(23):1-15.

[40] Alamsyah A, Rizkika W, Nugroho DD, Renaldi F, Saadah S. Dynamic large scale data on twitter using sentiment analysis and topic modeling. In 6th international conference on information and communication technology (ICoICT) 2018 (pp. 254-8). IEEE.

[41] Gurcan F, Cagiltay NE. Big data software engineering: analysis of knowledge domains and skill sets using LDA-based topic modeling. IEEE Access. 2019; 7:82541-52.

[42] Sundarkumar GG, Ravi V, Nwogu I, Govindaraju V. Malware detection via API calls, topic models and machine learning. In international conference on automation science and engineering (CASE) 2015 (pp. 1212-7). IEEE.

[43] Shahbazi Z, Byun YC. Topic prediction and knowledge discovery based on integrated topic modeling and deep neural networks approaches. Journal of Intelligent & Fuzzy Systems. 2021; 41(1):2441-57.

[44] Miles S, Yao L, Meng W, Black CM, Miled ZB. Comparing PSO-based clustering over contextual vector embeddings to modern topic modeling. Information Processing & Management. 2022; 59(3):1-11.

[45] Acharya S, Rawat U, Bhatnagar R. A low computational cost method for mobile malware detection using transfer learning and familial classification using topic modelling. Applied Computational Intelligence and Soft Computing. 2022; 2022:1-22.

[46] Chehal D, Gupta P, Gulati P. Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. Journal of Ambient Intelligence and Humanized Computing. 2021; 12:5055-70.

[47] Pathak AR, Pandey M, Rautaray S. Adaptive model for dynamic and temporal topic modeling from big data using deep learning architecture. International Journal of Intelligent Systems and Applications. 2019; 9(6):13-27.

[48] Mazzei D, Ramjattan R. Machine learning for industry 4.0: a systematic review using deep learning-based topic modelling. Sensors. 2022; 22(22):1-31.

[49] https://www.kaggle.com/datasets/crawford/20-newsgroups. Accessed 20 July 2023.

Gugulothu Venkanna is an Assistant Professor of Computer Science and Engineering at Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad. He earned his undergraduate degree in CSE in 2005 from Jawaharlal Nehru Technological University, Hyderabad, and completed his postgraduate studies in CSE at the National Institute of Technology, Warangal, in 2008. He is currently pursuing his Ph.D. at Jawaharlal Nehru Technological University, Anantapur, expected to be completed in 2023. His areas of interest include Data Mining, Artificial Intelligence, and Machine Learning.
Email: gugulothuvenkanna149@gmail.com

K.F Bharati is an Associate Professor in the Department of Computer Science and Engineering at JNTUA College of Engineering, Anantapuramu. She completed her undergraduate studies at Gulbarga University, Karnataka, her postgraduate studies at Visvesvaraya Technological University (VTU), Karnataka, and obtained her Ph.D. from JNTU Anantapuramu. Dr. Bharati is actively involved in research guidance. She is currently supervising 5 part-time Ph.D. scholars and 1 full-time Ph.D. scholar. Additionally, she has guided more than 60 M.Tech students and is currently mentoring 4 students. She has also supervised over 50 MCA students and is currently guiding 3 students in this program. Her research interests encompass Data Mining, Big Data, Machine Learning, and Data Analytics, with a focus on IoT research.
Email: kfbharathi@gmail.com

Appendix I

S. No.	Abbreviation	Description
1	CNN	Convolutional Neural Network
2	GAN	Generative Adversarial Network
3	HADDC	Hybrid Approach for Dynamic Document Clustering
4	LDA	Latent Dirichlet Allocation
5	MAE	Mean Absolute Error
6	ML	Machine Learning
7	MSE	Mean Square Error
8	NLP	Natural Language Processing
9	OO	Object Oriented
10	RF	Random Forest
11	SDI	Similar Document Identification
12	SVM	Support Vector Machine
13	TM-EDDC	Topic Modelling For Efficient And Dynamic Document Clustering
14	VHR	Very High Resolution