An integrated clustering and BERT framework for improved topic modeling

Lijimol George[1] · P. Sumathy[1]

Abstract Topic modelling is a machine learning technique that is extensively used in Natural Language Processing (NLP) applications to infer topics within unstructured textual data. Latent Dirichlet Allocation (LDA) is one of the most used topic modeling techniques that can automatically detect topics from a huge collection of text documents. However, the LDA-based topic models alone do not always provide promising results. Clustering is one of the effective unsupervised machine learning algorithms that are extensively used in applications including extracting information from unstructured textual data and topic modeling. A hybrid model of Bidirectional Encoder Representations from Transformers (BERT) and Latent Dirichlet Allocation (LDA) in topic modeling with clustering based on dimensionality reduction have been studied in detail. As the clustering algorithms are computationally complex, the complexity increases with the higher number of features, the PCA, t-SNE and UMAP based dimensionality reduction methods are also performed. Finally, a unified clustering-based framework using BERT and LDA is proposed as part of this study for mining a set of meaningful topics from the massive text corpora. The experiments are conducted to demonstrate the effectiveness of the cluster-informed topic modeling framework using BERT and LDA by simulating user input on benchmark datasets. The experimental results show that clustering with dimensionality reduction would help infer more coherent topics and hence this unified clustering and BERT-LDA based approach can be effectively utilized for building topic modeling applications.

Keywords Latent Dirichlet allocation (LDA) · Topic modeling · k-means clustering · Dimensionality reduction · Bidirectional encoder representations from transformers (BERT)

1 Introduction

As the online data volume is increasing exponentially, it is essential to understand the topics from the available data promptly and correctly. There is an increased demand for automatic topic modeling system due to the enormous volume of text data available in electronic form today and the limitations of human reading abilities. There are multiple techniques used to discover the topics from text, images, and videos [1]. Topic models have been extensively utilized in Topic Recognition and Tracking tasks, which helps to track, detect, and designate topics from a stream of documents or texts. This machine learning technique is widely used in NLP applications to analyze the unstructured textual data and automatically discover the abstract topics within it. There are various methods proposed to accurately identify topics [2–4] from huge text corpora.

This study is intended to model various text mining techniques to infer hidden semantic structures and hence the topics in a text corpus. Topic modeling experiments are conducted using the CORD-19 dataset. The topic modeling investigations are performed based on sentence embedding after pre-processing. A hybrid model of Bidirectional Encoder Representations from Transformers (BERT) and Latent Dirichlet Allocation (LDA) in topic modeling have been studied in detail.

✉ Lijimol George
  lijimol.george@tcs.com

  P. Sumathy
  sumathy.p@bdu.ac.in

[1] Department of Computer Science, Bharathidasan University, Tiruchirappalli 620 023, Tamil Nadu, India

2188

Int. j. inf. tecnol. (April 2023) 15(4):2187–2195

As an extended work, the applications of k-Means clustering in topic modeling have been studied. As the computational complexity of the Clustering Algorithms increases with the increased number of features, various dimensionality reduction methods are also be performed to improve the performance of the proposed system. Finally, an efficient integrated clustering and a combined BERT-LDA framework to perform automatic topic modeling is designed and implemented. The performance evaluation of the system is also carried out by computing Silhouette Score. The experimental results are found very promising, and it also unveils the opportunity for further research, precisely by applying hybrid methods to improve the efficiency of the proposed model.

The rest of the paper is organized as follows: Sect. 2 provides a quick review of related work around the topic modeling and clustering. Section 3 presents the proposed topic modeling methodology. In Sect. 4, the comprehensive experiments and results are illustrated. Lastly, Sect. 5 concludes the work and share the future research directions.

2 Review of related works

In this section, recent research findings related to topic identification including topic modeling and topic clustering are summarised. Topic modeling is an essential study area in Natural Language Processing (NLP). It aims to identify the most important topics from the documents, which help to understand about the document without reading the whole content. As the digital contents are growing rapidly and it's virtually impossible to read all the details to identify related topics. Reliable topic modeling algorithms help you to draw relevant conclusions quickly. A reliable topic modeling algorithm helps to get the relevant inferences quickly. Organizing, searching, and briefing large volume of textual information are the major tasks involved in NLP and the topic modelling can be used to resolve these problems to certain level. Some of the important techniques that are broadly used for topic modelling are detailed below.

The Vector Space Model (VSM) was widely used in the initial stage of text processing, in which a document is denoted as a vector. Latent Semantic Analysis (LSA) [5] came into practice to overcome the drawback of VSM, as it could not capture inter and intra-document statistical structure. It helps in examining relations between the set of documents and the terms it comprises by creating a set of conceptions related to the documents and terms. Latent Semantic Analysis looks at how a group of texts' terms relate to one another. Using singular value decomposition, it searches unstructured data for undiscovered connections between words and ideas (SVD). LSA can significantly reduce data when used to huge corpora, but as it uses a linear model, it

cannot account for nonlinear dependencies. Latent Dirichlet Allocation (LDA) overcomes the deficiencies of LSA [6]. LDA is the most widely used algorithm for topic modelling based on statistical analysis [7]. It aims to discover the topic belonging to a document, based on the specific words in it. Each latent topic of documents is distinguished by a probabilistic distribution around words and the word allocations of topics assign a common Dirichlet. There are multiple variations of LDA identified to get more accurate result based on different types of data, some are Correlated Topic Models, Supervised Topic Models, Nested hierarchical Dirichlet Process and Anchor-free correlated topic modeling [8–11]. Gupta et al. presented a scoring based LDA algorithm to effectively extract and conclude significant life events from social networking sites. [12]. Many research findings show that LDA with word embedding gain better performance. S. Limwattana et al. proposed Deep Word-Topic Latent Dirichlet Allocation (DWT-LDA) method for training LDA with word embedding [13]. As an alternative approach of assigning word topics to the Collapsed Gibbs Sampling process, a neural network with word embedding method is used.

BERT is a Natural Language Processing Model proposed by Google Researchers [14]. This topic modeling technique uses transformers (BERT embeddings) and class-based TF-IDF to generate dense clusters [16]. N. Peinelt et al. proposed tBERT, a topic-aware extractive and abstractive summarization approach that is built on Transformers' bidirectional encoder representations (BERTs) [15]. This method concentrated on topic mining and previously trained external knowledge to capture more accurate contextual representations. It can also effectively predict topics and construct summaries from social texts. The proposed method also uses BERT, FinBERT, RoBERTa, and DistilBERT for modeling the topics. Pre-trained models provide more accurate word and phrase representations, making it easier to understand topic in Bert-based models since stemming and lemmatization are not required.

Clustering and topic modelling combined techniques are used to improve the topic discovery process [17–19]. Recently, researchers started applying combined techniques to get more accurate topic predictions. The combined BERT-LDA model showed better result than individual LDA and BERT modeling. Zhang et al. proposed a flexible framework for combining topic models with BERT. According to their analysis, introducing LDA topics to BERT significantly enhanced performance across a variety of semantic similarity dataset prediction [20, 35]. When measured against a number of standard benchmarks, SBERT was found to significantly outperform state-of-the-art phrase embedding approaches [36, 37]. A quick glance about the topic modeling techniques used in recent research are mentioned in Table 1.

**Table 1** Summary of recent topic modeling research findings

Sl. No	Study	Data source	Methods	Intended use
1	Blei et al. 2003 [6]	Reuters -21,578	Latent Dirichlet Allocation (LDA)	Topic modeling based on statistical analysis
2	Blei et al. 2007 [11]	Pabo movie review data set	Supervised Topic Models	Statistical model of labelled documents to identify topics
3	Paisley et al. 2015 [9]	The New York Times Dataset	Nested hierarchical Dirichlet Process	Derived stochastic variational inference algorithm for the model, which enables efficient inference for massive collections of text documents
4	Fu et al. 2018 [10]	NIST Topic Detection and Tracking (TDT2), the Reuters-21,578, and the Reuters Corpus Volume 1 (RCV1) corpora	Anchor-free correlated topic modeling	Topic estimation criterion based on word-word association and its identifiability states were proven
5	Limwattana et al. 2021 [13]	Amazon customer reviews, Pantip	LDA, Word Embedding	Embedded Topic Model (ETM) that combines traditional topic models and word embeddings
6	T. Ma et al. 2022 [14]	CNN/Daily mail and XSum datasets	BERT	A topic-aware extractive and abstractive summarization model
7	N. Peinelt et al. 2020 [15]	The Microsoft Research Paraphrase dataset (MSRP), SemEval CQA	LDA, BERT	A flexible framework for combining topic models with BERT
8	Q. Xie et al. 2020 [20]	Chinese—China National Knowledge Infrastructure dataset, English- Web of Science	LDA, BERT	Topic evolution in monolingual and multilingual topic similarity relations

3 Proposed approach for topic modeling

This section presents the proposed methodology used in the research work for topic modeling along with the detailed description of the CORD-19 dataset used. There are 4 major components in the proposed integrated Clustering and BERT framework which includes feature extraction, topic modeling methods, dimensionality reduction and document clustering. In topic modeling a combined BERT-LDA model is proposed, and an improved topic modeling framework is implemented using various dimensionality reduction techniques. The reduced dimensionality results are clustered using a combined clustering framework based on k-means clustering algorithm is then implemented for mining a set of meaningful topics from the input text corpora. A high-level block diagram of the proposed approach is depicted in the Fig. 1. Each section mentioned in the block diagram are described in the following sub sections in detail.

3.1 Dataset used

For the experimentation purpose, the COVID-19 Open Research Dataset (CORD-19), which is the largest open dataset on COVID-19 research papers, is used [23]. The National Library of Medicine (NLM) at the National Institutes of Health (NIH), the Chan-Zuckerberg Initiative, Microsoft Research, IBM Research, Kaggle, and Georgetown's Centre
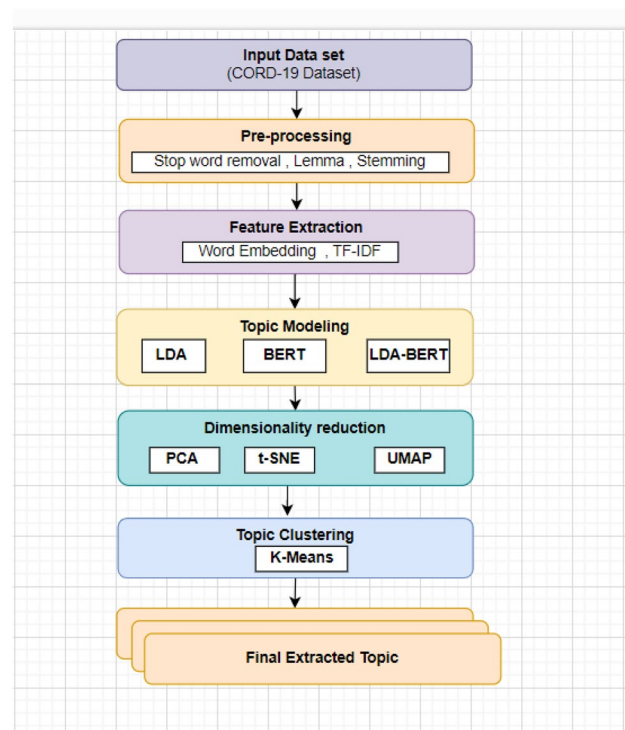
Fig. 1 Block Diagram of proposed integrated Clustering and BERT-LDA based topic modeling framework

2190

Int. j. inf. tecnol. (April 2023) 15(4):2187–2195

for Security and Emerging Technology worked together to develop the corpus of metadata and full text of COVID-19 publications and preprints, which have released daily.

Major features of the dataset used are mentioned in Table 2. CORD-19 is designed to enable the development of text mining and information retrieval techniques over its rich collection of metadata and structured full-text papers [24, 25]. This dataset is used in various text mining and corona related research in past [26, 27].

3.2 Data preprocessing

The data preprocessing techniques are used for knowledge mining which includes the transformation of raw data and making it appropriate for machine learning models. Real-world information is often inaccurate, incompatible, and/or missing in certain behaviors or trends and is probable to contain many mistakes. Preprocessing data is a proven technique for solving such problems as it prepares raw data for further processing. Initially, the input data is pipelined through Data Cleaning, Data Integration, Data Transformation, Data reduction, and Data discretization preprocessing techniques.

3.3 Feature extraction

The feature extraction techniques applied to the preprocessed document to extract the features from the document. There are various feature extraction techniques available in natural language processing. We have applied TF-IDF (Term Frequency – Inverse Document Frequency) to vectorize the document and extract the features. It is a frequently used statistical technique in information retrieval and natural language processing. It determines a term's importance within a document in relation to a group of documents. A text vectorization method converts words in a text document into importance numbers.

3.4 Topic modeling

Topic modeling techniques are applied on the vectorized documents obtained after the preprocessing and feature extraction process performed on CORD-19 dataset. The topic model output comprises of the probability of word 'w' in a given topic 't' and the probability of topic 'k' found in the given document 'N'. A combined BERT and LDA topic model is proposed. The Sentence Transformers from BERT is used for sentence embedding to provide better performance.

3.5 Dimensionality reduction

Dimensionality reduction is a significant step in text mining [21]. Dimensionality reduction enhances the performance of clustering methods by reducing dimensions so that text mining techniques process data with a reduced number of terms [28, 29]. The dimensionality reduction techniques including PCA, UMAP, and t-SNE are used in this study and performance of the system is compared based on the topic modeling output.

The PCA (Principal component analysis) is a applied to reduce the dimensionality of massive data sets, by altering a large collection of variables into a smaller one which retains the information in the massive data set. t-SNE (t-Distributed Stochastic Neighbor Embedding) is another important technique used for dimensionality reduction which is predominantly suited for high-dimensional datasets visualization.

UMAP (Uniform Manifold Approximation and Projection) is one of the promising dimensionality reduction techniques with solid mathematical foundations [22]. It is an effective tool with increased speed than t-SNE and preserve better global data structure. Also, it keeps a considerable portion of the high-dimensional structure into lower dimensional structure. The efficacy and utilization of UMAP in numerous scientific domains show how powerful the algorithm is. The UMAP technique-based topic modeling showed better result in this study.

Table 2 Features of the COVID-19 dataset

Sl. No	Features of the CORD-19: Largest open dataset on COVID-19 research papers	
1	First Release date	16 March 2020
2	Total number of documents as of Mar 2023	9,00,000 + academic articles about COVID-19, SARS-CoV-2, and related coronaviruses for 3use by the worldwide research community
3	Dataset description fields	cord_uid, publish time, title, abstract, Doi, authors, journal, full_text_file
4	Dataset download location	https://github.com/allenai/cord19
5	Data considered for current study	Published time > = Jan 2020
6	Total number of documents considered for current research	8,01,857
7	Parameters considered for the current research	Publish time, title, abstract

3.6 Clustering

It is an unsupervised method for forming similar groups of related data points. The clustering is used to group the topics so as to effectively categorize the topics within the clusters [28]. The k-Means algorithm is used for clustering. K-means unsupervised learning algorithm, groups the unlabeled dataset into different groups [30–33]. *k*-Means is a partitional based clustering method that is widely used in topic model clustering. The clustering results are evaluated using silhouette scores and the combined topic-model with reduced dimension clusters showed better results.

The algorithm of the proposed topic modelling approach based on integrated clustering and combined BERT-LDA is detailed in Algorithm -1.

Algorithm 1: Topic modelling based on integrated Clustering and combined BERT-LDA

Input   : COVID 19 Dataset
Output   : Cluster-informed categorized topics

Step 1: $\sum D_i$ is the input documents extracted from COVID-19 dataset,
            where $i = 1, \dots, n$ and n is the total number of documents
Step 2: Generate preprocessed documents $\sum D_p$ from $\sum D_i$ by implementing stop word
            removal, data cleaning, stemming and lemmatization.
Step 3: Extract the document features $\sum D_f$ from $\sum D_p$ using word
            embedding (TF-IDF)
Step 4: (i) Apply topic model algorithm LDA on $\sum D_f$ to get $\sum Tm\_LDA$.
            (ii) Apply the transformer-based (Sentence Transformers) topic model
                  algorithm BERT on $\sum D_f$ to get $\sum Tm\_BERT$.
Step 5: Apply combined BERT-LDA based topic modelling to get Tm_BERT_LDA.
Step 6: Apply dimensionality reduction on topics Tm_BERT_LDA
            based on PCA, t-SNE and UMAP algorithms to generate $\sum DR\_PCA$, $\sum DR\_$ t-
            SNE, $\sum DR\_$ UMAP
Step 7: Cluster the topics obtained from Step 6 after dimensionality reduction to get
            $\sum CT\_$ PCA, $\sum CT\_$ t- SNE, $\sum CT\_$ UMAP
Step 8: Compute Silhouette Score to evaluate the performance of the proposed model.

Abbreviations

D - Document
DP - Preprocessed documents
DF – Vectorized document
TM - Topic model
Tm_LDA – Topic model based on LDA algorithm
Tm_BERT – Topic model based on Bert algorithm
Tm_BERT_LDA – Topic model based on combined BERT-LDA
$\sum DR\_PCA$, $\sum DR\_$ t-SNE, $\sum DR\_$ UMAP – Topics obtained after Dimensionality reduction based on PCA, t-SNE and UMAP Algorithms.
$\sum CT\_$ PCA, $\sum CT\_$ t- SNE, $\sum CT\_$ UMAP - Clustered topic models obtained after dimensionality reduction based on PCA ,t-SNE and UMAP.

The following section describe the experimental setup and presents detailed analysis of results obtained.

4 Experimental results and discussion

4.1 Experimental setup

As previously stated, unsupervised clustering and transformer models are used in the study for topic modelling. The experiments are conducted to identify the topics from a set of large text corpora. The experiments have been performed on the CORD-19 dataset. The preprocessing techniques are applied on the collected subset of data. As a next step, the knowledge retrieval method that weighs a term's frequency (TF) and its inverse document frequency (IDF) are computed from the preprocessed document. The TF-IDF vectors are passed to the LDA and BERT Model to identify the topics. Then the dimensionality reduction of the embeddings is performed as the clustering algorithms poorly handle high dimensionality. The PCA, t-SNE and UMAP algorithms are used for the dimensionality reduction. The reduced dimensionality results are clustered using the *k*-means clustering algorithm and evaluated the results.

The Elbow Method is used to identify the optimum k that can be used for clustering purpose. Figure 2 shows the elbow method plot indicates the value of the cost function produced by different values of *k*.

The integrated development environment (IDE) named Spyder in open-source cross-platform is used for experimentation purposes. To perform the machine learning experiments, Scikit-learn and NLTK tools are used.

4.2 Results and discussion

The outcome of the results is evaluated based on Silhouette Score/Silhouette Coefficient. The Silhouette Coefficient is a metric applied to determine the goodness of a clustering method. The value ranges from -1 to 1. This indicated that

the clusters are perfectly separated from each other and distinguished. The Silhouette Coefficient of the topic cluster $S(j)$ is computed using the Eq. (1).

$$S(j) = \frac{(B(j) - A(j))}{MAX(B(j), A(j))} \tag{1}$$

where, A(j) is the average separation between that point and all other points belonging to the same cluster. *B(j)* is the average distance from that point to its cluster of all the points in the nearest cluster.

After computing the silhouette coefficient for each point, average it out to get the silhouette score. The Silhouette Scores obtained for different topic modelling methods with PCA, UMP, t-SNE algorithms used for dimensionality reduction are tabulated in Table 3. From the experimental results it is evident that the combined BERT-LDA model with UMAP outperforms other methods.

The Figs. 3, 4, and 5 shows the visualization of clustering result and silhouette plots for the best outcome obtained from LDA, BERT and combined BERT-LDA model as given in the Table 3. The topic clustering result and silhouette plot based on LDA is best derived with UMAP is shown in Fig. 3. Similarly, BERT technique provided better clustering results with UMAP, which is shown in Fig. 4.

Table 3 Silhouette Scores obtained for various models using different dimensionality reduction methods

Topic Models	Silhouette Scores		
	PCA	*t*-SNE	UMAP
LDA	0.33239	0.346303	0.37624
BERT	0.46211	0.369380	0.48761
BERT-LDA	0.50843	0.471261	0.51998

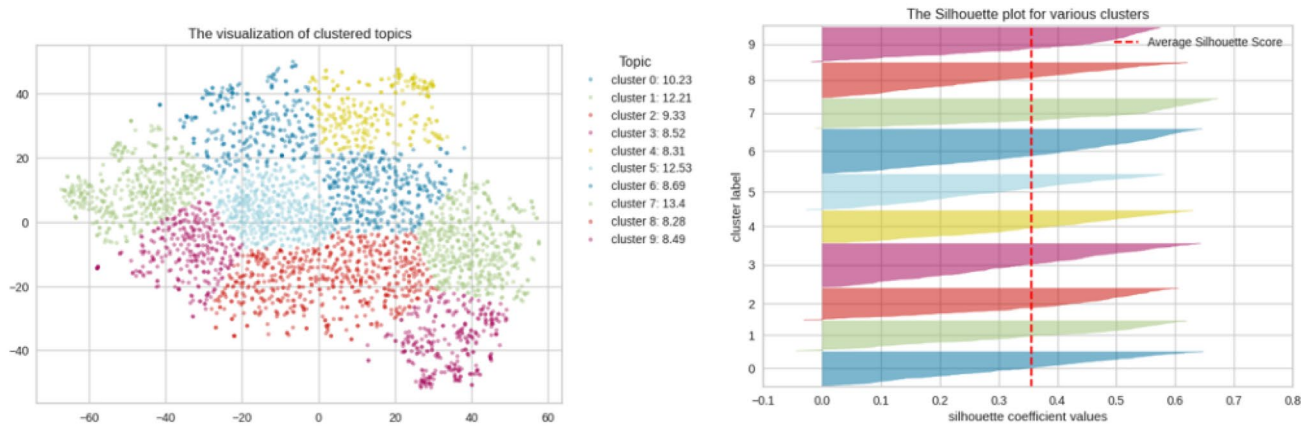**Fig. 2** Elbow method plot to identify optimum *k* value
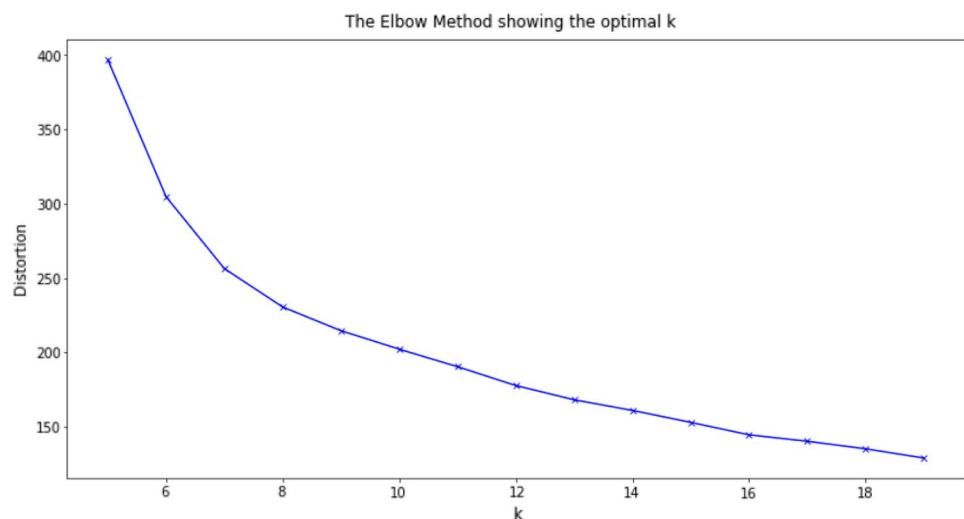
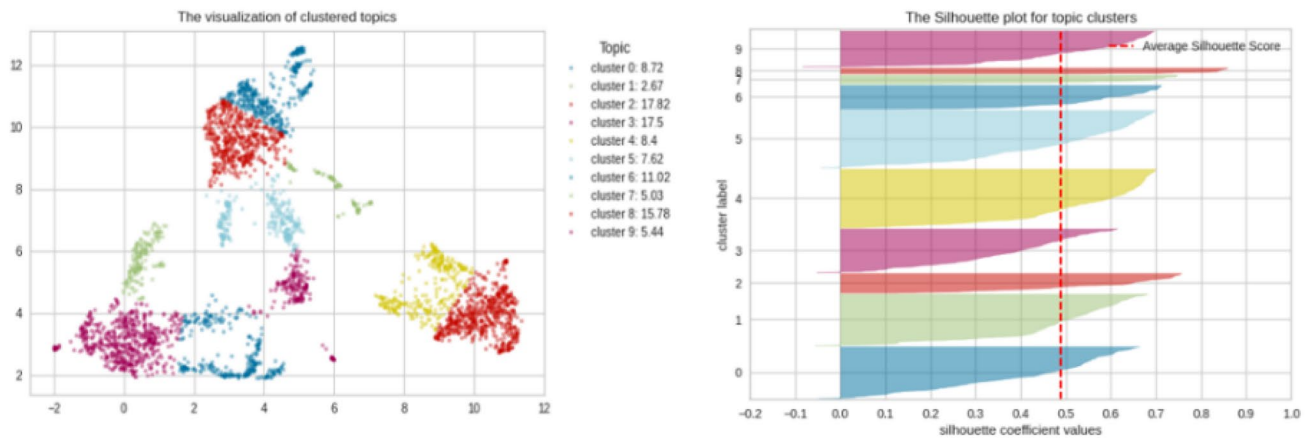**Fig. 3** Clustering result and Silhouette plot based on LDA with UMAP

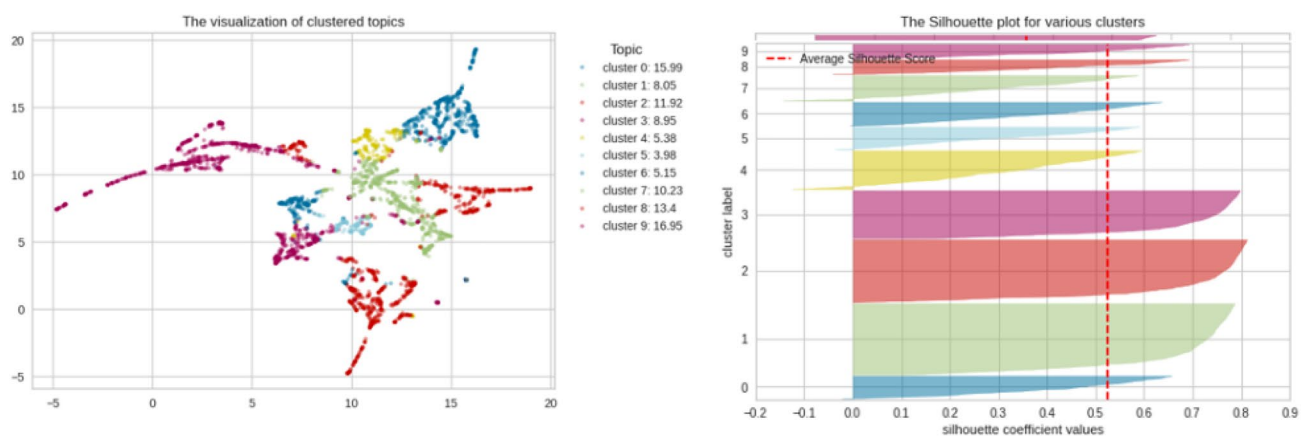**Fig. 4** Clustering result and Silhouette plot based on BERT with UMAP

Fig. 5 Clustering result and Silhouette plot based on BERT-LDA with UMAP

2194

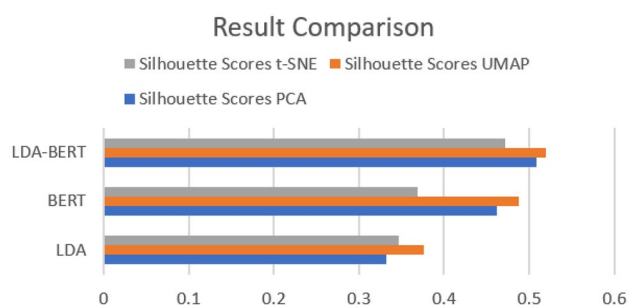Int. j. inf. tecnol. (April 2023) 15(4):2187–2195

Fig. 6 Result comparison with LDA, BERT and BERT-LDA

The combined BERT-LDA model presented best result on UMAP compared to all other techniques shown in Fig. 5.

4.3 Performance comparison of the proposed algorithm

The experimental results show that the proposed approach is performing better than the earlier algorithms reported in the similar type of research. The results are compared with already existing algorithms LDA and BERT. Figure 6 shows that the proposed integrated clustering and combined BERT-LDA framework performs better than the other methods under consideration. It may also be noted that the UMAP dimensionality reduction method is performed on top of combined BERT-LDA to get more precise context-based features for improved topic modeling.

5 Conclusion and future work

In this paper, topic modeling from text corpora has been studied in detail. A hybrid model of Bidirectional Encoder Representations (BERT) from Transformers and Latent Dirichlet Allocation (LDA) in topic modeling with unsupervised clustering have been applied. The PCA, t-SNE and UMAP-based dimensionality reduction methods are used as the clustering algorithms are computationally ineffectual with the higher number of features. The dataset CORD-19 is used after preprocessing and the hybrid model with UMAP dimensionality reduction showed comparatively better results for this input data. The reduced dimensionality results are clustered using the *k*-means clustering algorithm and evaluated the results. The *Elbow Method* is used to identify the precise number of clusters The model was developed by combining the probabilistic subject assignment vector of LDA taken from the LDA model and the sentence vectors taken from BERT model. The hybrid method helps to preserve the semantic information and created the contextual topic information.

Finally, a unified clustering-based framework using BERT- LDA is proposed and implemented for mining a set of meaningful topics from the input text corpora. The experiments are conducted to demonstrate the effectiveness of the cluster-informed topic modeling framework applying BERT- LDA. The experimental results show that clustering would help infer more coherent topics and therefore this unified clustering and BERT-based approach can be effectively utilized for developing topic modeling applications.

As a future work, the BERT variants and the Deep Neural Networks (DNNs) models can be used to infer topics. The work can be extended to verify the effectiveness of the proposed topic modeling framework with the advanced transformers and deep learning-based algorithms reported in the literature.

References

1. Qian S, Zhang T, Xu C, Shao J (2016) Multi-modal event topic model for social event analysis. IEEE Trans Multimed 18:233–246. https://doi.org/10.1109/TMM.2015.2510329
2. Rubayyi A, Khalid A (2015) A survey of topic modeling in text mining. Int J Adv Comput Sci Appl 6(1):147–153
3. Kherwa P, Bansal P (2018) Topic modeling: a comprehensive review. ICST Trans Scalable Inf Syst 7:24
4. Zhang F, Gao W, Fang Y, Zhang B (2020) Enhancing short text topic modeling with FastText embeddings. 2020 Int Conf Big Data Artif Intell Internet Things Eng (ICBAIE). https://doi.org/10.1109/ICBAIE49996.2020.00060
5. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach learn 42(1–2):177
6. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993
7. Campbell JC, Hindle A, Stroulia E (2015) Latent Dirichlet allocation: extracting topics from software engineering data. Art Sci Anal Softw Data 3:139–159

8. Blei DM, Lafferty JD (2005) Correlated topic models, advances in neural information processing systems 18 (NIPS 2005). NIPS Press, pp 147–154
9. Paisley J, Wang C, Blei DM, Jordan MI (2015) Nested hierarchical Dirichlet processes. IEEE Trans Pattern Anal Mach Intell (PAMI) 37(2):256–270
10. Fu X, Huang K, Sidiropoulos ND, Shi Q, Hong M (2018) Anchor-free, correlated topic modeling. IEEE Trans Pattern Anal Mach Intell 41(5):1056–1071
11. Blei D, McAuliffe J (2007) Supervised topic models. Advances in neural information processing systems. Springer, pp 121–128
12. Gupta M, Gupta P (2019) Research and implementation of event extraction from twitter using LDA and scoring function. Int J Inf Tecnol 11:365–371. https://doi.org/10.1007/s41870-018-0206-0
13. Limwattana S, Prom-on S (2021) Topic modeling enhancement using word embeddings, 2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–5. https://doi.org/10.1109/JCSSE53117.2021.9493816.
14. Ma T, Pan Q, Rong H, Qian Y, Tian Y, Al-Nabhan N (2022) T-BERTSum: topic-aware text summarization based on BERT. IEEE Trans Comput Soc Syst 9(3):879–890. https://doi.org/10.1109/TCSS.2021.3088506
15. Peinelt N, Nguyen D, Liakata M (2020) tBERT: topic models and BERT joining forces for semantic similarity detection. BERT, pp 7047–7055
16. MGrootendorst2020BERTopic: leveraging BERT and c-TF-IDF to create easily interpretable topicsZenodo:10.5281/zenodo.4381785Grootendorst M (2020) BERTopic: leveraging BERT and c-TF-IDF to create easily interpretable topics. Zenodo: https://doi.org/10.5281/zenodo.4381785
17. Seifollahi S, Piccardi M, Jolfaei A (2021) An embedding-based topic model for document classification. ACM Trans Asian Low-Resour Lang Inf Process 20:1–13. https://doi.org/10.1145/3431728
18. Sharaff A, Jain M, Modugula G (2022) Feature based cluster ranking approach for single document summarization. Int J Inf Tecnol 14:2057–2065. https://doi.org/10.1007/s41870-021-00853-1
19. Deshpande AR, Lobo L (2013) Text summarization using clustering technique. Int J Eng Trends Technol 4(8):3348–3351
20. Xie Q, Zhang X, Ding Y, Song M (2020) Monolingual and multilingual topic analysis using LDA and BERT embeddings. J Informetr 14(3):101055
21. Lacoste-Julien S, Sha F, Jordan MI (2008) DiscLDA: discriminative learning for dimensionality reduction and classification. Advances in neural information processing systems. Springer, pp 897–904
22. McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. J Open Source Softw 3:861
23. Lu Wang L, Lo K, Chandrasekhar Y et al (2020) CORD-19: the covid-19 open research dataset. ArXiv [Preprint]. arXiv:2004.10706v2. PMID: 32510522; PMCID: PMC7251955.
24. Wang X, Song X, Li B, Guan Y, Han J (2020) Comprehensive named entity recognition on cord-19 with distant or weak supervision. arXiv:2003.12218v5 [cs.CL]
25. Jain R, Bellaney B, Jangid P (2021) Information extraction from CORD-19 using hierarchical clustering and word bank, 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1–5. https://doi.org/10.1109/ICCCNT51525.2021.9580126.
26. Dutta P, Shah N, Saha S (2021) A multi-objective optimization-based clustering approach for CORD-19 scholarly articles, 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, pp. 1393–1398. https://doi.org/10.1109/SMC52423.2021.9658719
27. Jeyasree SKM, Vijayasree G, Geetha R (2022) Implementation of structural topic modelling for abstract mining from CORD – 19. Proceedings of the 13th international conference on soft computing and pattern recognition (SoCPaR 2021). SoCPaR 2021. Lecture notes in networks and systems, vol 417. Springer, Cham, pp 8–10
28. Otoo EJ, Shoshani A, Hwang S (2001) Clustering high dimensional massive scientific datasets. J Intell Inf Syst 17:147–168
29. Lossio-Ventura JA, Morzan J, Alatrista-Salas H, Hernandez-Boussard T, Bian J (2019) Clustering and topic modeling over tweets: a comparison over a health dataset, 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1544–1547
30. Kumar KM, Reddy AR (2017) An efficient k-means clustering filtering algorithm using density based initial cluster centers. Inf Sci 418:286–301
31. Lei Z, Wu LD, Lei Lei, Huang YY (2006) Incremental K-means method base on initialisation of cluster centers and its application in news event detection. J China Soc Sci Tech Inf 25(3):289–295
32. Alhawarat M, Hegazi M (2018) Revisiting K-means and topic modeling a comparison study to cluster Arabic documents. IEEE Access 6:42740–42749
33. Rashid J, Shah SMA, Irtaza A (2020) An efficient topic modeling approach for text mining and information retrieval through k-means clustering. Mehran Univ Res J Eng Technol 39(1):213–222. https://doi.org/10.22581/muet1982.2001.20
34. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410
35. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics, 2019, [online] Available: https://arxiv.org/abs/1908.10084. https://doi.org/10.18653/v1/D19-1410
36. Yin B, Zhao M, Guo L, Qiao L (2023) Sentence-BERT and k-means based clustering technology for scientific and technical literature, *2023 15th International Conference on Computer Research and Development (ICCRD)*, Hangzhou, China, pp. 15–20. https://doi.org/10.1109/ICCRD56364.2023.10080830.
37. Suryadjaja PS, Mandala R (2021) Improving the performance of the extractive text summarization by a novel topic modeling and sentence embedding technique using SBERT," 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Bandung, Indonesia, pp. 1–6. https://doi.org/10.1109/ICAICTA53211.2021.9640295.