Int. j. inf. tecnol. (January 2023) 15(1):435–445
https://doi.org/10.1007/s41870-022-01123-4

ORIGINAL RESEARCH

Extracting information and inferences from a large text corpus

Sandhya Avasthi[1] · **Ritu Chauhan**[2] · **Debi Prasanna Acharjya**[3]

Received: 22 March 2022 / Accepted: 14 October 2022 / Published online: 20 November 2022

Abstract The usage of various software applications has grown tremendously due to the onset of Industry 4.0, giving rise to the accumulation of all forms of data. The scientific, biological, and social media text collections demand efficient machine learning methods for data interpretability, which organizations need in decision-making of all sorts. The topic models can be applied in text mining of biomedical articles, scientific articles, Twitter data, and blog posts. This paper analyzes and provides a comparison of the performance of Latent Dirichlet Allocation (LDA), Dynamic Topic Model (DTM), and Embedded Topic Model (ETM) techniques. An incremental topic model with word embedding (ITMWE) is proposed that processes large text data in an incremental environment and extracts latent topics that best describe the document collections. Experiments in both offline and online settings on large real-world document collections such as CORD-19, NIPS papers, and Tweet datasets show that, while LDA and DTM is a good model for discovering word-level topics, ITMWE discovers better document-level topic groups more efficiently in a dynamic environment, which is crucial in text mining applications.

Keywords Topic model · Topic embedding · Embedded topic model · Scientific documents · Twitter data · Probabilistic machine learning

✉ Ritu Chauhan
rituchauha@gmail.com

[1] Amity University, Noida, India

[2] Centre of Computational Biology and Bioinformatics, Amity University, Noida, India

[3] Vellore Institute of Technology, Vellore, India

1 Introduction

The Internet has been embraced by most of the world and the world is moving toward Industry 4.0 which will revolutionize the way the industry work. The real-time data transmission of tools and applications through the internet has been seen as one of the major parameters to measure performance. The Internet provides the ability to collect and share data for understanding the efficiency and deficiency of its owners and developers. These data can be scientific, biological, operational, and social media by nature which shows the diversity of different datasets. Among various sources contributing to text, information is news headlines, tweets, social media posts, blog posts, user comments, news articles, scientific articles, etc. To meet the objectives, efficient machine learning methods, and algorithmic models are required for accurate data interpretability. Some of the texts mainly tweets, are considered short texts and, on the other hand, news and scientific articles are long texts. Analysis of both short and long texts is equally important as both are ubiquitous. Among the countless methods, theories, and applications in the text mining field, such as document clustering, text classification, information extraction, named entity recognition, text analytics, and so on, methods for detecting inherent themes and semantic structure in large-scale text collection have attracted the attention of both statisticians, analysts, and academicians [1–3]. The other well-known approach of this kind is topic modeling, typically determined using a probabilistic model called the Latent Dirichlet Allocation (LDA). Topic Modeling (TM) techniques discover semantic themes and perform statistical analysis from collections of large-scale text documents [4]. Mostly performed without human intervention, TM is an unsupervised machine learning approach. The topic modeling finds topics distributed over documents and word distribution over topics [5].

Many topic modeling algorithms give thematic topics from text corpus successfully if the text is long, but topic quality reduces in the case of short texts. Each topic can be represented by the top ten most probable words. Many application areas for topic models are marketing, law agency, political science, forensics, and digital libraries. Some other important applications are information retrieval, computational biology, recommendation systems, and computer vision. The document collections from the application areas are huge so organizing and analyzing such massive collections of documents are quite a tedious task. In recent times many variants of LDA are being used e.g., the Topic model to capture correlations between the topics, to classify documents, to represent multilingual documents, and to analyze the evolution of documents over time [6–8]. In the example, (device, hardware, software, RAM, keyboard, price, configuration, system, windows, office), the top ten words are chosen. Based on the words one can say that the most appropriate topic label is "computer System". In the given example, since words are coherent so choosing a suitable label for the topic is not difficult. The words in topics generated by topic models are not as coherent as they should be. While performing topic modeling on a large text corpus it is always good to have a good representation of the topics or labels for each topic [9, 10]. The interpretation of the topic becomes easy with the help of suitable labels and so the techniques for automatic labeling the topics are gaining popularity. In Fig. 1, the example of groups of the top ten words and their topic is mentioned.

A lot of attention was given to methods like topic modeling and information extraction of textual corpora implemented to extract thematic structure and useful patterns from large text collections. Summarizing long publications like news, essays, and books is an effective use of TM. On the other hand, as microblogs like Twitter grew in popularity, so did the importance of being able to analyze brief texts. The traditional methods such as Latent Dirichlet and Probabilistic Latent Semantic Analysis have proved to be useful in finding underlying themes or topics from the corpus, but these models are not suitable in changing environments where document collection size keeps on increasing due to real-time updates in databases. These non-parametric models are hard in computing posterior distributions and inference over the topics. The LDA model and many of its extended versions assume "bag-of-words" to represent documents, this assumption ignores word order and fails to capture semantic regularities in corpora [11]. Additionally, the traditional methods work at the document level and use global context information [12], which may not be useful or semantically coherent.

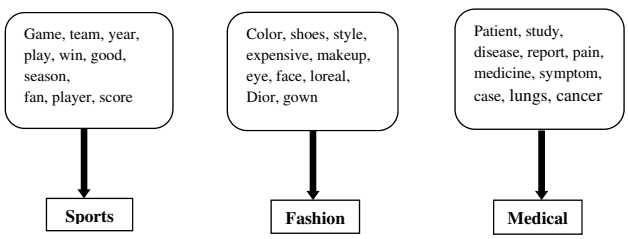

Fig. 1 Three groups of ten words each and the topic representing the group

This contribution in this research paper is given as follows:

- A topic modeling framework is proposed for incremental data, which analyzes massive document collections and extracts the topics.
- An efficient topic model for extracting the topics from short and long texts in an incremental setting is proposed.
- A comparative analysis with LDA, DTM, and ETM methods using three different datasets is performed.

The proposed Incremental Topic Model with Word Embedding (ITMWE) expands Embedded Topic Model (ETM) while integrating with the Dynamic topic model and LDA. ITMWE handles large document collections over time and scales very well by updating the topics with additional documents. The proposed model incorporates the benefits of Word embedding and the Dynamic Topic Model to give semantic structure between words and discover topics to represent document collections. The ITMWE models the similarity between words directly through the generative process. The goal in current research is to build an Incremental Topic Model with word embedding that can leverage word similarity in incremental settings. Many of the previous works are based on incremental document collection and a Dynamic Topic Model that evolves [13, 14]. The proposed model builds a topic model based on features through a dynamic topic model and word embeddings in an incremental setting. The model can identify meaningful topics by maintaining latent topics incrementally which makes it efficient in terms of time complexity. Figure 2 describes various steps in preprocessing large-scale text databases and other processes in topic modeling of the text corpora.

There are five sections in this paper; Sect. 1 introduces topic modeling and application areas. Section 2 examines related research on different topic models and their effectiveness in discovering relevant subjects. The backdrop of the topic model is discussed in Sect. 4, and the suggested topic modeling technique and assessment parameters are presented in Sect. 5. Section 6 presents all of the findings and comments, and Sect. 7 wraps up the study.

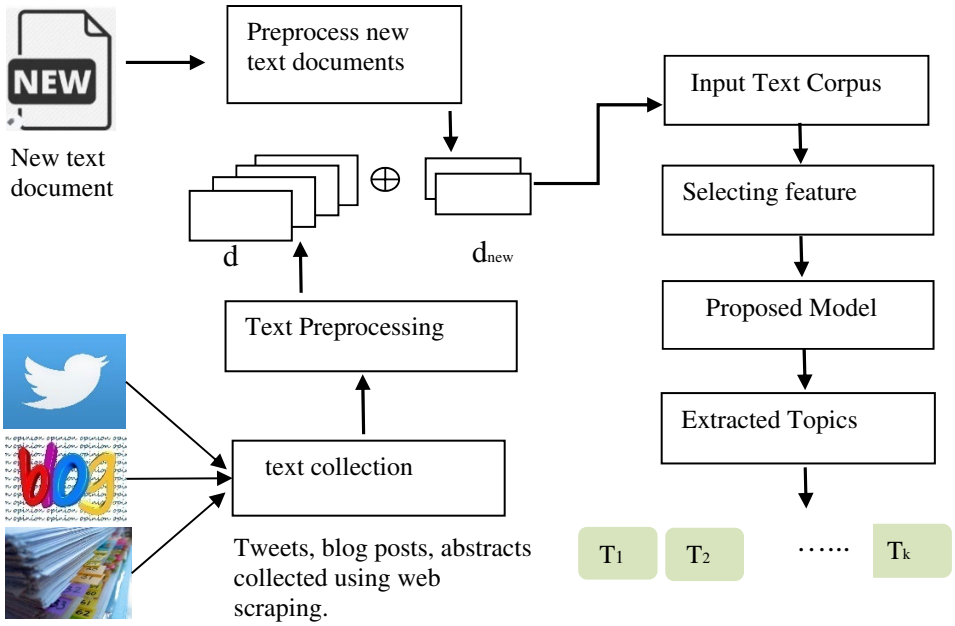

Fig. 2 Proposed Topic modeling framework to extract T1, T2…Tk topics

2 Related works

The review of the literature started with exploring terms and topics related to machine learning and text mining. The review of literature comprises more than 260 papers from online databases of Springer, IEEE Explore, Pubmed, and ACM. The main search keywords were "machine learning", "unsupervised machine learning", "text mining", "information extraction", "text corpus processing", "topic modeling", "word vectors","n-gram model","latent Dirichlet allocation" and "topic coherence". All the articles were downloaded. An initial screening of the papers eliminated around 200 papers. The systematic review approach is illustrated in Fig. 3. From the initial set of 62 papers, only a few papers were selected for full-text review those have a focus on topic modeling studies and algorithms, the remaining papers were not considered due to incomplete results, nature of work, and use of the dataset. Table 1 in this section summarizes full-text review done on selected research papers on topic modeling methods.

LDA model is a static model which only discovers latent topics from text corpus that does not capture time evolution. The vocabulary from which topics are extracted is always fixed [15–17]. The paper [18] proposes a method that extracts topics capturing the evolution within topics in an organized document collection where the document is sequentially arranged. The various articles from Journal Science have analyzed that span during the hundred years using the Dynamic Topic model based on year-wise grouping. The topic discovery and analysis of such massive text data is a very tedious and time-consuming task. The common topic model like probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA) does not perform well on short texts due to the limitation of information related to word co-occurrence.

In the paper [18, 19] authors proposed, a dynamic model based on Brownian motion and identifies latent topics from a document sequence. The latent topics form a specific pattern that evolves. cDTM (continuous DTM) [19] is an extended version of discrete dynamic topic model (dTM) [18]. In the dynamic topic model, latent topics drift over time and are known as the mixed-membership bag-of-word model. The process brings new words and old words soon become obsolete The Wiener process prior is applied to achieve continuity on the topic matrices. The paper [20] proposes Gaussian processes (GP) as priors on topic matrices, that provides generalization keeping rich dynamics. The word embedding represents words in document collections and this low-dimensional way captures the semantics of the words in texts [21, 22]. Recently a lot of topic modeling variants have implemented word embedding to reduce

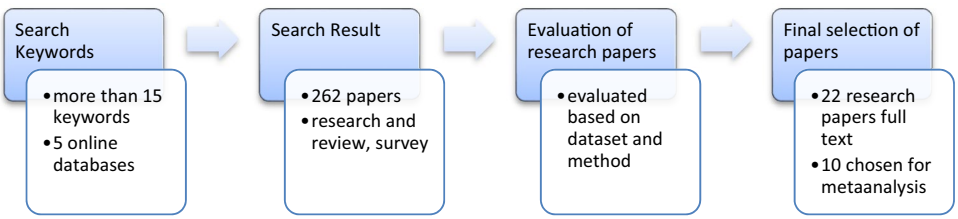

Fig. 3 Systematic Review of literature process flow

Table 1 Summary of various topic modeling techniques and their analysis

Year	Topic model	Evaluation method	Limitation	Dataset
1999	pLSA	Perplexity	Not a generative model for new documents	MED, LOB corpus
2003	Latent Dirichlet Allocation [15]	Perplexity, topic coherence	static and unable to capture correlations, fails in maintaining word-order	TREC AP corpus
2006	DTM [18]	Log-likelihood	Inference algorithm is not scalable, and performs poorly in capturing large topic dynamics	Science/JSTOR
2012	cDTM [19]	Predictive perplexity	It uses a fixed number of topics over time	TREC AP corpus
2013	Word2vec [21]	Accuracy	Inefficient in handling out of vocabulary (OOV) words	Google news corpus
2014	GloVe[23]	Accuracy	Memory usage is more due to use of matrix representation	Wikipedia dump
2017	RGloVe[25]	Precision, accuracy	Limited to Chinese corpus only	SINA news
2018	GeneralizedDTM	Predictive perplexity	Based on time-stamped data, does not include Geo-spatial information	NY corpus, NIPS
2020	ETM [1]	Topic quality	Non-dynamic model does not consider topic evolution over a period	Science, ACL corpus
2021	iVLDA [14]	Predictive perplexity	Unstable with a large number of topics (k)	NIPS, ENRON

sparsity in word representations. GloVe(Global Vector) model for word representation combines the goodness of the global matrix factorization model and local context window model [23]. GloVe model is based on statistical information and trains only the nonzero elements in the word-word cooccurrence matrix not on the entire sparse matrix of a large corpus [24]. An improvement over GloVe comes in the form of RGloVe[25] where the cosine similarity metric between entity vectors provides the measure for entity occurrences and converges easily.

In [26] authors proposed a Hyperspace Analogue to Language (HAL) word representation technique based on a matrix of the term-to-term type where rows and columns represent different words in a text corpus. The cell value in the matrix is the frequency count of the term-to-term pair. The drawback in HAL is that frequency count (number of times words co-occur together) has a large effect on similarity even though the pair does not provide any semantic relatedness. A scalable variational inference algorithm called skip-gram smoothing and skip-gram filtering [27] was proposed that was trained jointly over time. This algorithm gives a generalized embedding for historical texts which is sequential incorporating word and context vectors to drift through time. Another model that learns time-aware embeddings and solves the problem is known as the "alignment problem" [28]. Since LDA works on a fixed vocabulary, a model iVLDA [14] is proposed in the paper that follows Dirichlet process based on an incremental vocabulary. iVLDA identifies new words at the start of modeling process and adds those words into the vocabulary.

Table 2 Variable and symbols used

Symbol	Description
D	Number of documents
N	Number of words in the corpus
w	Words in corpus
w_i	ith word in the corpus
z	Topic assignment
z_i	ith topic in documents
α	Dirichlet prior
β	Dirichlet prior
θ	Probability of topic
ϕ	Probability of words given topic
σ^2, σ	Variance and standard deviation
w_{dn}	nth word in document d

3 Material and Methods

A brief review of the topic models is discussed in this section, which forms the basis of the proposed topic model and algorithm. The ITMWE incorporates three main ideas LDA, DTM, and word embedding. The variables and symbols used in this section are given in Table 2.

Let us consider D as document collections, where the vocabulary V comprises all distinct words from the documents. Let w_{dn} denote the n_{th} word in the d_{th} document.

3.1 LDA

The LDA model represents documents as multinomial distribution of topics and each topic as a distribution over many words. The earlier model Probabilistic Latent Semantic Analysis (pLSA) generates topics where different parameters are considered using documents, the limitation is overfitting, but LDA overcomes pLSA limitations by using two Dirichlet distributions [15]. The LDA can achieve a low value for perplexity as compared to pLSA but creates confusion as to how perplexity is related to retrieval tasks and other applications [18]. To learn, a good estimate of the number of topics for documents is required and gives a document vector as result.

	Generative Process in Dynamic Topic Model
1	Select topic $\beta_t \mid \beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I)$.
2	Select $\alpha_t \mid \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$.
3	For each document in collection:
	(a) Select $\eta \sim N(\alpha_t, a^2 I)$.
	(b) For each term/word:
	i. Select $Z \sim Mult(\pi(\eta))$.
	ii. Select $W_{t,d,n} \sim Mult\left(\pi(\beta_{t,z})\right)$.

Fig. 4 The generative process in DTM

However, the LDA suffers from the same problem as the Bag-of-words model which disregards any structure within documents between words.

3.1.1 Limitations in LDA

The topic models and their computational complexity poses a concern for model efficiency. In this part, we will discuss the overall cost to run the topic model LDA and ETM [1]. Assuming that number of topics K is set based on specific criteria initially, for instance in the ratio of the total size of documents [29]. With the huge number of documents, LDA has two main demerits: overfitting problems and high time complexity.

3.2 Dynamic Topic Model (DTM)

The Dynamic Topic Model is a variant of the LDA model and captures the evolution of topics from documents in sequential order. The paper [3] explains the model by showing implementation on articles dataset from Science journal collection over 100 years. The DTM determines the evolving topics throughout the years by applying an efficient approximate posterior inference technique. The DTM and LDA are all batch algorithms that scan the entire dataset and then make an inexact variational approximation before each update of the model.

The document collections are segregated year-wise to show dynamic behavior, and then the k-component topic model is applied to each such part. The topics associated with part 't' evolve from the topics associated with part 't-1'. For each time part 't', a K-component model with V words is considered where $\beta_{t,k}$ refers to the V-vector of natural parameters for the topic k in time 't'. The steps in the generative process of the Dynamic topic model is given in Fig. 4.

	Generative Process in ETM
1	Get topic proportions $\theta_d \sim LN(0, I)$
2	For each word n in the document:
	(a) Draw topic assignment $z_{dn} \sim Cat(\theta_d)$
	(b) Draw word $w_{dn} \sim softmax\left(\rho^T \alpha_{z_{dn}}\right)$

Fig. 5 Generative Process in ETM

The mean for parameters represents the multinomial distribution and is denoted by π. The mapping $\beta_i = \log(\pi_i / \pi_v)$ gives an ith component of natural parameters because Dirichlet cannot be used in sequential modeling. The main drawback of the Dynamic Topic Model is its use of discretization of time into different periods, also it is not considering increments in document size.

3.2.1 Limitation in dynamic topic model

DTM is not able to capture the rise and fall in the popularity of a topic. The inference algorithm in DTM is not scalable so it does not perform well in capturing large topic dynamics.

3.3 Word Embeddings and Topic model

Word embedding is being used in natural language processing extensively. The word embedding algorithm processes text corpus performs training based on certain parameters and returns vector representations of words in corpus reflecting their semantic structure [20]. Such word representation in vector form makes mathematical operations easy to perform on text corpus, even subtraction is possible (Madrid-Spain + France = Paris). When the difference between words is calculated, this enables one to find the semantic relation between words in the corpus.

Words that occur in the same context are represented by vectors close to each other. When using word embeddings, the Topic Model can extract information from a huge number of texts, also known as the "corpus" by embedding it into vector representations. This is not true for bag-of-word models, which may damage the efficiency of the model because not a lot of data is available [21–23]. The method maximizes the classification of a word based on another word in the same sentence and training complexity is proportional to the maximum distance between the words. As explained in, pivot and target word pair (j, i) are extracted when they co-occur in a moving window scanning across the corpus. The pivot word predicts the nearest target word.

The Embedded Topic model (ETM) uses word vectors to represent text documents and successfully improves the performance of the Latent Dirichlet Allocation method in terms of topic coherence and perplexity for both short and large documents [1]. Let ρ is an $L \times V$ matrix that contains embeddings in L dimensions of all the words in vocabulary where each column $\rho_V \in \mathbb{R}^L$ represents v^{th} term in the vocabulary. In ETM, through embedding matrix ρ each topic β_k can be defined by,

$$\beta_k = softmax\left(\rho^T \alpha_k\right). \tag{1}$$

The generative process in ETM includes word embedding α_k as done in LDA is shown in Fig. 5.

In step 1, *LN* notation is called logistic-normal distribution that converts Gaussian random variables to the simplex. Cat(.) denotes the categorical distribution. The ETM extracts meaningful topics from embedding space which is semantically related word assigned to similar topics.

3.3.1 Issues in word embedding

The main issue in the use of word embedding is dealing with out-of-vocabulary words. If a certain word does not exist in the word embedding phase, the model will fail in interpreting such words. In the domains, where lots of noisy and sparse data is there, this is a serious issue [30] and it becomes complex to implement the algorithm. Another limitation in word embedding is separating opposite word pairs such as "black" and "white". The word pairs like these are usually semantically very close in vector space hence reducing the performance of word vectors in tasks such as sentiment analysis [31–33].

4 Proposed method for topic modeling

The ITMWE model utilizes word embedding representations for new documents in an incremental environment as well as word vectors from old text documents. This model represents vocabulary in L-dimensional space that is like traditional word embeddings and each document can be represented by K latent topics. As done in ETM, ITMWE uses word embedding in its generative process and performs better than DTM and ETM. To find the probability of a word in a topic is the product of word embedding and topic embedding is calculated and normalized at every incremental step. A dataset with D documents $\{w_1, \ldots, w_D\}$ and D_{new} documents included in period T. The model is fitted by finding posterior distribution over the model latent variable.

At a particular stage in the extraction process, there are three main components in the database: a document from the previous stage (d), topics z from d, and a new document set (d_{new}). The algorithm in this section explains various steps in finding the latent thematic structure.

The ITMWE improves DTM and LDA model by adding the random variable from topic z from the previous stage. The generative process forms its basis on new documents d_{new} and probability distribution $p(d)$ of the new documents and old documents. The representation for document 'd' is given as a mixture of both new topics ($z_{new} = 1 \ldots Z_{new}$) and topics (z = 1…Z) from the previous stage. The process for generating document 'd' is interpreted as follows:

- From probability distribution p(d), choose a document d.
- For each word from the N-words in document d,
- -Choose a pair (z, z_{new}) based on conditional distribution $p(z, z_{new}|d)$ representing a document in the previous stage and incremental stage.
- -Choose a word based on conditional distribution $p(w|z, z_{new})$ representing the new topics and previous topic set for words.

Algorithm	Incremental Topic Model with Word Embedding (ITMWE)
Input	Document $\{w_1, w_2, \ldots w_D\}$ and their timestamps $\{t_1, t_2, \ldots t_D\}$ Initialize various hyperparameters, d_{new} represents new documents added in a specific period (T).
1	*Do the steps 1,2, 3…… iterations*
2	Get topic embeddings and latent means samples $\eta \sim q(\eta|\tilde{w}) and\ \alpha \sim q(\alpha)$
3	Compute topics $\beta_k{}^{(t)} = Softmax(\rho^T \alpha_k{}^{(t)})$ for k=1…K and T
4	Get a new documents batch (d_{new})
5	For d in d_{new} do
6	Sample the topic proportions $\theta_d \sim q(\theta_d|\eta_{t_d}, w_d)$
7	For n (each word) in the document d do
8	Estimate $p(w_{dn}|\theta_d) = \sum_k \theta_{dk} \beta_{k,w_{dn}}{}^{(t_d)}$
9	end
10	end
11	Estimate gradient corresponding to variational parameters
12	Make changes in variational parameters and update model
13	*end*

The novelty of the proposed model is based on considering prior word embeddings but only learns new patterns and embeddings from newly added documents. The proposed method is efficient in training and better based on the topic coherence value.

5 Experiments

5.1 Datasets and preprocessing

Two publicly available datasets and one dataset of collected tweets are used for performing experiments.

- The first dataset is the CORD-19 dataset [34].
- The second dataset is the NIPS papers dataset [35].
- The third dataset is the Tweets dataset collected during the covid-19 pandemic (TC19) [36].

CORD-19 dataset is prepared to deal with issues related to the COVID-19 pandemic situation by the White House and other research agencies [34] that is freely available to all research communities providing useful data/metadata about COVID-19, SARS-CoV-2, and similar health issues. The NIPS dataset contains data from papers presented at the Neural Information Processing Systems (NIPS) conference published between 1987 to 2016 [35]. The collection of scientific papers provides a diverse range of topics from the field of machine learning, neural networks, and optimization methods. The third dataset [36] is a dataset of collected tweets that contain tweet information between July 2020 to September 2020. To collect tweets, Twitter API and web scraping tools were used. This dataset is a large collection of text documents of short text.

Various preprocessing methods were applied for each dataset such as tokenization, hashtag removal, removing numbers, punctuation marks, stop words, and URLs. We also filtered stop words, words having a length less than 3, and words with a frequency of more than 60 percent. We used 30 topics (k = 30) for various experiments with three datasets.

5.2 Quantitative measure

The quality and coherence of topics extracted from the proposed method are measured using Topic Coherence and Topic diversity metrics. Topic coherence measures how different words or terms in the corpus fit in a topic and provides interpretability [37–39], the expression is given in Eq. 2. It provides the average pointwise mutual information of two words/terms that are randomly drawn from document collections as given by Eq. 2. The topic coherence measure can be used to automatically measure the quality of topics, and this filters out topics that cannot be interpreted [40].

$$Topiccoherence = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{45}\sum_{i=1}^{10}\sum_{j=i+1}^{10} f(w_i^{(k)}, w_j^{(k)}) \tag{2}$$

where $\left\{w_1^{(k)}, w_2^{(k)}, \ldots \ldots . w_3^{(k)}\right\}$ denotes the top-10 most likely words in the topic k. Here, $f(.,.)$ is the normalized pointwise mutual information [34],

$$f\left(w_i, w_j\right) = \frac{\log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \tag{3}$$

$P(w_i, w_j)$ is the probability of words w_i and w_j that occurs together in a text collection. High mutual information (MI) between co-occurring words is considered good and such topics are coherent. The other metric called; Topic Diversity is the ratio of distinct words in the top 25 words from various topics. Diversity close to zero indicates redundant topics. To find the overall quality of the group of words occurring in each topic, Topic Diversity and Topic Coherence values are multiplied. With the stated learning rate, log-likelihood ratings are produced for all the unseen documents. The model with the highest log-likelihood score is thought to be the best. Perplexity, also known as predictive likelihood, is a metric for determining how well a model can predict a sample. The loglikelihood of text documents with subjects generated by the topic model is used to calculate it.

6 Results and discussions

All three datasets utilized in this study yielded good interpretable topics because of the studies. Due to the pandemic COVID-19 predicament, the CORD-19 dataset has become a notable text dataset in recent times and is being used in several machine learning tasks. Python Gensim library is used for LDA, DTM, and Word embedding model, as well as several other python libraries for text mining and preprocessing.

Figure 6 shows the word embeddings obtained using the NIPS dataset. The word embeddings graphically depict how semantically close words are in the texts. Because the NIPS papers collection is a collection of scientific research articles, we can discern three primary categories in the embeddings in this section, such as themes like “computation”, “algorithm”, and “learning”. The word embeddings in the CORD-19 dataset are presented in Fig. 7.

The results in form of the top 10 words using the various model we have discussed so far are given. All the words/terms from topic1 discovered from the three datasets is shown in Table 3. The topic coherence measure of topics inferred from the CORD-19, TC19 and NIPS datasets

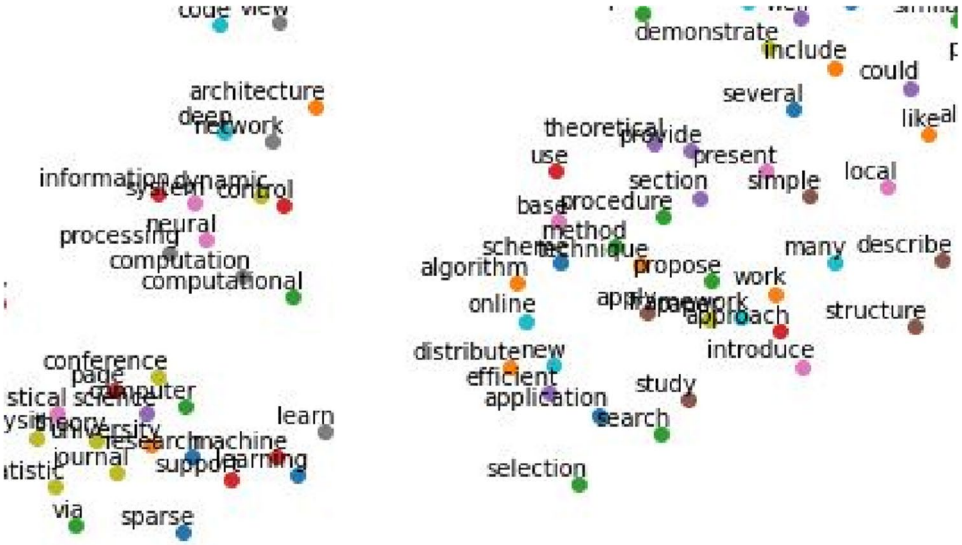

Fig. 6 Word embeddings through t-SNE plot drawn from NIPS papers collections

is illustrated through Table 4. The results given in Table 2 clearly indicates that the proposed model ITMWE outperforms the other models significantly for the TC19 dataset on topic coherence metric, which is significantly higher than long text topic models such as LDA and DTM.

The CORD-19 dataset contains articles and abstracts about coronavirus, SARS-CoV-22, and other related viruses. For the experiment, we took out papers published between November 2019 to decemeber'2020 and discovered key topics discussed in the abstracts of the papers. The topic 1 extraction has words such as 'inflammatory',' immune',' induction',' damage',' liver',' lung' which are very coherent. The topic 1 words extracted from the NIPS dataset through ITMWE are 'model',' neural',' function',' learn',' datum'

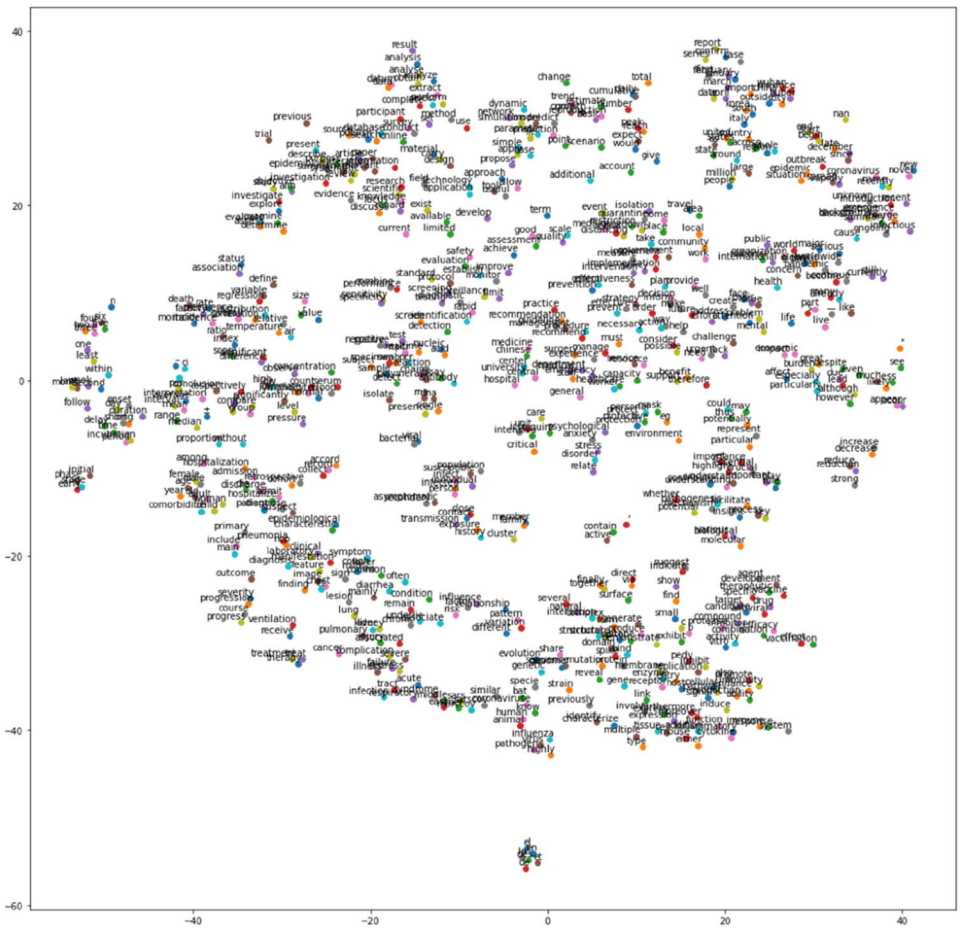

Fig. 7 Word embeddings in CORD-19 dataset plotted through t-SNE

Table 3 Top ten words in topic 1 learned by Models from CORD-19, NIPS, and TC19 dataset

Method	LDA	DTM	ETM	ITMWE
Top 10 words for NIPS (Topic 1)	Character	Function	Structure covariance	Model
	Word	Class	Pattern	Neural
	Network	Weight	Bias	Function
	Set	Layer	Estimate	Learn
	Training	Use	Noise	Problem
	Dimensionality	Result	Neural	Set
	High	Time	Datum	Network
	Different	Network	Condition	Result
	Learn	Training	Different	Natural
	Sequence	Example		Datum
Top 10 words for CORD-19 (Topic 1)	Model	Infection	Health	Lung
	Epidemic different	Virus	Public	system
	Impact	Disease	Program	inflammatory
	Base	Cause	System	innate
	Effect	Human	Care	immune
	Predict	Novel	Risk	induction
	Reduce	Report	Response	inflammation damage
	Population	Spread	Covid	Liver
	Also	Also	Provide	induce
		May	Medical	
Top 10 words for TC19 (Topic 1)	School	Covid	Lockdown	Covid
	Dead	State	Keep	Patient
	Home	Study	Daily	Break
	Stay	Record	Continue	Could
	Call	Trial	Official	Infection
	Allow	Symptom	Second	Remember
	Care	High	Area	Get
	Head	Safety	Vaccination	Hospital
	Reopen	Concern	Staff	Protest
	Person	Seem	Also	Order

etc. The topic coherence and topic diversity of models and ITMWE for all three datasets is put together in Table 2. The topic quality value is highest for ITMWE as compared to other models for all three datasets used in experiments.

Table 4 Topic coherence, Topic Quality values were observed for various models for CORD-19, NIPS, and TC19 datasets

Dataset	Method	TC	TD	TQ
CORD-19	LDA	0.4579	0.600	0.274
CORD-19	DTM	0.6311	0.229	0.144
CORD-19	ETM	0.2298	0.821	0.188
CORD-19	ITMWE	0.4221	0.836	0.352
NIPS	LDA	0.3305	0.552	0.182
NIPS	DTM	0.3011	0.183	0.055
NIPS	ETM	0.5340	0.732	0.390
NIPS	ITMWE	0.5766	0.884	0.509
TC19	LDA	0.5802	0.752	0.436
TC19	DTM	0.4987	0.321	0.160
TC19	ETM	0.5023	0.795	0.399
TC19	ITMWE	0.5102	0.898	0.457

Model is better if metric values are high

Another evaluation metric log-likelihood is used to evaluate LDA, DTM, ETM models with our proposed model. Figure 8 provides the relationship between several topics (k = 20, 30, 40, 50, 100) and topic coherence values for all four models used in the experiments. Part (b) in Fig. 8, illustrates the relationship between log-likelihood measure and document size. The figure clearly shows topic coherence value increases as the number of topics (k) increases but then decreases after reaching a value of 30–40.

7 Conclusion

Topic modeling techniques discover a diverse range of topic terms automatically from a large text corpus and work at the core of many text mining applications. We propose an incremental topic model using word embedding to retrieve latent topics for both long and short-text document collections. The experiments were performed on a topic model using three different corpora publicly available. The topic modeling framework will provide retrieval of topics and themes hidden in texts in incremental text databases. The model

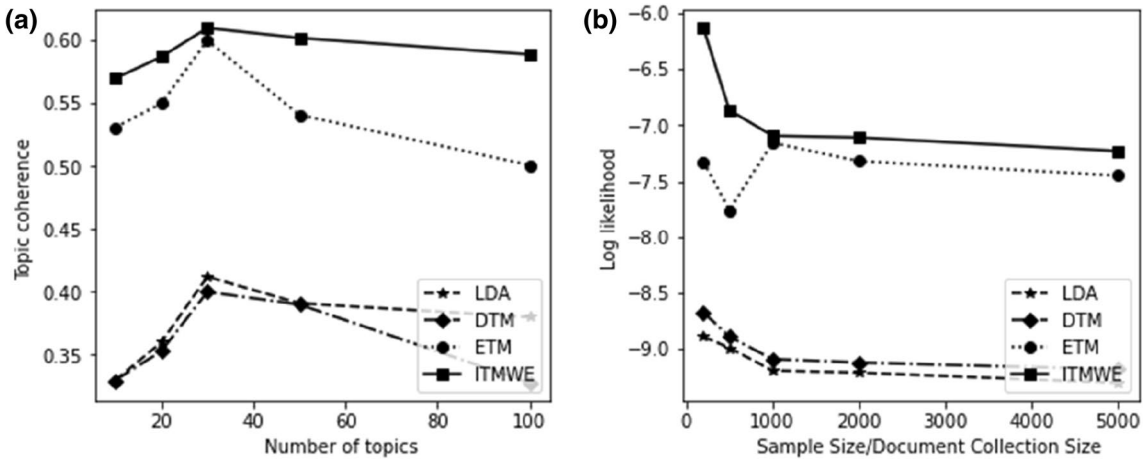

Fig. 8 **a** Topic quality measure depicted for all four models, **b** Log-likelihood measure for increasing document collection size

is effective in both short and long text documents yielding high mean topic coherences and topic diversity. The models are evaluated using Topic coherence, Topic Quality, and Topic diversity metrics. The ITMWE interprets good quality topics from all three text datasets used for implementation purposes. It is discovered that the ITMWE learns a wider range of topics than the ETM while taking much less time to fit. The limitation of the proposed model is choosing a suitable label for the inferred topics automatically, therefore, the future work will include developing strategies and techniques to generate a label for inferred topics automatically for any large-scale document collections. The model can further be applied to multilingual text corpus.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Dieng AB, Ruiz FJ, Blei DM (2020) Topic modeling in embedding spaces. Trans Assoc Comput Linguistic 8:439–453
2. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2019) Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 78(11):15169–15211
3. Qiang, J., Chen, P., Wang, T., & Wu, X. (2017, May). Topic modeling over short texts by incorporating word embeddings. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 363–374). Springer, Cham.
4. Sbalchiero S, Eder M (2020) Topic modeling, long texts and the best number of topics. Some Problems and solutions. Qual Quant 54:1095–1108
5. Sbalchiero, S. (2018). Topic Detection: A Statistical Model and a Quali-Quantitative Method. In *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences* (pp. 189–210). Springer, Cham.
6. Giordan, G., Saint-Blancat, C., & Sbalchiero, S. (2018). Exploring the History of American Sociology Through Topic Modelling. In *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences* (pp. 45–64). Springer, Cham.
7. Li Y, Rapkin B, Atkinson TM, Schofield E, Bochner BH (2019) Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. Qual Life Res 28(6):1441–1455
8. Kholghi, M., De Vine, L., Sitbon, L., Zuccon, G., & Nguyen, A. (2016). The benefits of word embeddings features for active learning in clinical information extraction. *arXiv preprint* arXiv:1607.02810.
9. Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint* arXiv:1605.02019.
10. Avasthi, S., Chauhan, R., & Acharjya, D. P. (2022). Topic Modeling Techniques for Text Mining Over a Large-Scale Scientific and Biomedical Text Corpus. In International Journal of *Ambient Computing and Intelligence* (vol 13, Issue 1).
11. Avasthi, S., Chauhan, R., & Acharjya, D. P. (2021). Techniques, Applications, and Issues in Mining Large-Scale Text Databases. In *Advances in Information Communication Technology and Computing* (pp. 385–396). Springer, Singapore.
12. Xun, G., Li, Y., Zhao, W. X., Gao, J., & Zhang, A. (2017, August). A correlated topic model using word embeddings. In *IJCAI* (pp. 4207–4213).
13. Hashimoto T, Shepard DL, Kuboyama T et al (2021) Analyzing temporal patterns of topic diversity using graph clustering. J Supercomput 77:4375–4388
14. Wang M, Yang L, Yan J, Zhang J, Zhou J, Xia P (2019) Topic model with incremental vocabulary based on belief propagation. Knowl-Based Syst 182:104812
15. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
16. Gupta M, Gupta P (2019) Research and implementation of event extraction from twitter using LDA and scoring function. Int J Inf Technol 11(2):365–371
17. Visvam Devadoss AK, Thirulokachander VR, Visvam Devadoss AK (2019) Efficient daily news platform generation using natural language processing. Int J Inf Technol 11(2):295–311

18. Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113–120).
19. Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. *In proceedings of the Twenty-Fourth conference on Uncertainty in artificial intelligence (UAI'08).* AUI Press,579–586.
20. Jähnichen, P., Wenzel, F., Kloft, M., & Mandt, S. (2018, March). Scalable generalized dynamic topic models. In *International Conference on Artificial Intelligence and Statistics* (pp. 1427–1435). PMLR.
21. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *ICLR Workshop Proceedings.* arXiv:1301.3781.
22. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2,NIPS. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.
23. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).
24. Brochier, R., Guille, A., & Velcin, J. (2019, May). Global vectors for node representations. In *The World Wide Web Conference* (pp. 2587–2593).
25. Chen Z, Huang Y, Liang Y, Wang Y, Fu X, Fu K (2017) RGloVe: an improved approach of global vectors for distributional entity relation representation. Algorithms 10(2):42
26. Lund K, Burgess C (1996) Producing high-dimensional semantic spaces from lexical co-occurrence. Behav Res Methods Instrum Comput 28(2):203–208
27. Bamler, R., & Mandt, S. (2017). Dynamic word embeddings. *arXiv preprint* arXiv:1702.08359.
28. Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining* (pp. 673–681).
29. Fountain, T., & Lapata, M. (2011). Incremental models of natural language category acquisition. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).
30. Gao W, Peng M, Wang H, Zhang Y, Xie Q, Tian G (2019) Incorporating word embeddings into topic modeling of short text. Knowl Inform Syst 61(2):1123–1145
31. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F (2019) A survey of word embeddings for clinical text. J Biomed Inform 4:100057
32. Meshram S, Anand Kumar M (2021) Long short-term memory network for learning sentences similarity using deep contextual embeddings. Int J Inf Technol 13(4):1633–1641
33. Adjuik TA, Ananey-Obiri D (2022) Word2vec neural model-based techniqueto generate protein vectors for combating COVID-19: a machine learning approach. Int J Inform Technol 19:1–9
34. Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... & Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. *ArXiv*.
35. Perrone, V., Jenkins, P. A., Spano, D., & Teh, Y. W. (2016). Poisson random fields for dynamic feature models. *arXiv preprint* arXiv:1611.07460.
36. COVID-19 Tweets dataset available at https://www.kaggle.com/datasets/sandhyaavasthi/covid19-tweetsjuly2020december2020
37. Lau, J. H., Newman, D., & Baldwin, T. (2014, April). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530–539).
38. Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).
39. Avasthi, S., Chauhan, R., & Acharjya, D. P. (2021). Information Extraction and Sentiment Analysis to Gain Insight into the COVID-19 Crisis. In *International Conference on Innovative Computing and Communications* (pp. 343–353). Springer, Singapore.
40. Avasthi, S., Chauhan, R., & Acharjya, D. P. (2021). Processing large text corpus using N-gram language modeling and smoothing. In *Proceedings of the Second International Conference on Information Management and Machine Intelligence* (pp. 21–32). Springer, Singapore.