Topic Modeling Using Latent Dirichlet allocation: A Survey

UTTAM CHAUHAN, Vishwakarma Government Engineering College
APURVA SHAH, Maharaja Sayaji Rao University of Baroda

We are not able to deal with a mammoth text corpus without summarizing them into a relatively small subset. A computational tool is extremely needed to understand such a gigantic pool of text. Probabilistic Topic Modeling discovers and explains the enormous collection of documents by reducing them in a topical subspace. In this work, we study the background and advancement of topic modeling techniques. We first introduce the preliminaries of the topic modeling techniques and review its extensions and variations, such as topic modeling over various domains, hierarchical topic modeling, word embedded topic models, and topic models in multilingual perspectives. Besides, the research work for topic modeling in a distributed environment, topic visualization approaches also have been explored. We also covered the implementation and evaluation techniques for topic models in brief. Comparison matrices have been shown over the experimental results of the various categories of topic modeling. Diverse technical challenges and future directions have been discussed.

CCS Concepts: • **Computing methodologies** → **Topic modeling**; **Latent Dirichlet**; *Mixture modeling*; *Latent variable models*; *Bayesian network models*; Maximum likelihood modeling; Markov decision processes; • **Mathematics of computing** → *Gibbs sampling*; *Markov processes;*

Additional Key Words and Phrases: Topic modeling, latent dirichlet allocation, probabilistic model, statistical inference, gibbs sampling

ACM Reference format:
Uttam Chauhan and Apurva Shah. 2021. Topic Modeling Using Latent Dirichlet allocation: A Survey. *ACM Comput. Surv.* 54, 7, Article 145 (September 2021), 35 pages.
https://doi.org/10.1145/3462478

1 INTRODUCTION

Information retrieval is the most necessitated field due to rapidly increasing sources of information over the internet. Information is made available in the form of web pages, documents, images, videos, audios, and many more. The remarkable improvement has been made in the performance of techniques for text summarization and document classification. Text summarization has received notable attraction among the researcher community. As a result, there is an enormous amount of research work being done for text summarization, text extraction, dimensionality reduction, and so on. The extensive text collection or corpus can be understood easily and quickly if it is transformed into some low dimensional subspace, such as topic structure [59, 154].

Authors' addresses: U. Chauhan, Computer Engineering Department, Vishwakarma Government Engineering College, Chandkheda, Ahmedabad - 382424, Gujarat - India; email: ug_chauhan@gtu.edu.in; A. Shah, Department of Computer Science, Faculty of Technology, The Maharaja Sayaji Rao University of Baroda, Vadodara - 390005, Gujarat - India; email: apurva.shah-cse@msu.ac.in.

0360-0300/2021/09-ART145 $15.00
https://doi.org/10.1145/3462478

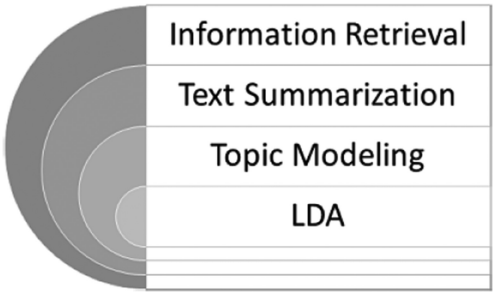


Fig. 1. Topic modeling using LDA in the view of information retrieval.

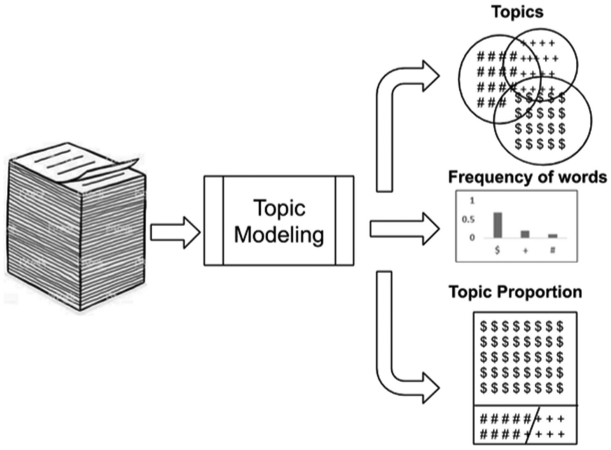


Fig. 2. Topic modeling technique.

The topic modeling applications span across many fields. Specifically, the literature covers topic modeling techniques using LDA. Figure 1 depicts that topic modeling is one of the best techniques for text summarization and information retrieval. It presents the information to the user in a compact and understandable form. Topic models have attained much interest among information retrieval researchers community in recent years. Topic modeling techniques map the text collection to low dimensional topic subspace, which is a cluster of words, known as a topic. Topic models are statistical models to uncover the hidden structure of the text data. It gives the provision to navigate through extensive text collection, digital library, web content, and so on. It can explain the corpus in a variety of ways, such as topic proportion in a document, how many documents the specific topics span across, what fraction of the collection falls into various categories of the themes, and so on [8].

Topic models can be mentioned as stochastic models, which refers to statistical procedures for learning the latent semantic structures of a massive text collection. It is often used as a text analysis tool for the discovery of unseen semantic aspects in a corpus. The topic modeling technique produces three-fold output when applied on corpus: a set of topics, frequency of words in a topic, and proportion of topics in a document. Figure 2 depicts the topic modeling technique.

A document may comprise a few topics, and one would expect to occur a semantically relevant set of words in that particular document. For instance, the words “Bank” and “Loan” may appear

more often in documents about a "Financial bank." Similarly, "Stream" and "Bank" may appear in a document about the "River." A document characteristically comprises multiple topics in different proportions. Thus, in a document that is 90% about the topic "Financial bank" and 10% about the topic "River," there would be a majority of words about the topic "Financial bank" words than the topic "River" words.

Though there are several reviews work on topic modeling, the survey in this article exhibits the novel way of representation. Moreover, we reviewed the existing study in-depth and presented a summarized form in the tabular format. Alghamdi and Alfalqi surveyed the literature work in the light of two aspects, topic modeling methods and evolution of topics over time [2]. They did not cover numerous relevant aspects such as evaluation and visualization of topic models, implementation tools and techniques, and so on.

Ali DAUD et al. showed the review over the extension of the LDA in the chronological order [27]. The review included extensively basic concepts of topic models. However, the author enhanced the review by categorizing Directed Probabilistic Topic Models into five categories. Also, topic interpretability has been examined in the presence of polysemy. In this article, too, topic models in poly language and distributed environment have not been gone through thoroughly.

Jelodar et al. presented ubiquitous survey work considering articles published between 2003 and 2016. The authors reviewed the most significant articles associated with LDA [55]. The whole review article has been organized concerning the year of publication, but they have not considered their characteristics to categorize them. Debortoli et al. have presented a tutorial on topic modeling based on its working [30]. They showcased the inherent challenges of topic modeling, instead of challenges exposed by the extension of LDA.

We considered topic models research tasks coming from all directions. We attempted to taxonomies topic modeling based on their features, such as hierarchical nature, dataset type, the language of the dataset, centralized or distributed computing, and so on. Besides, a far and wide variety of underlying domains for topic modeling exploration in our survey made a remarkable difference from previous review work. We also explored the tools and technology section for the implementation view. Moreover, research articles on the visualization of topics were reviewed with different angles. We also presented the linguistic challenges faced by topic models while having applied them in multilingual settings. Moreover, topic modeling evaluation techniques were presented, which added extra flavor in the work.

The article has been organized as follows: Section 2 provides brief about related works. Section 3 mentions about working of LDA. Section 4 discusses about topic models across the numerous domains and categories. It also briefs the topic modeling for the non-text data. Section 5 explores variety of topic visualization approaches. At the end, Section 6 focuses on sources for the implementation of topic models. Most importantly, Section 7 discusses the research issues and future trends of topic modeling.

2 BACKGROUND

If a document is to be considered a **Bag of Words (BoW)**, then few anchor words would be adequate to express the theme of the document. **Latent Semantic Indexing (LSI)** was considered the basic technique for thematic discovery from the text archives [31]. LSI transforms the original dataset in different spaces so documents and terms about the same concept can be mapped. Words and documents have many to many relationships. So difficulties might be aroused when a word has multiple meanings, and this is exceptionally regular in the large text data. LSI accomplishes this by **Singular Value Decomposition (SVD)**. The concept is represented by words that

occur together and words having only one meaning. The last simplification seems unrealistic and imposes a limitation on LSI. It occurs that the resulting dimension becomes difficult to interpret.

Even though LSI can find topics from the large set of documents, it does not describe the document generation process. The limitation is overcome by **probabilistic Latent Semantic Indexing (pLSI)** [49]. LSI maps the word with a concept, while pLSI maps word with concepts with a specific probability. It builds the model in a more meaningful way, by assigning each word to a topic sampled from a multinomial distribution over topics. As a result, a document would be a mixture of multiple concepts or topics, a model where the document represents exactly one concept. Asuncion et al. proved that pLSI performs better than LSI [4]. Though pLSI is proficient for generating documents, it does not fit into the fully generative model. pLSI does not have the provision to explain documents that are not part of the given collection.

Blei et al. first introduced **Latent Dirichlet allocation(LDA)**, a generative model, which is capable to model topics for unseen documents, too [12]. The LDA is a probabilistic graphical model, as it can find the proportion of the one variable given value of other variable [143].

The wide variety of topic models were developed across an extensive range of domains and languages. Asuncion et al. analyze the connection between Gibbs sampling [16, 41, 77, 109, 118, 181], variational inference, and maximum a posterior [4], Spherical Admixture Model for L_2 normalized data [116]. LDA is an unsupervised text analysis technique, though it also works for the labeled document [90], a supervised variation for LDA. Wallach et al. dismissed the restriction of the bag of words [144].

Topic modeling does not consider the order of document in the corpus. The work for considering the time of document generation was investigated by Ma et al. [87] for discovering topics of great interest, time gap analysis for three different text sources [56], time-dependent clustering of streaming data [60], to understand news immersion in the collection of news articles over the time [54, 62, 71]. However, Wang et al. proposed an n-gram topic model framework for learning not only topics but phrases also [150]. To speed up the sampling process for the biterm itself, researchers have offered the extended version known as FastBTM [45, 171].

Web resources have also been part of interest for LDA topic modelers. The topic model for Wikipedia link disambiguate was developed [125], link recommendation by uncovering hidden topics [88], learning health-relevant topics from social media [105], a recommendation system for Twitter [65]. On the same line, how topics spread geographically in the city of London [68] was examined. **Topic-Sentiment Mixture (TSM)** was proposed to model the topic in the Weblog archive [92].

Big data analytics caused wide usage of parallel computing techniques and distributed systems [89]. Topic discovery from a massive collection of documents finds time complexity due to learning thousands of topics from trillions of documents. As a result, Researchers integrated distributed and parallel computing aspects with LDA for faster topic inference. Hsiang et al. [169] applied the MapReduce technique for learning topics, which resulted in the distributed version of LDA. Parallel Latent Dirichlet allocation proved itself good scalable [152]. The variation inference is used with parallel LDA for scalability [175]. Against Hadoop Map-Reduce, Sayadi et al. [121] implemented distributed LDA on Spark. LDA also was employed for modeling topics from online streaming for the infinite size of vocabulary [48, 174] utilized variation Bayes for online LDA. Furthermore, Online LDA and incremental LDA were developed for streaming data [21].

LDA was integrated with neural network [96, 157, 172], and collaborative filtering. Furthermore, topic models Zipf distribution [173], topic models for online streaming data [48, 166, 174], the temporal topic models, which considered a time of documental, also were evolved and attained the performance level [32, 56, 60, 62, 87, 148, 149, 167].

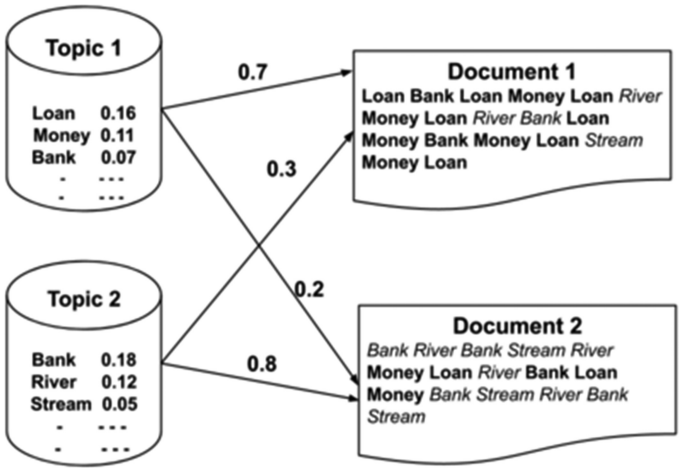


Fig. 3. The Generative process of LDA. The topic proportion for document 1 is 70% and 30% for topic 1 and topic 2, respectively. Similarly, document 2 has topic proportion as 20% and 80% for topic 1 and 2, respectively. For each topic, there is a multinomial distribution over words, which indicates the probability of getting sampled word for that specific topic. A document would be generated according to the proportion of topic and the probability of words.

3 LATENT DIRICHLET ALLOCATION – A GENERATIVE PROCESS

LDA is a mixture models, because documents are considered as a mixture of topics [12, 49, 50, 129]. They are also known as admixture because its segments are itself mixture of other segments [46]. Specifically, in the topic models, documents are a mixture of topics, and in turn, topics are a mixture of words. The LDA generative model allows a set of observations to be explained by a group of unobserved variables [42]. To generate the new document, as explained by the generative procedure of LDA, it samples the topic from a multinomial distribution. The probability that a specific topic gets selected depends on some prior belief of distribution over topics. Once the topic is sampled, the words are sampled for that topic. A topic is a probability distribution over vocabulary words. Hence, the number of times a word can appear in the document depends on how high probable that word in the given topic. This process is iterated as many times as several words to have in the document. The generative process of LDA has been depicted in Figure 3.[1] It generates two documents from two topics. The documents have a different proportion of topics, and topics have a diverse probability for words. The proportion of all topics sum up to 1 in the specific document. There are some assumptions as mentioned below for generative process.

(1) A document is a bag of words. The order of words is not counted.
(2) Each document is a mixture of topics. That is, for K topics, every topic takes some proportion in the document.
(3) Each topic is a mixture of words.

With these assumptions, documents can be generated as mentioned below. The root of this theory can be understood by making use of Bayesian Network. Bayesian network is a type of a probabilistic model that derives the probabilistic relationship between random variables involved in the process [8, 66]. The Bayesian network can be depicted by the acyclic graph whose node represents the random variables and edges represents the dependencies between the variables. The process

[1]Figure 3 has been rebuilt from Figure 2 in Reference [129].

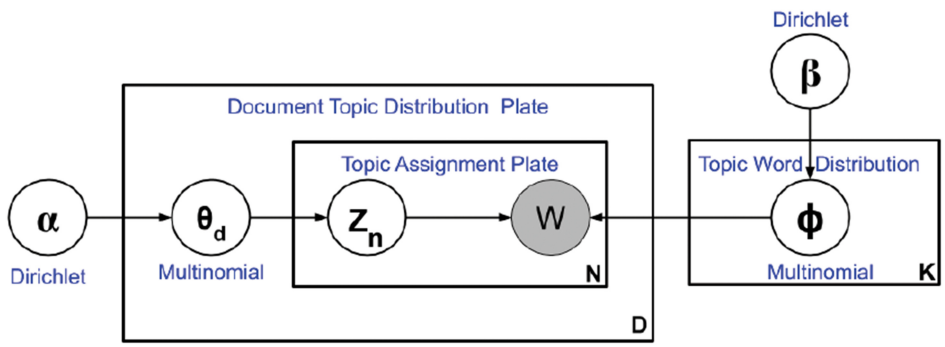


Fig. 4. Plate notation for LDA generative algorithm [12].

is to be repeated for each document. The number of words in the documents are distributed normally. The document consists of N words. Figure 4 depicts the generative process for LDA using plate notations. It seems reasonable to produce a documents from a given topic distribution and a word-topic distributions. There are terms that need to be interpreted solely. "Latent" is a Greek word, which means the "hidden." It specifies the hidden structure (topics) are to be divulged given the documents collection. Algorithm 1, shown below, describes the document generative process using LDA.

ALGORITHM 1: Generative Algorithm for LDA

Input: Dataset, K topics, Hyperparameter α and β
Output: Topic files, Topic Word distribution, Document Topic distribution
for *All topics* $k \in 1, K$ **do**
  // sample the probability distribution over words for each topic
  sample mixture components $\phi_k \sim Dir\beta$;
end
for *all documents* $m \in 1, M$ **do**
  // proportion of topics for each document
  sample mixture proportion $\theta_m \sim Dir\alpha$;
  // Length of documents in the corpus is normally distributed
  sample document length $N_m \sim Poisson\xi$;
  for *all words* $n \in 1, N_m$ *in document* m **do**
    // assign the topic to each word
    sample topic index $Z_{m,n} \sim Mult\theta_m$;
    // identify the word identity from a probability distribution over words
    sample term for words $W_{m,n} \sim Mult\phi_{Z_{m,n}}$;
  end
end

Gibbs Sampling. LDA shows the generative process of documents. In practice, we have to infer parameter value given a set of documents. One has to apply one of the statistical inference techniques for that. Gibbs sampling is one of the methods to achieve a chain of observations that are approximated by a specified multivariate distribution. Gibbs sampling is useful when a joint distribution is not known or is difficult to sample directly, but the conditional distribution is known and easy to sample from References [41, 129]. The sequence of observations can be used to approximate the marginal distribution of a random variable or group of random variables. Gibbs sampling

is associated with the Bayes theorem. It is a randomized method and being used as an alternate of deterministic approach such as the Expectation-Maximization algorithm. It can be considered as a general architecture for sampling from a large group of random variables. Gibbs sampling fits very nicely in the Bayesian network when the posterior distribution is to be inferred from the prior distribution and observed random variables.

4 SOPHISTICATED TOPIC MODELS

The NLP research community often prefers to explore diverse datasets to discover the patterns. The type of text may vary from one corpus to another corpus. The primary objective is to learn the latent aspects of the corpus. The corpus might vary in the variety of factors, such as the total number of documents, the average length of the document, vocabulary size, noise in the corpus, and so on. Additionally, the researchers may consolidate some domain knowledge or linguistic aspects in the corpus. The dimensions of the corpus may play an indispensable role.

4.1 Plain Topic Models

Topic Models for Research Articles and Authors. The topic modelers initially preferred to model topics over text corpus straightforward. Author topic model [120] is regarding building a set of documents, with the available group of authors, and a probability distribution over words for each author of the group. The suggested *Author-Topic Model (ATM)* employs the advantages of both, the topic mixture model and the author model. The work helps find the appropriate reviewers of research papers. Given the abstract of the article and list of authors in addition to known past collaborators, then it outputs a list of profoundly relevant authors who can be reviewers for this article. Researchers applied symmetric KL divergence to find the dissimilarity between any two authors.

A similar kind of research was attempted to portray the affinity between authors and research work. The cocitation analysis may expose the mapping between a set of authors [82]. The objective is to find out alike interest are of research. The proposed work was not a replacement of the co-citation analysis, but it can be put into the execution. The technique was tested for the research articles, which were collected from the broad range of journals of the category **"Library and Information Science" (LIS)** category. Likewise, Mimno and McCallum offered Author-Persona-Topic to model the association between a research paper and their best match reviewers. It fetched the set of experts(based on past data) for the evaluation of the given research article [93].

Topic Modeling over software domain. The research scientists concentrated on the software engineering domain to model the topics. Linstead et al. discovered topics from 1,555 different software projects [76]. The millions of lines of code were used for modeling topics over them. They suggested that the technique could be applied for software refactoring and project organization like tasks. In a relevant study, LDA was employed to determine the bug in the source code [84] by Lukins et al. A bug report based query was shaped and decided the portion of source code to correct it. Thomas et al. practiced LDA for modeling the evolution of software over the time [133]. They used the software repository of JHotDraw and jEdit. Sun et al. to maintain software [130]. They introduced a method named **MSR4SM (Mining Software Repositories for Software Maintenance)** targeting two jobs of software maintenance, feature allocation, and change analysis. In a different study, Corley et al. [25] analyzed the difference between two consecutive versions of software repositories using the online topic model. They mainly analyzed changesets-based and snapshot-based for **feature location techniques (FLT)** and **developer identification techniques (DIT)**. Tian et al. proposed an LDA-based technique, named LACT, to analyze source code as probabilistic mixture [134]. Chen et al. modeled the software quality metrics were as topics over

Table 1. Plain Topic Model

Authors	Pub. Year	Source of dataset	No. of Documents	No. of Tokens	Vocabulary Size	No. of Topics
Rosen et al. [120]	2004	NIPS Research Articles	1,740	23,01,375	13,649	100/200/400
Griffiths et al. [42]	2004	PNAS Abstract Articles	28,154	30,26,927	20,551	50/100/200/400/500/600/1000
Blei et al. [11]	2007	JSTOR Articles	16,351	57,00,000	19,088	100
Farrahi and Gatica-Perez [36]	2011	Location based mobile data	800,000 hours of data	NA	NA	200
Lu et al. [82]	2012	ARIST Research Papers	5,227	NA	NA	50
Cohen et al. [23]	2014	Patient Note	8,557	60,00,000	NA	50
Vorontsov et al. [139]	2015	NIPS	1,740	22,01,375	13,639	100
Kayi et al. [58]	2017	CT clinical reports	3,705	NA	1501	5–150
Heo et al. [47]	2017	Journal papers (MEDLINE)	170,099 Papers	NA	1501	10/20/30/40/50
Hagen et al. [43]	2018	WtP petitions	3,344 petitions	324,594	37,663	10–110

software code files [20]. Asuncion et al. worked for the trace the software development activities by implementing software traceability using topic modeling [5]. They implemented several tools to assess the advised routine.

Topic Modeling for Scientific articles. The scientific domain was chosen to model the topics to find out the emerging field of the science [42]. It discovered hidden topics and exposed the dominating area of research in science. The study aimed at examining the most attractive annexes of the study in science. On the other end, it was found the fact that some areas lost their essential in the field of science. Similarly, Heo et al. modeled topics for MEDLINE journal research papers for discovering topics from the bioinformatics field [47]. Yau et al. worked for topic modeling over scientific publications [168]. Similarly, Kayi et al. applied topic modeling techniques to classify computed tomography reports [58]. They suggested three different models: **Aggregate Topic Classifier(ATC), Condence-based Topic Classifier(CTC)**, and **Similarity-based Topic Classifier(STC)**.

Topic Modeling over clinical data. Cohen et al. confirmed that the text summarization task might result in poor outcomes when redundant data are present in the corpus [23]. The inherent noise and redundant data must be got rid of before presenting them to the method. The analysis of the influence of redundant data on the quality of the topic model was examined. The target domain of the dataset comprises of patient notes. Authors have elucidated that if a few words are copied in a document from another document, then topic distribution might not be as good as it should be.

Farrahi et al. investigated the pattern for daily routine based on location [36]. They practiced the plain LDA for Author-Topic Model to reveal the unseen patterns of the real-life dataset collected by cellphones. Hagen et al. advised an architecture for training and evaluating LDA [43]

for analyzing e-petitions records. The primary objective was to synchronize computer automation and human interpretation by integrating the precise direction of human decisions in LDA training, evaluation, and interpretation. The research findings were compared using multiple sources including manual content analysis. Besides, topic modeling for the economic history journal was developed by Reference [155].

Model for correlating topic. Vorontsov et al. addressed the limitation of Dirichlet priors, which are not capable to produce vectors with zero elements [139]. As a result, Dirichlet priors do not get aligned with the sparsity. The authors have proposed a novel non-Bayesian and semi-probabilistic approach, which is different from a fully probabilistic approach. In addition to this, ARTM used EM in place of Bayesian inference and simplified the inference process. The authors have performed a test: sparsing, smoothing, and decorrelation. It was concluded that though it simplified the process, the model learned topics faster and equally interpretable at the same time.

The relationship among the inferred topics was found out in the **Correlated Topic Modeling(CTM)** [11]. The advantage of modeling correlation among topics has better predictive accuracy. The Dirichlet distribution of the document was replaced by a logistic normal distribution. The prediction process considered after observed some portion of the corpus. It supports those results; authors reasoned that CTM could predict the topic of the word better due to the topic relationship, while LDA did not get these features, as it operates independently on each word. Table 1 depicts the comparison of the important parameters of the above discussed sophisticated topic models.

4.2 Hierarchical Topic Models

The hierarchical structure forms the tree of topics. Each node of the tree exhibits the topic. As the evaluator traverse the tree from the root node to the leaf node, it finds that the topic gets more specialized from the generalized view. There are noteworthy exploration endeavors for the hierarchical structure formation integration with LDA. Blei et al. showed how the topic could be learned in the form of a hierarchical structure with the help of Bayesian nonparametric methodology [9]. They proved that no prior knowledge of the structure of data or hierarchy is required. The Chinese restaurant process was practiced to define prior distribution on the data. The tests were set up for three different datasets. The outcomes of the trials demonstrated multiple levels of abstraction in the tree structure.

The exchangeability assumption does not stay feasible if the data points are dependent on the corpus under examination. The occasion becomes stronger when we model the topics on the temporal dataset specifically. Kim et al. displayed the idea of **distance-dependent Chinese restaurant Process (ddCRP)** to model the dependency among the documents in the corpus. They have considered the non-exchangeability of the token in the documents and also documents in the corpus [61]. The recommended method captured the phenomena of sequential data patterns. The research team found the proposed technique performed better than HDP and LDA. Kim et al. also worked for labeled data using the HDP [63] in the supervised mode of topic modeling technique. They offered the **Hierarchical Dirichlet Scaling Process (HDSP)** technique, which scaled the topic proportion for those topics that were tightly associated with labeled data.

Zhu et al. proposed an unusual variation of hierarchical LDA having continuous words instead of discrete, named **Sparse Coding Latent Dirichlet Allocation(SCLDA)** [182]. The different parameters of the model such as corpus, documents, and latent topics were associated among themselves, so the suggested model was generalized to LDA. The proposed method could was applied either by a generative framework or a discriminative framework. The SCLDA was applied for object classification and natural scene categorization. The SCLDA outperformed LDA in the set of experiments. Zhang et al. applied HDP multiple times (EvoHDP) to expose cluster evolution

Table 2. Hierarchical Topic Models

Author	Pub. Year	Hierarchical topic formation method	Inference technique	Remarks/Method
Teh et al. [131]	2005	Chinese Restaurant Process (CRP)	Gibbs sampling Collapsed Gibbs sampling	Each group of data comprises of number of components
Chang et al. [18]	2009	Undirected link modeled as a binary variable	Expectation Maximization	Documents modeled and then links between documents were modeled
Blei et al. [9]	2010	Nested Chinese Restaurant Process (nCRP)	Gibbs sampling Collapsed Gibbs sampling	hLDA is more interpretable than LDA
Kim et al. [61]	2011	distance dependent CRP (ddCRP)	Gibbs sampling	The exchangeability assumption is released
Kim et al. [64]	2012	reccursive CRP (rCRP)	Gibbs sampling	Performed better than nCRP
Zhu et al. [182]	2012	Sparse coding	Expectation Maximization (EM)	Words in the model are not discrete but continuous
Kim et al. [63]	2014	Gamma representation of HDP	Variational Bayes (VB)	HDSP performed better prediction compared to HDP labeled LDA and Partially labeled LDA
Wang et al. [147]	2015	Granular computing	Gibbs sampling	Incorporate partial supervision from incomplete knowledge of domain expert
Li et al. [74]	2017	Hierarchical Dirichlet Process-Inverse Regression (HDP-IR)	Stochastic Variational Inference	The nonparametric approach to overcome uncertainity of topic structure

from associated time-varying document collection [176]. They showed that evoHDP predicted better compared to HDP. Similarly, Wang et al. divided the corpus into a sequential group based on the time stamp. They modeled topics for each group and demonstrated how the topic evolves from one group to another over time [10, 146].

As data becomes available streamingly, the newly arrived data have to be adapted for incremental learning. On the relevant track, the HDP was improved by working with an incremental approach. Wang et al. reshaped HDP by assimilating two ideas [147]. The expert knowledge was put for the learning process like supervised learning. They remarkably contributed by introducing information granules, which are a collection of information organized in the groups. A group was composed of similar or coherent information. On the parallel track of research, the associations between the documents were modeled to view the organized network of documents [18]. They proposed **Relational Topic Model (RTM)**; the technique seems to be applicable for recent trends such as social network, citation of papers, navigation of webpages through hyperlinks, and so on. RTM provided link prediction more accurately compared to the state-of-the-art technique. The knowledge-based hierarchical topic model was proposed to form the topics hierarchy of multiple domains [159].

Li et al. proposed a novel nonparametric topic model that integrated the **Inverse Regression (IR)** with the **Hierarchical Dirichlet Process (HDP)** [74]. The HDP-IR offered a couple of advantages: (1) It resolved the indecision for the number of topics and (2) It improved the predictive performance by extending Inverse Regression with **Sufficient Dimension Reduction (SDR)**. Table 2 summarizes the variety of hierarchical topic models proposed.

4.3 Multilingual Topic Models

The world is considered the home of all nations and peoples living interdependently. As a result, people need the ability to access multiple languages. The topic modelers have also been working with parallel corpus in multiple languages. The researchers have explored many languages using topic modeling techniques.

Zha et al. worked for word alignment using bilingual dictionary [178]. The topic inference was carried out in bilingual settings. Generally, the statistical Machine translation system works on sentence-pair irrespective of whether they are from document-pair or not. But with the approach of admixture model or statistical inference, a document in source language can be translated into the target document as a whole. The translation is mapped by considering essential topics discovered by that document-pair. Similarly, Mimno et al. targeted multiple languages in the work for machine translation [95]. The technique was proposed for topic modeling techniques in polylingual circumstances. Authors exploited this corpus to find the ability of the model for discovering similarities between vocabularies across languages. The topic inference was carried out for the various number of topics. The work was extended by Peng et al., who modeled the topics for code-switch social media across a pair of languages (English-Spanish and English-Chinese). This combination of languages in the same perspective promoted associations among the word types for numerous languages using the shared topics [107].

Zosa et al. grabbed the advantages of both, topic models for multiple languages [95] and tracking the evolution of topics over the time [10, 33], and offered Multilingual Dynamic Topic Mode [183]. They managed to prove the ability of the proposed model to detect the pick or fall of the significant event during the specific time slice by executing experiments on parallel and comparable corpora. However, Yang et al. did not follow the typical assumption that documents are comparable to different languages intrinsically. Targeting the low-resources languages, researchers presented the topic model for many languages [164]. They could infer better semantically coherent topics in comparison to previous multilingual topic models.

Zhang et al. proposed an idea of entity linking between two languages to overcome the issue of a word-to-word mapping. They learned topics for English and Chinese language [177]. Researchers prepared a framework to demonstrate the process step-wise. A similar research task for event linking was proposed for English and Dutch language [28]. They did not practice machine translation or intermediate dictionary. Instead, they modeled topics for two languages under study. One more piece of work of the related category for retrieval of information based on LDA across numerous language is Reference [141].

Liu et al. devised a framework for modeling topics over multiple languages accomplished the objective of bilingual dictionary built-up [79]. They first modeled parallel topics over comparable documents. In other words, the comparable topic corpus was formed from a comparable document corpus. These parallel topic corpora were used to extract a bilingual dictionary. The approach did not need the seed of a bilingual dictionary. A pioneering model, **Cross-Lingual Query Log Topic Model (CL-QLTM)**, systematically incorporated web search data in different dialects by jointly employing cross-lingual vocabularies. The co-occurrence relations of words was also used in the query log [57].

Table 3. Multilingual Topic Models Using LDA

Authors	Pub. Year	Languages	datasets	Remarks
Zhao and Xing [179]	2008	English Chinese	General newswire	Statistical machine translation
Mimno et al. [95]	2009	Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish.	Proceedings of European parliament	Machine translation and analyzing topic trends across the language
		Welsh, German, Greek, English, Farsi, Finnish, French, Hebrew, Italian, Polish, Russian, and Turkish.	Wikipedia	
De Smet and Moens [28]	2009	English Dutch	Wikipedia	Event linking
Zhang et al. [177]	2013	English Chinese	Wikipedia	Entity linking across the languages by modeling bilingual topics
Peng et al. [107]	2014	English-Spanish, Chinese	Twitter data Weibo data	Utilized code switched social media to discover correlations across languages
Liu et al. [79]	2015	Japanese, English, Chinese, French	Wikipedia	Built the bilingual dictionary using parallel topic corpus
Vulic et al. [142]	2015	English, Dutch, Spanish, French, Italian	Wikipedia	Cross-lingual event-centered news clustering
Jiang et al. [57]	2016	English Chinese	Search engine query log	Analyzed search query log in bilingual settings
Zosa et al. [183]	2019	German, English Finnish, Swedish	News articles	Topic evolution over time for poly languages
Yang et al. [164]	2019	English-Chinese English-Sinhalese	Wikipedia Disaster Response	Topic models for low-resources languages

On the same corridor, Vulić et al. discussed topic modeling in the multilingual environment set up in Reference [142]. The primary objective was achieved to provide a sketch of thematic structure in the multi-language environment, based on the experimental work on bilingual topic modeling. The accomplishment inferred that the bilingual topic modeling technique could be generalized in a multilingual background.

Ganguli et al. proposed a framework for cross-lingual information retrieval [39]. The technique worked for retrieving the documents from the target language while the query is given from the source language. Vulic et al. built a **cross-lingual information retrieval (CLIR)** modal with document aligned comparable corpora [141] to defeat the problem of low resource languages in case of source-target language pair. The authors used only the per-document topic distribution and per-topic word distribution for developing the CLIR model. The work was centered around query document retrieval when the query asked in one language and the answer of the query in another language. Table 3 exhibits numerous topic models in the poly language setup.

4.4 Topic Models in Distributed Environment

Researchers tested LDA in a distributed or parallel environment, too [123, 152, 169]. The key objective was to summarize the document collection quicker by diminishing the inference time of modeling topics. Smola and Narayanamurthy proposed a general architecture for inferring topics over multiple workstations in parallel mode [127]. The authors found that the same can be applied for other probabilistic models such as hierarchical and n-grams models.

The document(s) summarization using LDA was attempted for single-node in the work References [19, 44]. The multi-document summarizer based on MapReduce architecture was expressed in the work Reference [98]. Summarization process time was found proportional to the number of workstations. Additionally, They found improvement in the scalability and semantic similarity of words in a topic.

Newman et al. exploited the inference process by executing in distributed setup, at the same time updating the topic count globally [100]. To achieve the goal, the dataset was segregated across the CPUs, and the inference was performed in parallel and distributed mode. AlSumait et al. proposed an extension for tracking the evolution of topics over time, known as On-line LDA [3]. The research emphasized on keeping the model up-to-document whenever a new document or set of documents becomes available. The composite version of both of the techniques as mentioned earlier, **Distributed online LDA (DoLDA)**, was introduced in Reference [6]. It was found that DoLDA was twice faster than online LDA experimentally. At the same time, perplexity remained almost stable. Zhai et al. proved The LDA for streaming data may run with vocabulary length infinitely [174].

Nallapati et al. developed parallel LDA with the Expectation-Maximization process [99] to deal with the oversized corpus. The distributed environment was set for parallel inference on multiprocessors. The set of assessments brought out the measurement of speed and scalability. Both of the factors got improved over the non-parallel inference. The researchers also uncovered the ways for advanced improvisation in the performance.

Mimno and McCallum [94] presented **Dirichlet Compound Multinomial LDA (DCMLDA)** to settle the issue of organizing a large scale digital library. Unlike the work in Reference [99], a global topic cluster is formed in the end. Smyth et al. offered an asynchronous learning approach for two unsupervised methods LDA and **Hierarchical Dirichlet Process (HDP)** [128]. Instead of updating topic-word counts globally, Gibbs sampling is applied to their local data and transfer their information in a local asynchronous mode with other processors. The proposed method produced almost similar outcomes as if the inference process was carried out on a single-processor.

Wang et al. employed parallel LDA on MPI and MapReduce surroundings [152] for scalability measure. They proved that PLDA could be made functional for gigantic text collection. The researchers presented two PLDA implementations based on the AD-LDA algorithm, first on MPI [132] and second on MapReduce [29]. The inference process followed parallel Gibbs sampling, in which the training set was divided and distributed equally on P processors.

Table 4. Parallel/Distributed LDA

Author	Pub. Year	Dataset	Parallelism	# Processors # Nodes
Newman et al. [100]	2009	KOS, WIKIPEDIA, NIPS, PUBMED, NEWSGROUP	Giggs sampling inference process	Processors up to 3,000
Wang et al. [152]	2009	Wikipedia The forum set	Parallel LDA using MPI and MapReduce	16 to 256 nodes 64 to 1,024 nodes
Liu et al. [81]	2011	NIPS Wiki-20T Wiki-200T	data placement, pipeline processing, word bundling, and priority-based scheduling	64 to 2,048 nodes
Bak et al. [6]	2012	Tweet Collection	Variation inference	8 core on each node 60 nodes
Nagwani et al. [98]	2015	Federal Court of Australia 4,000 legal cases	MapReduce/Hadoop for document filtering	Four processors Four nodes
Yu et al. [169]	2015	Enron, NYTimes, PubMed Amazon UMBC	Nomadic Framework for Parallel LDA	20 core on each node 32 nodes
Moon et al. [97]	2018	KOS, NIPS Enron, NYTimes	Multiple vector processors (VLDA)	192 cores per MP
Tian et al. [135]	2019	KOS, NIPS Enron NYTimes, PubMed4M	Nomadic Framework for Parallel LDA	6,126 processor 12 cores

Liu et al. extended the PLDA with PLDA+ by introducing four different strategies [81] aiming at load balancing, scalability, and lessened communication bottleneck concerns. PLDA+ outperformed PLDA over these technical issues. Likewise, the issues of parallelizing of inference methods on GPU were addressed in Reference [162]. It was also capable of balancing the load on each multiprocessor and resolved memory access conflicts. Again, the novel procedure established itself better scalable at the same time producing equivalent coherent topic models. Moon et al. proposed an advanced GPU-based LDA version that could attain better performance with respect to time of inference by forcing the less data movement [97]. Tian et al. achieved the 3 to 12 times faster inference by taking assistance of data-level and thread-level parallelism [135]. The limitations of existing models, which are high-memory demands and imbalance workload, were overcome by priority-aware scheduling approach. Table 4 presents the summarized form of topic models in distributed, parallel, and online mode.

4.5 Topic Model with Prior Embedded Information

Incorporating domain knowledge into the corpus is an innovative research direction. To address the intricacy of topic models, researchers approached to incorporate characteristics of the language under study. It has also been attempted to give weight to the terms for better orientation of topic models. Moreover, features such as co-occurrences of words, synonyms of words are well-thought-out for topic models.

Accentuating the utilization of the characteristics of words to increase the topic coherence was introduced by Petterson et al. for modeling topics [108]. The assessments were carried out to enhance the topic quality by incorporating the dictionary information. According to the authors, it is ubiquitous to have a dictionary of words with synonyms. The topics can be derived more coherently if similar words tend to be centric to a common topic. For a single language, lexical similarity, distribution similarity, and synonyms could be integrated into the standards LDA. Lexical similarity also was proposed for the multilingual setting.

There was an essential work submitted by Wilson et al. for weighting the terms of a bag of words [158]. The work contrasted in the sense that terms would be carrying distinctive weight unlike equally weighted terms in all previous endeavors of the topic modeling. The main goal was to prevent the influence of high-frequency words from spreading across many topics. It was observed experimentally that by incorporating the weighting scheme in the Gibbs sampling remarkable improved the performance of the model.

Likewise, Hu et al. presented the iterative technique to refine the set of words of a specific topic [53]. The authors wanted to improve the interpretability of topics by recursively integrating constraints. It modeled the topics numbers of times, and in each iteration, some constraints were added based on the output of the previous run. By adding constraints, they proved that groups of words are formed as they must be. On-site human knowledge was incorporated into the topic models. The topic should be composed of semantically relevant words ideally. There might be the scenario that synonyms of words fall into another set rather than it should appear. Ma et al. accentuated on the use of a synonym for the extraction of features of the products [86]. Due to the diversity of product features for users' preferences, features were extracted automatically. Again, it was not independent but worked jointly with the LDA approach.

Das et al. suggested that if semantically relevant words groups are known in advance, then they must be grouped into a topic, too [26]. The authors examined the word embedding effects in unsupervised settings. The proposed technique, that embedded words, was not part of the existing topic model. The proposed approach could assign high topical probability, which is a topical term but not available in the dictionary [26]. Researchers proved proposed techniques captured better semantic coherence compared to the classical LDA approach. Moreover, the inferred set of topics could be used for second-order applications. The sentiment classification for tweets with embedded word features carried forward to an innovative concept. To embed words, topics were inferred using LDA [117]. The words of topics were used for further classification of tweets. Li et al. proposed a technique that dealt with the limitation of topic modeling for the short text [72].

Yang et al. proposed the complete structure, known as factor graph framework, to leverages the embedding of prior knowledge for the modeling topics over oversize collection [165]. They presented the technique **Sparse Constrained LDA (SC-LDA)** a robust extension of the plain LDA technique, for proficiently integrating prior knowledge into LDA. They showed SC-LDA attained both better interpretability of topics and quicker inference of the topics at the same time. Li et al. extended the **Dirichlet Multinomial Mixture(DMM)** model by modeling topic number as Poisson distribution [72]. The proposed model was named PDMM, which was further extended by auxiliary word embedding to recommend the semantically associated words under the same topic. Brody and Lapata offered a model that generated words from local topics selected based on a context window around the ambiguous word [14].

Bunk et al. have made use of word embedding for modeling topics in the proposed work [15]. They merged word embedding and LDA and named the proposed method, WELDA. Though both the aspects differ fundamentally, as LDA works on bag-of-words and Word embedding considers the small window of contextual words, authors smartly combined them to uplift the semantic coherence of the topics. Before coming up with the utilizing the strength of both the idea, they

Table 5. Topic Model with Prior Embedded Information

Authors	Pub. Year	How embedding works.
Brody and Lapata [14]	2009	A model that generated words from local topics chosen depends on a context window around the vague word.
Petterson et al. [108]	2010	Side information (features) was used as prior to β. The prior probability was learned based on words distribution over topics in the context of dictionary, synonyms collection, thesauri, etc.
Wilson et al. [158]	2010	The weight was assigned to the high-frequency words, which did not have the presence on the list of stop words.
Ma et al. [86]	2013	LDA was combined with synonym collection to excerpt product features from online product review.
Das et al. [26]	2015	Gaussian distribution replaced by parameterized, which considers opaque word embeddings instead of word type.
Yang et al. [165]	2015	Proposed a factor graph architecture, Sparse Constrained LDA (SC-LDA) for including prior word correlation and document label LDA.
Ren et al. [117]	2016	First, topics were formed using LDA. The words of topics were used to classify Twitter's tweets using SVM.
Li et al. [72]	2017	Auxiliary word embedding to support the semantically linked words under the same topic during the sampling process. DMM model was extended by modeling topic number as Poisson distribution (PDMM).
Xun et al. [161]	2017	Embedding has been done at word level and utilized it with topic modeled using multivariate Gaussian distribution.
Bunk et al. [15]	2018	Word Embedding (WE) has been combined with LDA to accelerate the inference process, having enhanced interpretablility.
Dieng et al. [33]	2019	Exploited embedded (ETM) topic model and dynamic topic model (DTM).

were tested individually and the outcome was not encouraging in terms of words associativity. Besides, the authors could have experimented with varying sizes of contextual words window. The interpretability could have been analyzed with the varying size of context words window.

The further advancement of topic modeling techniques, the correlated topic model incorporated the word embedding theory. The authors proposed a **Correlated Gaussian Topic Model (CGTM)**, which amalgamated the idea of word-level correlation and topic level correlation to enhance the semantic associativity of the topics [161]. Moreover, they had to apply multivariate Gaussian distribution, because the input data were semantically relevant word embedding instead of a sequence of discrete words. The research fellows did not mention the total inference time of the model including the word embedding process using the word to a vector representation.

Blei et al. proposed the **Embedded Topic Model (ETM)**, taking advantage of specialties of topic model and word embedding [34]. As the embedding procedure fits the words with a similar meaning to the same representation, it helps discover topics with better interpretability. Nonetheless, embedding space enhances the semantic associativity among the words of the topic, and it may increase computation cost, as the word to vector transformation is required.

Topic evolution over time was investigated by Blei et al. considering three-time slice [10]. On a similar track of the research, Dieng et al. extended the examination by bringing the innovation in the existing **dynamic topic model (DTM)** and suggested the **dynamic embedded topic model (DETM)** [33]. DETM integrated the idea of word embedding with modeling the topic evolution over time. The investigational results showed that the interpretability of topics increases, in shorter topic formation time considerably. The span of the test could have been extended to measure the time and semantic coherence of topics with a varying number of iterations of the inference process.

4.6 Topic Modeling for Short Text

Usually, documents such as news articles, research papers, and legal reports, abstracts of research papers are considered as large documents. However, the dataset might contain only short documents, such as tweets. The topic modelers have practiced the topic models on the short text dataset. Besides, it may be shaped in a precise way to learn them correctly [23].

Cheng et al. discovered topics for a collection of short texts [21]. As short texts are prevailing on the web in the form of social networking websites, microblogging websites, questions in the QA system, and much more. In addition to the short text sparsity, the authors found two other challenges: the frequency of the word in the original short text was not very discriminative, and ambiguity is more due to limited content. The merging of short texts into a document was pointed out as a convenient solution. Table 6 shows the summary of various topic models over short text.

Ren et al. performed Twitter sentiment classification by embedding words into the tweets [117]. The incorporation of words was accomplished through topic discovery using LDA. To boost the performance, the topic enhanced words were linked with **Support Vector Machine (SVM)**. Further, LDA was applied to find a cluster of similar users. Then the most relevant hashtag was recommended to the users [180]. The Hashtag-LDA was a merged approach to model words and hashtags in the microblog. Likewise, Ramage et al. proposed a topic model to characterize Twitter users on their style, status, and social characteristics of posts [112].

Tweets might have time-stamped and geographic location identifiers. Lansley et al. used tweets made from London to learn topics [68]. The task was impressive to uncover the behavior of Twitter across inner London. Roberts et al. worked around tweeter dataset to analyze the hidden emotions [119]. They learned how various emotions were distributed in tweets. The topics inferred by LDA were used as features for inputting in binary SVM classifiers.

Vo et al. used the title of the scientific document to model topics over them. In turn, topics were enriched by incorporating the external text. Three comprehensive datasets were verified against three machine learning procedures, SVM, KNN, and naive Bayes [137]. Following the same, topic-enhanced word embedding was introduced in Reference [117] for sentiment classification of Twitter data. The task ended up successfully by including SVM, a traditional classification method to outperform other state-of-the-art methods. TWILLITE, a recommendation system, was implemented for Twitter data [65]. LDA was used for recommending a certain number of top users to follow several top tweets.

The **Ailment Topic Aspect Model (ATAM)**, another augmentation of LDA, was proposed to unhide the health topics from the collection of tweets [105]. The ATAM filtered comprehensive Twitter data based on health-specific keywords in a supervised manner. An ample number of studies related to health targeted various zones of health and diseases for illness detections. Some of them were performed for specific illness detection. ATAM revealed new health issues without any prior knowledge. The work did not approach any specific illness. Likewise, Instagram posts and comments were modeled with the help of LDA in Reference [78].

Inherently, the short text undergoes the problem of very inadequate word co-occurrence, which leads to the sparsity issue. As the topic comprises words that are a very small fraction of the

vocabulary, the topic-term matrix becomes too sparse. Similarly, it happens with the document-topic matrix. The **biterm topic model (BTM)**, proposed by Yan et al., overcomes the sparsity difficulty at document-level [45]. FastBTM accelerated the inference process on the foundation of BTM [21].

Quan et al. suggested that sparsity phenomena cannot be solved reasonably by transforming the short text to pseudo-documents. They offered a solution by presenting **self-aggregation based topic model (SATM)**, which assumed that short texts are generated as a part of long documents itself, hence to capture the relationship between such short text snippet and the relevant large documents is sufficient to model topics over the short text [111]. The authors compared their findings to LDA and unigram but did not compare with any other short text topic modeling approach. Contrary, Zuo et al. proposed the **Pseudo-document-based Topic Model (PTM)** to learn topics over short text that also conquered the issues of overfitting and expensive computation. The researchers analyzed the short text and pseudo documents association matter in detail and suggested the optimal solution. On this track, they enhanced and extended PTM by **Sparsity-enhanced PTM (SPTM)**. They showed the benefits of SPTM when pseudo documents are small in size and one-to-one mapping between pseudo documents and short text [185].

The original PageRank algorithm was customized to unveil the influence of users on other users [126]. The link structure of Twitter users was modeled based on "Who follows who" along with the like-minded users' topics. Weng et al. modeled the behaviors of users by followers of users and to whom they follow. Though research could get succeed to find out the link structure among the followers, they could not figure out the unusual behavior. It left space for carrying the research forward on the same track. The examination targeted mainly two terminologies concerning the social behaviors of the Twitterers: "reciprocity" and "homophily." Bhattacharya et al. applied akin theory to learn user interest in Twitter social network [7]. The researchers figured out that a user always follows the expert of the specific field. To achieve this point, they learned the subject expertise of Twitter users first and then deduced the followers for the particular experts. In Reference [153], a framework was proposed to discover the topics of Donald Trump's followers on Twitter. Furthermore, the "likes" on those topics were modeled alongside weights.

On the analogous route, Hong et al. predicted the popularity of the tweets by considering the factors such as the number of retweets, contents of the message, metadata, how good the user's profile, and so on [51]. The work also attempted to derive how quickly and farther the message may get propagated. The event detection from streaming data is a vital and inspiring research task. Cordeiro investigated for detecting events from streaming tweets by linking wavelet analysis with topic inference summarization [24].

The LDA was applied with Gibbs sampling for inferring topics from tweets. The interest of users was modeled as topics in two ways: first by the methodology of suggested and by exploring the profiles itself. The topics were found more interpretable in the case of suggested methodology.

Lin et al. addressed the issue of sparsity by applying the "Spike and Slab" prior, at the same time using weak smoothing prior to smoothing prior [75]. They took care of selecting relevant topics and words by smoothing prior, also not selecting irrelevant topics and words by weak smoothing prior. As a result, the sparsity issue was targeted from two sides. Though they claimed that topic quality can be improved, they also agreed to the threat that Bernoulli selectors may cause trouble in multinomial distribution.

Li et al. exploited neural network-based language modeling to fit word embedding for modeling topics over the short text [73]. Researchers overcome the sparsity issues of short text topic models by incorporating semantically relevant words with the help of word embedding. They proposed GPU-DMM, which samples the semantically relevant words using Generalized Poly Urn built on

Table 6. Topic Model for Short Text

Author	Pub. Year	Dataset/Method	Remarks
Weng et al. [156]	2010	3,200 tweets of Singapore based Twitterers/LDA	The PageRank algorithm has been referred
Hong et al. [51]	2011	10,612,601 messages and 2,541,178 users/LDA	To predict how quickly message will becomes breaking news
Cordeiro [24]	2012	13,651,464 tweets/LDA	Detecting event from streaming tweets by linking wavelet analysis
Roberts et al. [119]	2012	7,000 tweets/LDA	Emotions such as "Anger," "Happiness," "Stress," etc., discovered from tweets
bhattacharya et al. [7]	2014	3,200 tweets/L-LDA	Labeled LDA applied on the expertise of the specific field
Paul and Dredze [105]	2014	Health-relevant tweets from 2009 to 2012/ATAM	ATAM having embedded words relevant to health domain
Cheng et al. [21]	2014	Tweets2011 Collection of tweets/BTM	Short text of social networking and microblogging websites, questions in the QA system
Kim et al. [65]	2014	12,098,339 tweets from 8405 Twitter users/LDA-VEM	Recommending a certain number of top users to follow several top tweets
Ren et al. [117]	2016	20,000 tweets/LDA	Words were incorporated in tweets and classified using SVM
Lansley et al. [68]	2016	289,240 tweets/LDA	Uncovers the behavior of Twitter for London
Zhao et al. [180]	2016	UDI-Twitter Crawl-Aug2012-Tweets/TTM	Tag-topic model determined the most probable tags for topics
Wang et al. [153]	2016	US2016 Twitter dataset/LDA	The "likes" on those topics were modeled with weights
Li et al. [73]	2016	BaiduQA 179,042 QA pairs Snippet 12,265 snippets	Auxilary word embedding using NN language Model (GPU-DMM)
Li et al. [72]	2017	BaiduQA 648,514 QA pairs Snippet 12,340 snippets	Poisson DMM (PDMM) and extended version based on GPU (GPU-PDMM)

Dirichlet Mixture Model and achieved a better semantically coherent topic set. In separate work, Li et al. addressed the limitation that short text maps only one topic. In their proposed **PDMM (Poisson distribution DMM)** model, they released the restriction that a short text always comprises only one topic by allowing the short text to be represented by a small number of topics [72]. On this occasion, too, they found better interpretable topics and better text classification accuracy. They extended PDMM by integrating with GPU Model and offered another variation named GPU-PDMM. Nonetheless, the integrated version received better accuracy, but put more computation power on the other side. Qiang et al. surveyed over the various short text topics models and displayed the comparison of their performance in different scenarios [110]. Plus, they developed an open-source JAVA library called STTM. The easy-to-use programming interface and the designed module for the evaluation of methods made STTM more countable in the short text topic modeling field.

4.7 Modeling Topics over Non-textual Data

The application of LDA has not only covered textual data, but it has also expanded itself over audio, video, and image data collection. As shown in Table 7, LDA has been put into practice to solve many issues relevant to non-text datasets, such as video fingerprinting, image quality improvement, grouping the audio files, and so on.

Table 7. Topic Models over Audio, Video and Image Domain

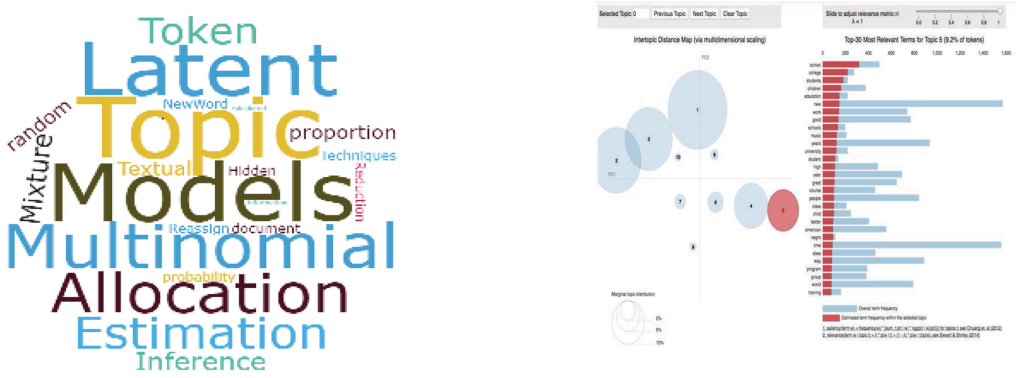
Authors	Pub. Year	Method Name	Domain
Liu et al. [80]	2012	Attribute-Restricted Latent Topic Model	Video
Vretos et al. [140]	2012	LDA with ScaleIn variant Features Transform	Video
Elguebaly and Bouguila [35]	2013	Reversible jump Markov chain Monte Carlo	Video
Hu et al. [52]	2014	Gaussian Latent Dirichlet allocation	Audio
Yang et al. [163]	2014	Hierarchical Variant of Latent Dirichlet allocation	Video
Kooiji et al. [67]	2015	LDA and Mixture of Gaussians	Image
Zou et al. [184]	2016	Locally Consistent Latent Dirichlet allocation	Video
Niu et al. [104]	2017	Dirichlet allocation with Mixture of Dirichlet Trees	Image
Yuan et al. [170]	2019	Gaussian Latent Dirichlet allocation	Image

4.7.1 Topic Modeling over Video Dataset. Copy detection, replica identification, or content-based similarity findings are challenges issues for video datasets. Vretos et al. used LDA along with facial image features to develop a video fingerprinting framework [140]. The content-based video retrieval novelty was introduced using LDA to fill the semantic gap. Liu et al. proposed the **Attribute-Restricted Latent Topic Model (ARLTM)** for person re-identification, which recognizes the specific person from the collection of videos taken from multiple cameras [80]. **Reversible jump Markov chain Monte Carlo (RJMCMC)** technique experimented for human action video categorization, pedestrian detection, and face recognition [35] by Elguebaly and Bouguila. The work for human action recognition was also performed with an innovative latent semantic learning method based on structured sparse representation [83].

Likewise hierarchical topic model for text data, Yang et al. attempted the same successfully using h-LDA for recognizing human action [163]. Instead of capturing and recognizing human space from video collection, the **Locally Consistent Latent Dirichlet allocation (LC-LDA)** model learned collective patterns using tracklets and bag-of-words as low-level features [184].

Yuan et al. introduced **Gaussian Latent Dirichlet allocation (GLDA)** in the diversity induced image retrieval domain [170]. Then GLDA was joined with spectral clustering to improve relevance and diversity of retrieval outcomes.

4.7.2 Topic Modeling over Images Dataset. Generally, an image is considered as a document when LDA is applied to the collection of images. In this work, the research team segmented the images and then applied GLDA on each image segment instead of on the whole image as a document. To improve the coherence of the image cluster, they reorganized the images in the cluster. Simultaneously, the images with lower similarity were moved to another cluster where those are more similar. Notably, the reorganization of images in the cluster was performed based on metadata of those images. The proposed approach improved performance in all aspects such as coherence, diversity, and relevance. Kooij et al. defined an innovative technique for detecting many objects in a collection of images [67]. They released the restriction of one-to-one correspondence between image and objects identity. They have projected a creative regularized **Semi-Supervised Latent Dirichlet allocation (r-SSLDA)** for learning visual concept classifiers [83]. The proposed



(a) Topic as a Word Cloud

(b) LDAvis - Gensim Topic Model Visulization

Fig. 5. Topic visualization word clouds and LDAvis.

technique capitalized both supervised and unsupervised topic modeling approaches. Moreover, recognizing multiple objects from unclear detection [67] was carried out with the help of a stochastic process.

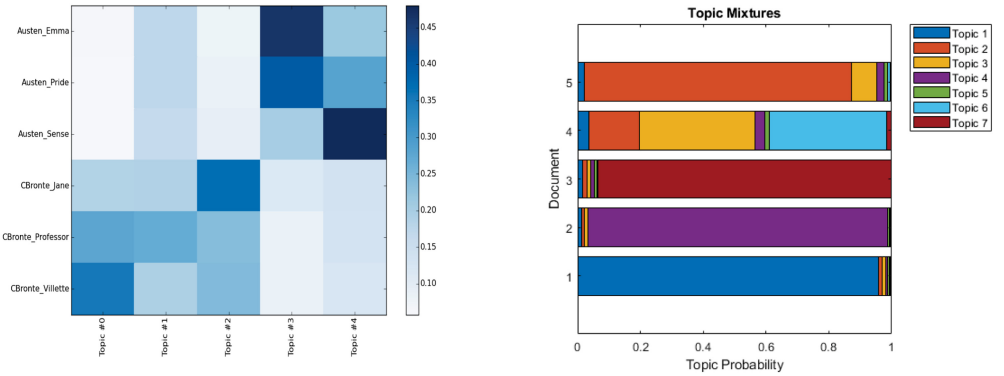
Niu et al. performed object class identification and localizing the crucial segments of the image object in the unsupervised enviornment [104]. They leveraged the prior-knowledge available for the image collection. They identified Must-Links images to resolve the issue of polysemy in visual words. The authors proposed **Latent Dirichlet allocation with Mixture of Dirichlet Trees (LDA-MDT)**, which included Must-Links for object discovery. The Must-Links influenced only one or some topics of the model, instead of all the topics. As a result, they could achieve better semantic coherence of the visual topics.

4.7.3 Topic Modeling over Audio Dataset. LDA confirmed accomplishment for modeling over audio related data, too. Hu et al. proposed a topic model for audio retrieval [52]. The standard LDA treats the topic distribution over words, while Gaussian LDA considered audio features for topic distribution. The Gaussian-LDA was found better performing compared to plain LDA.

5 TOPIC VISUALIZATION

Topic models are used to examine a text collection, and the revealed topics must represent the corpus under study. The aforementioned research effort advocates that the quality of a topic is often determined by the coherence [1, 101, 145] of its constituent words and its relative importance [1, 69] to the analysis task in comparison to other topics [70]. The set of discovered topics is presented to the end-users in some graphical format. A range of data visualization methods become available, such as Word clouds, stacked bar charts, and pie charts, some of the simple graphical representations of topics. These visualization methods are thought-provoking because of the high dimensionality of the model. The number of topics learned in the model is so much smaller than the number of discrete words in the vocabulary. In this section, we examined some topic visualization techniques offered by different research scholars.

Word cloud is a cluster of different words with varying font sizes. The font size indicates the weight of words in that particular cluster. Researchers have represented topics as word clouds [68, 126, 136]. They used the probability of words as the weight in the word cloud. Figure 5(a) depicts the salient words of Topic Models as a cloud. Sievert and Shirley proposed a collaborating and web-based environment for the visualization of discovered topics [124]. Figure 5(b) depicts the



(a) Heatmap - Topic proportion in documents

(b) Topic proportion with stacked bar chart

Fig. 6. Topic visualization heatmap and stacked bar chart.

topic visualization. The development tool supported the bird-eye view of all topics. Moreover, it allowed examining the term-topic association profoundly for each topic. A web-based interface for visualizing the topic was proposed by Reference [38], named as TopicVis. It displayed topics predominantly in the form of a pie chart and stacked bar charts.[2] The area covered by the specific color is proportional to the proportion of the topic in that document. In addition to the stacked bar chart, a heatmap is also used to show the topic proportion in the document. Figure 6(a) shows an example of heatmap representation of topic.[3] Heatmap looks like a matrix, where row and column correspond to document and topic, respectively. Based on the chosen color palette, the darkness of color indicates the influence of a specific topic in that document.

Termite, a result-oriented visual analysis tool, suggested for judging the quality of the topic model [22]. Termite presented a tabular layout of topics and terms associated with the topics. It enabled users to compare terms intra-topics and inter-topics. The researchers proved, with a set of illustrations, that Termite enhanced the interpretability of topics and model as a whole. A more interactive system, TopicNets [40], was offered for analyzing topic model efficacy. Authors provided a complete package for an effective analysis of an individual topic, topic distribution for a document or set of documents, and also for the whole corpus, and so on. It was also empowered by providing features such as filtering, searching, interactive, and drill-down. The analyst could search with various options and parameters.

6 TOPIC MODELS TOOLS

Practitioners have developed a comprehensive collection of techniques to recapitulate a variety of datasets. Though, efficient techniques are required to be designed to meet the expectation of the current era of unstructured text analysis. We need different techniques and methods on various stages of the topic inference process from the corpus under study. To practice for different languages, the implementation of such a module needs some customization. Likewise, topic visualization is also an essential part to be integrated as part of the topic modeling implementation.

6.1 Implementation

We discuss diverse tools, techniques, and technologies for modeling the topics, visualizing them, and tune the inference process according to the need. Blei et al. implemented a C language code

[2] https://i.stack.imgur.com/tCw8E.png.
[3] https://i.stack.imgur.com/tCw8E.png.

of **latent Dirichlet allocation (LDA)**, a model of discrete data that is fully described in Reference [12]. The lda-j, the Java version of lda-c, has also been implemented.[4] Another Java-based LDA has been developed by Gregor Heinrich.[5] Likewise, a C/C++ development of LDA using Gibbs Sampling technique for parameter estimation and inference is also available.[6] Google Code service hosts the source code of various technologies and algorithms.

Ramage et al. proposed a supervised version of LDA including partially labeled LDA and labeled LDA [113, 114], the Python implementation is available for both the approaches.[7] To speed up the inference process, SparseLDA has been implemented.[8] It proved itself 20 times quicker than plain LDA [166]. Lda-go[9] is a Gibbs sampling training and inference of the Latent Dirichlet allocation model written in Google's Go programming language.

Wang et al. and Lui et al. proposed parallel LDA for large-scale applications and implemented corresponding module.[10] Using MapReduce and Hadoop distributed environment, A package has been employed for massive text collection summarization.[11] This package has been developed by a Cloud Computing Research Team in University of Maryland, College Park. The PLDA has been enhanced by OpenMP and MPI-Based Parallel Implementation of LDA.[12] The package delivers the combined features of multithreading and distributed computing. OpenMP determines multithreading functionality. The multiple-node (distributed computing) functionality is dependent on MPI. The parallel algorithms, implemented in this package, are based on the work proposed by Newman et al. [102, 103]. Table 8 presents the various tools for topic models.

6.2 Tools

6.2.1 MAchine Learning for LanguagE Toolkit - MALLET. MALLET [91] is a topic modeling package implemented in Java. In addition to topic modeling, it can be used for document classification, information extraction, a variety of natural language processing tasks, machine learning applications, and text summarization. Furthermore, it offers sequence tagging with the help of algorithms such as Hidden Markov Models, Maximum Entropy Markov Models, and Conditional Random Fields. The MALLET topic model package comprises a tremendously fast and extremely extensible implementation of Gibbs sampling, proficient techniques for document-topic hyperparameter optimization, and tools for deducing topics for new documents given the trained model.

6.2.2 Stanford Topic Modeling Toolbox (TMT). TMT has been written in an old version of Scala using the linear algebra library [17]. It offers services to the humanities and social scientists who want to examine the immense accumulation of documents. It delivers features such as input/output from/to spreadsheets, train the model, select parameters for the model, and topic generation. In addition to plain LDA, it gives the facility of Labeled LDA and parallels LDA to summarize the dataset.

6.2.3 Matlab Topic Modeling Toolbox 1.4. Mark Steyvers developed a complete package of topic modeling in Matlab.[13] The author has made it free to use for scientific purposes. The toolbox

[4] http://www.arbylon.net/projects/.
[5] http://www.arbylon.net/projects/LdaGibbsSampler.java.
[6] http://gibbslda.sourceforge.net/.
[7] https://shuyo.wordpress.com/2013/07/24/python-implementation-of-labeled-lda-ramage-emnlp2009/.
[8] https://github.com/ankazhao/python-sparselda.
[9] https://code.google.com/archive/p/lda-go/.
[10] https://code.google.com/p/plda/.
[11] https://github.com/lintool/Mr.LDA.
[12] https://code.google.com/p/ompi-lda/.
[13] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

Table 8. Tools and Techniques

Techniques/Tools	Inference Method	Language
MALLET	Gibbs sampler	Java
Gensim	Variational Bayes	Python
Scikit-Learn	Variational Bayes	Python
TMT-Matlab	Gibbs sampler	Matlab
Stanford-TMT	Collapsed Variational Bayes	Scala
Yahoo LDA	Variational Bayes	C++
Mahout LDA	Collapsed Variational Bayes	Java
R package for LDA	Collapsed Gibbs sampler	R
ompi	Gibbs sampler	C++
Mr. LDA	Variational Bayes	Java
tomotopy	Collapsed Gibbs Sampling	Python

comprises plentiful scripts for modeling topics over the text collection. The various flavors of topics modeling were offered, such as **Author Topic Model (AT)**, HMM-LDA, Collocation Model.

6.3 Packages and Libraries

6.3.1 Gensim. Gensim [115] is a potential topic modeling toolkit implemented in Python. It deals with data streaming and robust incremental algorithms, unlike other tools that operate in batch mode alone. It is explicitly designed to summarize massive datasets. It uses NumPy and SciPy Python libraries for the implementation. Gensim is employed across diverse fields, such as health, insurance, social media, document organization, document classification, and so on.

6.3.2 Scikit Learn. Scikit learn is an easy-to-use and proficient Python library for text analysis [106]. LDA implementation of sci-kit learn uses variational inference to sample from a tractable approximation of a topic model's posterior distribution. It provides online and batches update approaches.

6.3.3 R Package for LDA. The package mainly incorporates supervised LDA, correlated LDA, and the mixed-membership stochastic block model. The statistical inference process has been implemented through fast collapsed Gibbs sampling using C language. Further, input/output methods used stereotypically in topic models and posterior distribution assessments was accommodated in the package. It offers a variety of functions for the LDA and the relevant model.

6.3.4 Yahoo LDA. Yahoo LDA makes available a fast C++ module for the statistical inference technique. It adds the provisions for multi-core parallelism and multi-node parallelism using Hadoop cluster. As it is capable of working on a distributed architecture, it has got the potential to discover a large number of topics from the immense collection of documents.[14]

6.3.5 Mahout for LDA. The statistical inference may follow either Gibbs sampling or Variational Bayes inference. In turn, the Gibbs sampling technique was extended by Collapsed Gibbs sampling. The Mahout implementation for LDA benefits from the combination of features from both of the procedures named **Collapsed Variational Bayes (CVB)**. The algorithm usages two methodologies to marginalize out parameters when computing the joint distribution.[15]

[14] https://github.com/sudar/Yahoo_LDA.
[15] https://mahout.apache.org/users/clustering/latent-dirichlet-allocation.html.

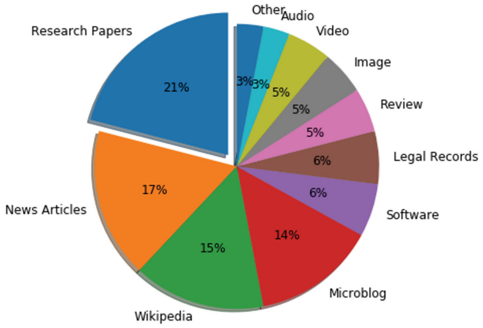


Fig. 7. Contribution of various domains for modeling topics over them.

6.3.6 Tomotopy. In addition to Gensim and Scikit learn, Python has been enriched by the package tomotopy.[16] Tomotopy is a Python extension of **tomoto (Topic Modeling Tool)**, a Gibbs-sampling based library originally developed in C++. Tomotopy uses Collapsed Gibbs-sampling to infer the latent topics. Moreover, Tomotopy exploits multicore CPUs with a SIMD instruction set, which makes the topic inference process faster compared to Gensim. The latest version of tomotopy provides a variety of topic modeling in addition of LDA, such as Labelled LDA, Supervised LDA, Correlated Topic Modeling, and several others.

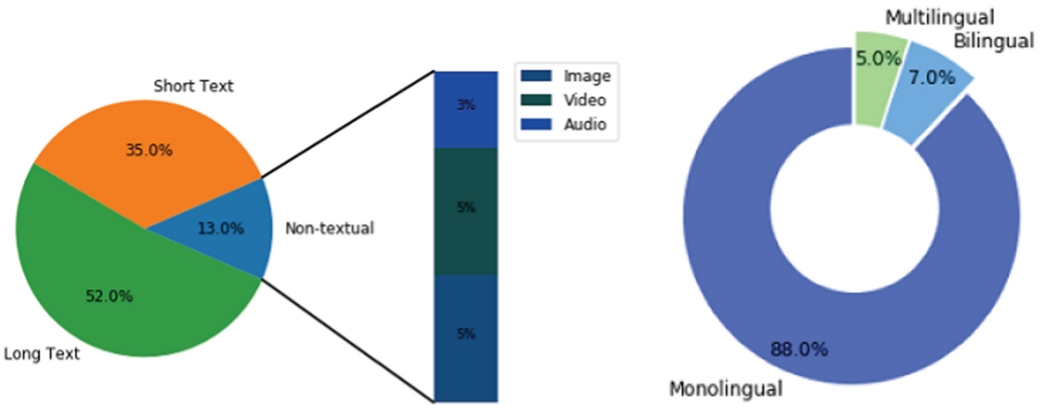
7 ISSUES AND RESEARCH TRENDS

Although LDA underwent a variety of extensions far and wide, several segments still demand more introspection. The preprocessing such as stopwords elimination, stemming, lemmatization, and so on, requires to be developed for different languages. They appear an easy part of text summarization, however, its absence or presence plays a crucial role in the final output, because careful consideration of such preprocessing tasks gives more meaningful topics in a quick time. However, the exclusion of such steps may result in the extensive size of the vocabulary. As a result, the inference process demands more computation power. Besides, linguistic characteristics such as synonyms, polysemy, homonymy, hyponymy, and so on, have not been given that importance. These linguistic characteristics may give better semantic coherence of topics.

Moreover, the preprocessing components are vital in the topic modeling. NLP research tasks generally involve language-specific preprocessing actions. Instead of removing language-specific stopwords, Schofield et al. proposed topic models preceded by corpus-specific stopwords [122]. Li et al. suggested a novel way to remove noise for modeling topics over the short text. They introduced an idea of common topics to collect noisy words. In Reference [160], a novel approach was proposed to locate general words automatically from the corpus. Moreover, instead of analyzing unigram for discovering topics, bigram, trigram, or n-gram may exhibit some interesting, surprising, or unexpected outcomes [138].

Once topics are modeled, there would be a question of the quality of those individual topics. Topic modelers got to determine how individual topics are meaningful and useful. The perplexity and semantic coherence are familiar methods to evaluate the topic model based on the unseen set of documents. Also, the set of inferred topics might comprise mixed topics, chained topics, undistinguishable topics, and illogical topics. Besides, topic model quality also can be diagnosed with the measurement of the parameters, such as the difference between a document and token frequencies, distance from corpus distribution, prominence within documents, burstiness [13].

[16] https://pypi.org/project/tomotopy/.



(a) Proportion of textual and non-textual datasets

(b) Topic Modelings in monolingual and multilingual setup

Besides, the posterior inference is used to infer the topics of the corpus. It approximates the word distribution for topics and topic distribution for documents. Though Gibbs sampling and variational inference are well-known and being used widely, the researchers also may choose Laplace approximation or variational approximation [12]. The convergence in Gibbs sampling necessitates more time compared to variational inference; the curious scholars should look for other choices. Furthermore, LDA has been being practiced in its plain form across a variety of domains and applications. However, LDA have been extended by applying **non-negative matrix factorization (NMF)** [85], non-Markov model [146, 149], non-parametric [37], and asynchronous parameter settings [128, 151].

Moreover, we analyzed the interest of the topic modelers concerning the domains; more attraction was found towards textual corpus against the non-textual. The researchers have shown more attachment towards the domains of research articles, news articles, and Wikipedia articles for modeling topics, it can be observed in Figure 7. They have been charming because of their volume and variety. The ease of extraction from the internet could also be one of the reasons for getting attention from the NLP community. Similarly, micro-blogging portal data has been sought after for modeling topics over the short text, as they can be fetched easily with the use of APIs. However, large document datasets have been still enjoying the position on top; Figure 8(a) supports the fact. Likewise, monolingual topic modeling has received major attraction against bilingual and polylingual work, as depicted in Figure 8(b).

However, there is a wide room available to carry the research on over the non-textual datasets. Though the building block of a non-textual dataset is a challenging task (the word is the building block in case of the textual dataset). It may help uncover many hidden characteristics. Multilingual topic modeling also keeps the door open for the researchers to take the research task to the new height. Though parallel corpora helped in the alignment of topics, one can use the results in other NLP tasks such as Machine Translation, Word Sense Disambiguation, and so on. Additionally, there could be the integration of OCR with LDA by fetching out the text (handwritten or typed) and modeling topic over the accumulated text. One may explore the by aligning long text and equivalent short text using LDA to prepare the extractive summary of the long text articles. Furthermore, Figure 8 exhibits the trends of topic modeling over the course of time since its invention. It conveys that the researchers always look for something innovative while modeling topics. Right from modeling topics over the corpora, it was followed by hierarchical topic modeling, embedded topic modeling, and so on. The future of the research inclined towards the neural topic models.

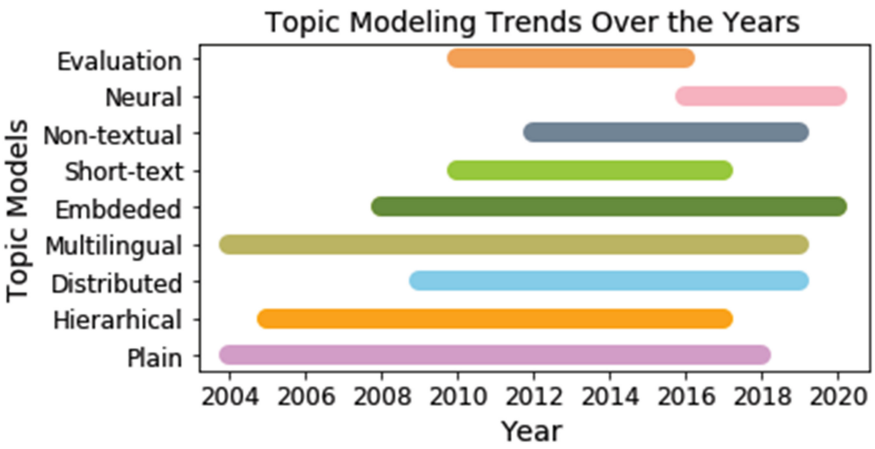


Fig. 8. Trends of topic modeling.

8 CONCLUSION

In this survey article, we delivered a concise overview of topic modeling and LDA. We reviewed the LDA application over a diversity of domains, such as research paper, clinical data, Twitter data for health care, and so on. There is no uncertainty that there are ample domains that still need to be explored using LDA. We studied that many authors tested LDA on the distributed environment for the parallel execution. The remarkable volume of work surveyed for topic modeling over numerous languages. We have seen that researchers also attempted to find two different language alignments. The investigators endeavored topic modeling by embedding external knowledge for more meaningful topic extraction. As interpretability of topic models matters a lot, topic modelers presented research work for topic model evaluation techniques and the efficacy of those techniques. They also explained how and up to what extent the automatic evaluation got aligned with the human judge. We briefly introduced several tools, coding, and software available. The researcher has also implemented its customized or extended methods of topic modeling using LDA. Moreover, topic modeling visualization techniques proved a very important unit of text summarization. We brushed up a set of research work covering a variety of ways to visualize inferred topics.

REFERENCES

[1] Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics*. 13–22.

[2] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.* 6, 1 (2015).

[3] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. 2008. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 8th IEEE International Conference on Data Mining*. IEEE, 3–12.

[4] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. 27–34.

[5] Hazeline U. Asuncion, Arthur U. Asuncion, and Richard N. Taylor. 2010. Software traceability with topic modeling. In *Proceedings of the ACM/IEEE 32nd International Conference on Software Engineering*, Vol. 1. IEEE, 95–104.

[6] D. K. JinYeong Bak and A. Oh. 2012. Distributed online learning for latent Dirichlet allocation. In *Proceedings of the NIPS Workshop on Big Learning*. 1–8.

[7] Parantapa Bhattacharya, Muhammad Bilal Zafar, Niloy Ganguly, Saptarshi Ghosh, and Krishna P. Gummadi. 2014. Inferring user interests in the Twitter social network. In *Proceedings of the 8th ACM Conference on Recommender Systems*. ACM, 357–360.

[8] David M. Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84. DOI:https://doi.org/doi:10.1145/2133806.2133826
[9] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM* 57, 2 (2010), 7. DOI:https://doi.org/10.1145/1667053.1667056
[10] David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 113–120. DOI:https://doi.org/10.1145/1143844.1143859
[11] David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *Ann. Appl. Statist.* (2007), 17–35.
[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, Jan. (2003), 993–1022. DOI:https://doi.org/10.1162/jmlr.2003.3.4-5.993
[13] Jordan Boyd-Graber, David Mimno, and David Newman. 2014. *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*. Vol. 225255. CRC Press, Boca Raton, FL.
[14] Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 103–111.
[15] Stefan Bunk and Ralf Krestel. 2018. WELDA: Enhancing topic models by incorporating local word context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. 293–302.
[16] George Casella and Edward I. George. 1992. Explaining the Gibbs sampler. *Amer. Statist.* 46, 3 (1992), 167–174.
[17] Jonathan Chang. 2012. Collapsed Gibbs sampling methods for topic models. R package: lda (version 1.3.2). http://cran.r-project.org/web/packages/lda/index.html.
[18] Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *Artificial Intelligence and Statistics*. PMLR, 81–88.
[19] Ying-Lang Chang and Jen-Tzung Chien. 2009. Latent Dirichlet learning for document summarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1689–1692. DOI:https://doi.org/10.1109/ICASSP.2009.4959927
[20] Tse-Hsun Chen, Weiyi Shang, Meiyappan Nagappan, Ahmed E. Hassan, and Stephen W. Thomas. 2017. Topic-based software defect explanation. *J. Syst. Softw.* 129 (2017), 79–106. DOI:https://doi.org/10.1016/j.jss.2016.05.015
[21] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. BTM: Topic modeling over short texts. *IEEE Trans. Knowl. Data Eng.* 26, 12 (2014), 2928–2941. DOI:https://doi.org/10.1109/TKDE.2014.2313872
[22] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 74–77. DOI:https://doi.org/10.1145/2254556.2254572
[23] Raphael Cohen, Iddo Aviram, Michael Elhadad, and Noémie Elhadad. 2014. Redundancy-aware topic modeling for patient record notes. *PloS One* 9, 2 (2014), e87555. DOI:https://doi.org/10.1371/journal.pone.0087555
[24] Mário Cordeiro. 2012. Twitter event detection: Combining wavelet analysis and topic inference summarization. In *Doctoral Symposium on Informatics Engineering*. 11–16.
[25] Christopher S. Corley, Kostadin Damevski, and Nicholas A. Kraft. 2020. Changeset-based topic modeling of software repositories. *IEEE Trans. Softw. Eng.* 46, 10 (2020), 1068–1080. DOI : 10.1109/TSE.2018.2874960
[26] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of the Meeting of the Association for Computational Linguistics*. 795–804.
[27] Ali Daud, Juanzi Li, Lizhu Zhou, and Faqir Muhammad. 2010. Knowledge discovery through directed probabilistic topic models: A survey. *Front. Comput. Sci. China* 4, 2 (2010), 280–301.
[28] Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM Workshop on Social Web Search and Mining*. ACM, 57–64.
[29] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. *Commun. ACM* 51, 1 (2008), 107–113.
[30] Stefan Debortoli, Oliver Müller, Iris Junglas, and Jan vom Brocke. 2016. Text mining for information systems researchers: An annotated topic modeling tutorial. *Commun. Assoc. Inf. Syst.* 39, 1 (2016), 7.
[31] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *J. Amer. Soc. Inf. Sci.* 41, 6 (1990), 391. DOI:https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
[32] Mohamed Dermouche, Julien Velcin, Leila Khouas, and Sabine Loudcher. 2014. A joint model for topic-sentiment evolution over time. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'14)*. IEEE, 773–778. DOI:https://doi.org/10.1109/ICDM.2014.82
[33] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545* (2019).
[34] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Ling.* 8 (2020), 439–453.

[35] Tarek Elguebaly and Nizar Bouguila. 2013. Simultaneous Bayesian clustering and feature selection using RJMCMC-based learning of finite generalized Dirichlet mixture models. *Sig. Process.* 93, 6 (2013), 1531–1546.
[36] Katayoun Farrahi and Daniel Gatica-Perez. 2011. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.* 2, 1 (2011), 3.
[37] Xianghua Fu, Kun Yang, Joshua Zhexue Huang, and Laizhong Cui. 2015. Dynamic non-parametric joint sentiment topic mixture model. *Knowl.-based Syst.* 82 (2015), 102–114.
[38] Debasis Ganguly, Manisha Ganguly, Johannes Leveling, and Gareth J. F. Jones. 2013. TopicVis: A GUI for topic-based feedback and navigation. DOI:https://doi.org/10.1145/2484028.2484202
[39] Debasis Ganguly, Johannes Leveling, and Gareth J. F. Jones. 2012. Cross-lingual topical relevance models. DOI:https://doi.org/10.1145/564405.564408
[40] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. 2012. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Trans. Intell. Syst. Technol.* 3, 2 (2012), 23. DOI:https://doi.org/10.1126/science.1178206
[41] Tom Griffiths. 2002. Gibbs sampling in the generative model of latent Dirichlet allocation. DOI:https://doi.org/10.1145/1401890.1401960
[42] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proc. Nat. Acad. Sci.* 101, suppl 1 (2004), 5228–5235. DOI:https://doi.org/10.1073/pnas.0307752101
[43] Loni Hagen. 2018. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Inf. Proc. Manag.* 54, 6 (2018), 1292–1307.
[44] Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 362–370.
[45] Xingwei He, Hua Xu, Jia Li, Liu He, and Linlin Yu. 2017. FastBTM: Reducing the sampling time for biterm topic model. *Knowl.-Based Syst* 132 (2017), 11–20.
[46] Gregor Heinrich. 2008. *Parameter Estimation for Text Analysis*. Technical Report. University of Leipzig. 1–32.
[47] Go Eun Heo, Keun Young Kang, Min Song, and Jeong-Hoon Lee. 2017. Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. *BMC Bioinf* 18, 7 (2017), 251.
[48] Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent Dirichlet allocation. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 856–864. DOI:https://doi.org/10.1.1.187.1883
[49] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 289–296. DOI:https://doi.org/10.1162/jmlr.2003.3.4-5.993
[50] Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.* 42, 1 (2001), 177–196.
[51] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting popular messages in Twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*. ACM, 57–58.
[52] Pengfei Hu, Wenju Liu, Wei Jiang, and Zhanlei Yang. 2014. Latent topic model for audio retrieval. *Pattern Recog.* 47, 3 (2014), 1138–1143. DOI:https://doi.org/10.1016/j.patcog.2013.06.010
[53] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Mach. Learn.* 95, 3 (2014), 423–469.
[54] Dongping Huang, Shuyu Hu, Yi Cai, and Huaqing Min. 2014. Discovering event evolution graphs based on news articles relationships. In *Proceedings of the IEEE 11th International Conference on e-Business Engineering (ICEBE'14)*. IEEE, 246–251.
[55] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools. Applic.* 78, 11 (2019), 15169–15211.
[56] Do-Heon Jeong and Min Song. 2014. Time gap analysis by the topic model-based temporal technique. *J. Informet.* 8, 3 (2014), 776–790. DOI:https://doi.org/10.1016/j.joi.2014.07.005
[57] Di Jiang, Yongxin Tong, and Yuanfeng Song. 2016. Cross-lingual topic discovery from multilingual search engine query log. *ACM Trans. Inf. Syst.* 35, 2 (2016), 9.
[58] Efsun Sarioglu Kayi, Kabir Yadav, James M. Chamberlain, and Hyeong-Ah Choi. 2017. Topic modeling for classification of clinical reports. *arXiv preprint arXiv:1706.06177* (2017).
[59] Muhammad Taimoor Khan, Mehr Durrani, Shehzad Khalid, and Furqan Aziz. 2016. Online knowledge-based model for big data topic extraction. *Comput. Intell. Neurosci.* DOI:https://doi.org/10.1155/2016/6081804
[60] Milad Kharratzadeh, Benjamin Renard, and Mark J. Coates. 2015. Bayesian topic model approaches to online and time-dependent clustering. *Dig. Sig. Process.* 47 (2015), 25–35. DOI:https://doi.org/10.1016/j.dsp.2015.03.010

[61] Dongwoo Kim and Alice Oh. 2011. Accounting for data dependencies within a hierarchical Dirichlet process mixture model. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, 873–878.
[62] Dongwoo Kim and Alice Oh. 2011. Topic chains for understanding a news corpus. *Comput. Ling. Intell. Text Process.*. DOI:https://doi.org/10.1007/978-3-642-19437-5_13
[63] Dongwoo Kim and Alice Oh. 2014. Hierarchical Dirichlet scaling process. In *Proceedings of the International Conference on Machine Learning*. 973–981.
[64] Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. 2012. Modeling topic hierarchies with the recursive Chinese restaurant process. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. 783–792. DOI:https://doi.org/10.1145/2396761.2396861
[65] Younghoon Kim and Kyuseok Shim. 2014. TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Inf. Syst.* 42 (2014), 59–77. DOI:https://doi.org/10.1016/j.is.2013.11.003
[66] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.
[67] Julian F. P. Kooij, Gwenn Englebienne, and Dariu M. Gavrila. 2015. Identifying multiple objects from their appearance in inaccurate detections. *Comput. Vis. Image Underst.* 136 (2015), 103–116.
[68] Guy Lansley and Paul A. Longley. 2016. The geography of Twitter topics in London. *Comput. Environ. Urb. Syst.* 58 (2016), 85–96. DOI:https://doi.org/10.1016/j.compenvurbsys.2016.04.002
[69] Jey Han Lau and Timothy Baldwin. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 483–487.
[70] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 530–539.
[71] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 497–506.
[72] Chenliang Li, Yu Duan, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2017. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Trans. Inf. Syst.* 36, 2 (2017), 11.
[73] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 165–174.
[74] Weifeng Li, Junming Yin, and Hsinchsun Chen. 2017. Supervised topic modeling using hierarchical Dirichlet process-based inverse regression: Experiments on e-commerce applications. *IEEE Trans. Knowl. Data Eng.* 30, 6 (2017), 1192–1205.
[75] Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong Cheng. 2014. The dual-sparse topic model: Mining focused topics and focused terms in short text. In *Proceedings of the 23rd International Conference on World Wide Web*. 539–550.
[76] Erik Linstead, Paul Rigor, Sushil Bajracharya, Cristina Lopes, and Pierre Baldi. 2007. Mining concepts from code with probabilistic topic models. In *Proceedings of the 22nd IEEE/ACM International Conference on Automated Software Engineering*. ACM, 461–464.
[77] Jun S. Liu. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* 89, 427 (1994), 958–966.
[78] Shuhua Liu and Patrick Jansson. 2017. Topic Modelling Analysis of Instagram Data for the Greater Helsinki Region.
[79] Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. 2015. Multilingual topic models for bilingual dictionary extraction. *ACM Trans. Asian Low-resour. Lang. Inf. Process.* 14, 3 (2015), 11.
[80] Xiao Liu, Mingli Song, Qi Zhao, Dacheng Tao, Chun Chen, and Jiajun Bu. 2012. Attribute-restricted latent topic model for person re-identification. *Pattern Recog.* 45, 12 (2012), 4204–4213.
[81] Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun. 2011. PLDA+: Parallel latent Dirichlet allocation with data placement and pipeline processing. *ACM Trans. Intell. Syst. Technol.* 2, 3 (2011), 26.
[82] Kun Lu and Dietmar Wolfram. 2012. Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *J. Amer. Soc. Inf. Sci. Technol.* 63, 10 (2012), 1973–1986.
[83] Zhiwu Lu and Yuxin Peng. 2013. Latent semantic learning with structured sparse representation for human action recognition. *Pattern Recog.* 46, 7 (2013), 1799–1809. DOI:https://doi.org/10.1016/j.patcog.2012.09.027
[84] Stacy K. Lukins, Nicholas A. Kraft, and Letha H. Etzkorn. 2010. Bug localization using latent Dirichlet allocation. *Inf. Softw. Technol.* 52, 9 (2010), 972–990.
[85] Minnan Luo, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander Hauptmann, and Qinghua Zheng. 2017. Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[86] Baizhang Ma, Dongsong Zhang, Zhijun Yan, and Taeha Kim. 2013. An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews. *J. Electron. Commer. Res.* 14, 4 (2013), 304. DOI:https://doi.org/10.1016/j.im.2015.02.002

[87] Hui-Fang Ma. 2011. Hot topic extraction using time window. In *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC'11)*. IEEE, 56–60.

[88] Masoud Makrehchi. 2011. Social link recommendation by learning hidden topics. In *Proceedings of the 5th ACM Conference on Recommender Systems*. ACM, 189–196.

[89] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H. Byers. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[90] Jon D. Mcauliffe and David M. Blei. 2008. Supervised topic models. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 121–128.

[91] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. (2002). Retrieved from http://mallet.cs.umass.edu.

[92] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, 171–180.

[93] David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 500–509.

[94] David Mimno and Andrew McCallum. 2007. Organizing the OCA: Learning faceted subjects from a library of digital books. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, 376–385.

[95] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 880–889. DOI:https://doi.org/10.3115/1699571.1699627

[96] Christopher E. Moody. 2016. Mixing Dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019* (2016).

[97] Gordon E. Moon, Israt Nisa, Aravind Sukumaran-Rajam, Bortik Bandyopadhyay, Srinivasan Parthasarathy, and P. Sadayappan. 2018. Parallel latent Dirichlet allocation on GPUs. In *Proceedings of the International Conference on Computational Science*. Springer, 259–272.

[98] N. K. Nagwani. 2015. Summarizing large text collection using topic modeling and clustering based on MapReduce framework. *J. Big Data* 2, 1 (2015), 6.

[99] Ramesh Nallapati, William Cohen, and John Lafferty. 2007. Parallelized variational EM for latent Dirichlet allocation: An experimental evaluation of speed and scalability. In *Proceedings of the International Conference on Data Mining Workshops (ICDMW'07)*. IEEE, 349–354.

[100] David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *J. Mach. Learn. Res.* 10, Aug. (2009), 1801–1828.

[101] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 100–108.

[102] David Newman, Padhraic Smyth, and Mark Steyvers. 2006. Scalable parallel topic models. *J. Intell. Commun. Res. Devel.* 5 (2006). DOI:https://doi.org/10.7551/mitpress/9486.003.0011

[103] David Newman, Padhraic Smyth, Max Welling, and Arthur U. Asuncion. 2008. Distributed inference for latent Dirichlet allocation. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 1081–1088.

[104] Zhenxing Niu, Gang Hua, Le Wang, and Xinbo Gao. 2017. Knowledge-based topic model for unsupervised object discovery and localization. *IEEE Trans. Image Process.* 27, 1 (2017), 50–63.

[105] Michael J. Paul and Mark Dredze. 2014. Discovering health topics in social media using topic models. *PloS One* 9, 8 (2014), e103408. DOI:https://doi.org/10.1371/journal.pone.0103408

[106] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825–2830.

[107] Nanyun Peng, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics*. 674–679.

[108] James Petterson, Wray Buntine, Shravan M. Narayanamurthy, Tibério S. Caetano, and Alex J. Smola. 2010. Word features for latent Dirichlet allocation. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 1921–1929.

[109] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 569–577.

[110] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*.
[111] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*.
[112] Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*.
[113] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 248–256.
[114] Daniel Ramage, Christopher D. Manning, and Susan Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 457–465. DOI:https://doi.org/10.1145/2020408.2020481
[115] Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*. ELRA, 45–50.
[116] Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney. 2010. Spherical topic models. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*. 903–910. DOI:https://doi.org/10.1007/s10955-009-9892-0
[117] Yafeng Ren, Ruimin Wang, and Donghong Ji. 2016. A topic-enhanced word embedding for Twitter sentiment classification. *Inf. Sci.* 369 (2016), 188–198.
[118] Philip Resnik and Eric Hardisty. 2010. *Gibbs sampling for the uninitiated*. Maryland Univ College Park Inst for Advanced Computer Studies.
[119] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. Citeseer, 3806–3813.
[120] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 487–494. DOI:https://doi.org/10.1016/j.nima.2010.11.062
[121] Karim Sayadi, Quang Vu Bui, and Marc Bui. 2016. Distributed implementation of the latent Dirichlet allocation on Spark. In *Proceedings of the 7th Symposium on Information and Communication Technology*. ACM, 92–98.
[122] Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 432–436.
[123] Karthick Seshadri, S. Mercy Shalinie, and Chidambaram Kollengode. 2015. Design and evaluation of a parallel algorithm for inferring topic hierarchies. *Inf. Proc. Manag.* 51, 5 (2015), 662–676. DOI:https://doi.org/10.1016/j.ipm.2015.06.006
[124] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 63–70.
[125] Bradley Skaggs and Lise Getoor. 2014. Topic modeling for Wikipedia link disambiguation. *ACM Trans. Inf. Syst.* 32, 3 (2014), 10.
[126] Alison Smith, Jason Chuang, Yuening Hu, Jordan Boyd-Graber, and Leah Findlater. 2014. Concurrent visualization of relationships between words and topics in topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 79–82.
[127] Alexander Smola and Shravan Narayanamurthy. 2010. An architecture for parallel topic models. *Proc. VLDB Endow.* 3, 1-2 (2010), 703–710.
[128] Padhraic Smyth, Max Welling, and Arthur U. Asuncion. 2009. Asynchronous distributed learning of topic models. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 81–88.
[129] Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handb. Latent Semant. Anal.* 427, 7 (2007), 424–440.
[130] Xiaobing Sun, Bixin Li, Hareton Leung, Bin Li, and Yun Li. 2015. MSR4SM: Using topic models to effectively mining software repositories for software maintenance tasks. *Inf. Softw. Technol.* 66 (2015), 1–12. DOI:https://doi.org/10.1016/j.infsof.2015.05.003
[131] Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2005. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 1385–1392.
[132] Rajeev Thakur, Rolf Rabenseifner, and William Gropp. 2005. Optimization of collective communication operations in MPICH. *Int. J. High Perf. Comput. Applic.* 19, 1 (2005), 49–66.
[133] Stephen W. Thomas, Bram Adams, Ahmed E. Hassan, and Dorothea Blostein. 2014. Studying software evolution using topic models. *Sci. Comput. Prog.* 80 (2014), 457–479. DOI:https://doi.org/10.1016/j.scico.2012.08.003

[134] Kai Tian, Meghan Revelle, and Denys Poshyvanyk. 2009. Using latent Dirichlet allocation for automatic categorization of software. In *Proceedings of the 6th IEEE International Working Conference on Mining Software Repositories*. IEEE, 163–166.

[135] Zhongyuan Tian, Harumichi Yokoyama, and Takuya Araki. 2019. Parallel latent Dirichlet allocation using vector processors. In *Proceedings of the IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 1548–1555.

[136] Calin Rares Turliuc, Luke Dickens, Alessandra Russo, and Krysia Broda. 2016. Probabilistic abductive logic programming using Dirichlet priors. *Int. J. Approx. Reas.* 78 (2016), 223–240. DOI:https://doi.org/10.1016/j.ijar.2016.07.001

[137] Duc-Thuan Vo and Cheol-Young Ock. 2015. Learning to classify short text from scientific documents using topic models with various types of knowledge. *Exp. Syst. Applic.* 42, 3 (2015), 1684–1698. DOI:https://doi.org/10.1016/j.eswa.2014.09.031

[138] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. 2015. BigARTM: Open source library for regularized multimodal topic modeling of large collections. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts*. Springer, 370–381.

[139] Konstantin Vorontsov and Anna Potapenko. 2015. Additive regularization of topic models. *Mach. Learn.* 101, 1–3 (2015), 303–323.

[140] Nicholas Vretos, Nikos Nikolaidis, and Ioannis Pitas. 2012. Video fingerprinting using latent Dirichlet allocation and facial images. *Pattern Recog.* 45, 7 (2012), 2489–2498. DOI:https://doi.org/10.1016/j.patcog.2011.12.022

[141] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2013. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Inf. Retr.* 16, 3 (2013), 331–368. DOI:https://doi.org/10.1007/s10791-012-9200-5

[142] Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Inf. Proc. Manag.* 51, 1 (2015), 111–147. DOI:https://doi.org/10.1016/j.ipm.2014.08.003

[143] Martin J. Wainwright, Michael I. Jordan et al. 2008. Graphical models, exponential families, and variational inference. *Found. Trends® Mach. Learn.* 1, 1–2 (2008), 1–305.

[144] Hanna M Wallach. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 977–984.

[145] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning*. 1105–1112.

[146] Chong Wang, David Blei, and David Heckerman. 2012. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298* (2012).

[147] Di Wang and Ahmad Al-Rubaie. 2015. Incremental learning with partial-supervision based on hierarchical Dirichlet process and the application for document classification. *Appl. Soft Comput.* 33 (2015), 250–262. DOI:https://doi.org/10.1016/j.asoc.2015.04.044

[148] Jin Wang, Xiangping Sun, Mary F. H. She, Abbas Kouzani, and Saeid Nahavandi. 2013. Unsupervised mining of long time series based on latent topic model. *Neurocomputing* 103 (2013), 93–103. DOI:https://doi.org/10.1016/j.neucom.2012.09.008

[149] Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 424–433.

[150] Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM'07)*. IEEE, 697–702.

[151] Xiang Wang, Kai Zhang, Xiaoming Jin, and Dou Shen. 2009. Mining common topics from multiple asynchronous text streams. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*. ACM, 192–201.

[152] Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang. 2009. PLDA: Parallel latent Dirichlet allocation for large-scale applications. In *Proceedings of the International Conference on Algorithmic Applications in Management*. 301–314. DOI:https://doi.org/10.1007/978-3-642-02158-9_26

[153] Yu Wang, Jiebo Luo, Richard Niemi, Yuncheng Li, and Tianran Hu. 2016. Catching fire via "likes": Inferring topic preferences of Trump followers on Twitter. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*.

[154] Yi Wang, Xuemin Zhao, Zhenlong Sun, Hao Yan, Lifeng Wang, Zhihui Jin, Liubin Wang, Yang Gao, Jia Zeng, Qiang Yang et al. 2014. Towards topic modeling for big data. *arXiv preprint arXiv:1405.4402* (2014).

[155] Lino Wehrheim. 2019. Economic history goes digital: Topic modeling the journal of economic history. *Cliometrica* 13, 1 (2019), 83–125.

[156] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: Finding topic-sensitive influential Twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. ACM, 261–270.

[157] Erik Wiener, Jan O. Pedersen, Andreas S. Weigend, et al. 1995. A neural network approach to topic spotting. In *Proceedings of the 4th Symposium on Document Analysis and Information Retrieval*.

[158] Andrew T. Wilson and Peter A. Chew. 2010. Term weighting schemes for latent Dirichlet allocation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. 465–473. DOI:https://doi.org/1857999.1858069

[159] Yueshen Xu, Jianwei Yin, Jianbin Huang, and Yuyu Yin. 2018. Hierarchical topic modeling with automatic knowledge mining. *Exp. Syst. Applic.* 103 (2018), 106–117.

[160] Yueshen Xu, Yuyu Yin, and Jianwei Yin. 2017. Tackling topic general words in topic modeling. *Eng. Applic. Artif. Intell.* 62 (2017), 124–133.

[161] Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A correlated topic model using word embeddings. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 4207–4213.

[162] Feng Yan, Ningyi Xu, and Yuan Qi. 2009. Parallel inference for latent Dirichlet allocation on graphics processing units. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 2134–2142.

[163] Shuang Yang, Chunfeng Yuan, Weiming Hu, and Xinmiao Ding. 2014. A hierarchical model based on latent Dirichlet allocation for action recognition. In *Proceedings of the 22nd International Conference on Pattern Recognition*. IEEE, 2613–2618. DOI:https://doi.org/10.1109/ICPR.2014.451

[164] Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2019. A multilingual topic model for learning weighted topic links across corpora with low comparability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. 1243–1248.

[165] Yi Yang, Doug Downey, and Jordan Boyd-Graber. 2015. Efficient methods for incorporating knowledge into topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 308–317.

[166] Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 937–946. DOI:https://doi.org/10.1145/1557019.1557121

[167] Liang Yao, Yin Zhang, Baogang Wei, Lei Li, Fei Wu, Peng Zhang, and Yali Bian. 2016. Concept over time: the combination of probabilistic topic model with wikipedia knowledge. *Exp. Syst. Applic.* 60 (2016), 27–38.

[168] Chyi-Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen. 2014. Clustering scientific documents with topic modeling. *Scientometrics* 100, 3 (2014), 767–786.

[169] Hsiang-Fu Yu, Cho-Jui Hsieh, Hyokun Yun, S. V. N. Vishwanathan, and Inderjit S. Dhillon. 2015. A scalable asynchronous distributed algorithm for topic modeling. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1340–1350.

[170] Bo Yuan, Xinbo Gao, Zhenxing Niu, and Qi Tian. 2019. Discovering latent topics by Gaussian latent Dirichlet allocation and spectral clustering. *ACM Trans. Multimedia Comput. Commun. Applic.* 15, 1 (2019), 25.

[171] Lele Yut, Ce Zhang, Yingxia Shao, and Bin Cui. 2017. LDA* a robust and large-scale topic modeling system. *Proc. VLDB Endow.* 10, 11 (2017), 1406–1417.

[172] Manzil Zaheer, Amr Ahmed, and Alexander J. Smola. 2017. Latent LSTM allocation joint clustering and non-linear dynamic modeling of sequential data. In *Proceedings of the 34th International Conference on Machine Learning*. JMLR.org, 3967–3976.

[173] Jianping Zeng, Jiangjiao Duan, Wenjun Cao, and Chengrong Wu. 2012. Topics modeling based on selective Zipf distribution. *Exp. Syst. Applic.* 39, 7 (2012), 6541–6546. DOI:https://doi.org/10.1016/j.eswa.2011.12.051

[174] Ke Zhai and Jordan Boyd-Graber. 2013. Online latent Dirichlet allocation with infinite vocabulary. In *Proceedings of the International Conference on Machine Learning*. 561–569.

[175] Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L. Alkhouja. 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 879–888. DOI:https://doi.org/10.1145/2187836.2187955

[176] Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. 2010. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1079–1088.

[177] Tao Zhang, Kang Liu, Jun Zhao, et al. 2013. Cross lingual entity linking with bilingual topic model. *Proceedings of the International Joint Conference on Artificial Intelligence*. 2218–2224.

[178] Bing Zhao and Eric P. Xing. 2006. BiTAM: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*. Association for Computational Linguistics, 969–976.

[179] Bing Zhao and Eric P. Xing. 2007. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. *Advances in Neural Information Processing Systems* 20 (2007), 1689–1696.

[180] Feng Zhao, Yajun Zhu, Hai Jin, and Laurence T. Yang. 2016. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. *Fut. Gen. Comput. Syst.* 65 (2016), 196–206.
[181] Huasha Zhao, Biye Jiang, John F. Canny, and Bobby Jaros. 2015. Same but different: Fast and high quality Gibbs parameter estimation. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1495–1502.
[182] Wenjun Zhu, Liqing Zhang, and Qianwei Bian. 2012. A hierarchical latent topic model based on sparse coding. *Neurocomputing* 76, 1 (2012), 28–35. DOI:https://doi.org/10.1016/j.neucom.2010.11.038
[183] Elaine Zosa and Mark Granroth-Wilding. 2019. Multilingual dynamic topic model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'19)*. 1388–1396.
[184] Jialing Zou, Qixiang Ye, Yanting Cui, Fang Wan, Kun Fu, and Jianbin Jiao. 2016. Collective motion pattern inference via locally consistent latent Dirichlet allocation. *Neurocomputing* 184 (2016), 221–231. DOI:https://doi.org/10.1016/j.neucom.2015.08.108
[185] Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2105–2114.

Received April 2020; revised March 2021; accepted April 2021