A Detailed Survey on Topic Modeling for Document and Short Text Data

S. Likhitha
Department of Information Science & Engineering
JSS Science & Technology University, Mysuru, India

B. S. Harish
Department of Information Science & Engineering
JSS Science & Technology University, Mysuru, India

H. M. Keerthi Kumar
JSSRF, JSS TI Campus
Mysuru, India

ABSTRACT
Text mining is one of the most significant field in the digital era due to the rapid growth of textual information. Topic models are gaining popularity in the last few years. A topic comprises of a group of words that are often take place together. Topic models are better performing techniques to extract semantic knowledge presented in the data. The various methods used for topic models are, LSA (Latent Semantic Analysis), PLSA (Probabilistic Latent Semantic Analysis), LDA (Latent Dirichlet Allocation). These methods gained popularity in extracting hidden themes from the document (corpus). Various topic modeling algorithms are developed to inquiry, summarize and extract hidden semantic structures of large corpus. In this paper, we present a detailed survey covering the various topic modeling techniques proposed in last decade. Additionally, we focus on different strategies of extracting the topics in social media text, where the goal is to find and aggregate the topic within short texts. Further, we summarize the various applications and quantitative evaluation of the various methods, with statistical and mathematical knowledge to predict the convergence of results.

Keywords
Text ining, Topic Modeling, Short Text, Semantic Analysis.

1. INTRODUCTION
From the last two decades, the massive amount of digital information growing rapidly over the internet due to advancement of computer science technology. The available digital information is in the form of text like web-pages, emails, study materials etc. Hence there is a need of tool to search, organize and understand the huge quantity of digital information. Topic modeling is a technique used to understand, summarize and organize the vast amount of textual data. Topic modeling is one of the major applications in Text mining [1][2]. Text mining is a multidisciplinary field, involving different aspects of information retrieval, text analysis, information extraction, clustering [3], categorization [4][5], visualization, database technology, and topic modeling [6]. Topic models are probabilistic statistical models that uncover the hidden thematic structure in document collections and provides a simple way to analyze large volumes of unlabeled text [2]. The main aim of the topic modeling is to find out the patterns of words in text and discover hidden structural words that runs through corpus by analyzing different pattern present in words. The topics contain the group (cluster) of words, which frequently occur together in the corpus [1]. In topic models, words are represented as a mixture of the topic and each topic is a probability distribution over the words in vocabulary. The model builds the relationships between the vocabulary and the documents through the thematic structure and learns the structural relationships. Topic models are made used in document as well as short text data analysis. The survey summarizes the topic model application for both document and short text data. The document data are filled with long text data, summarizing and finding latent structure through finding the terms in the corpus. Mapping the large set of words by analyzing the pattern or frequency of co-occurrences words based on its probability distribution of words in the corpus [1]. The Micro blogs have become an important platform for people to share instant messages, online publishing, and acquire knowledge. Short texts are popular in today's web, especially with the emergence of social media where people share their ideas and opinions [7]. Compared to short text data, document data will have less difficulties in analyzing the topics because document doesn't possess short form of words. The short text data are filled with more short form words which is more sparse, noisy words and non-informative words. Unlike long text, a challenging task in short text data is inferring topics, which contains more short form words and becomes a crucial task for resulting accurate topics. To overcome these problems, one best approach for discovering the thematic structure from the short text is Topic Modeling [8].

Phan et al., [8] proposed a framework to focus on the combination of subject's topics and feature fundamental topics of the short and document reports, helping to survive challenges like: synonyms, hyponyms, vocabulary mismatch, for better characterization, grouping, coordinating, and positioning. The classical problem is to represent words. To overcome these basic problem Latent Semantic Indexing (LSI) was proposed, which uses the linear algebra method to map the latent topics by performing matrix decomposition using Singular Value Decomposition (SVD) [9].

Latent Semantic Analysis (LSA) was proposed [10] to automate the analysis of the indexing words. The core idea is vectors-based representation of hidden semantic context using SVD and to leverage the context words to capture the hidden concepts. LSA lacks in handling different significance and polysemy words, unfortunately it fails to update the new document immediately [10]. To overcome the drawbacks of LSA, Hofmann [11] introduced a Probabilistic Latent Semantic Analysis (PLSA) in the year 1999. The idea is to find a probabilistic model to generate hidden topics with observed variable with respect to contexts in the corpus with the usage of a dictionary words. Mainly two formulations have overcome by PLSA. The first formulation is symmetric formulation model by applying the Bayes' standard to transform the restrictive likelihood and next followed by asymmetric formulation model with respect to joint distribution of terms obtained by summing over all possible realizations of corpus. Finally, the model chose the latent topic based on the distribution of co-occurrence words using

conditional probability [12]. PLSA is more adaptable than LSA to demonstrate yet at the same time it has significant drawbacks. Since PLSA has no prior parameters like alpha and beta to model word distributions in corpus, it also results in over fitting when number of documents are increased linearly [13]. To address the issues with PLSA and LSA, Beli et al., in the year 2003 proposed a novel LDA model, which is statistical (Bayesian) probability model that generates document-topic and word-topic distribution with the usage of Dirichlet priors such as alpha and beta. These priors are hyper-parameters which are used to estimate document-topics density and topic-word density respectively [14]. LDA model tries to capture the hidden semantic structure with its correlated words based on the distribution of topics over vocabulary in the observed documents [4]. Conventional methods like LDA and PLSA are unsuccessful for short text, due to lack of words in the text. Straightforwardly using these models elevates to sparse topic modeling for the short text data. Researchers have developed a model based on the strategy for leaning the observed words in short corpus [4], [15] [16]. To hypothesize the issues of conventional topic model for the short text data, one popular topic modeling algorithm is used i.e. Dirichlet Multinomial Mixture Model (DMM), an apparently less prevalent topic model, expect that a corpus can just have a single topic, which seems to be more progressively fitting supposition for short content [17]. Baseline topic models uses Gibbs sampling method to infer the topics. It is one of the sampling conditional algorithm which uses Markov Chain Monte Carlo (MCMC). Gibbs sampler works by distributions of variables, posterior distribution as its target distribution to identify the topic in the documents [9] [18] [19].

Earlier reviews are concentrated on various baseline models and different applications of the topic model using multimedia tools [6] [20] [21]. However, this survey work differs with other existing literatures. In this survey we focused on the amount of work made with short text as well documents data. Especially focused on last one decade works done with micro blog data.

This survey work presents the following objectives.

- To focus on short text information, which is the most prominent information on the web and distinctive procedures used for the portrayal of words.
- Listing diverse datasets that are worked with various challenges in short content as well as documents.
- Various topic modeling approaches with modification of parameters and with different representation forms.
- Explored various challenges and applications related to topic modeling.
- Lastly, the overview of topic modeling methods with its evaluation metrics.

2. PRELIMINARY STEPS OF TOPIC MODELING

The data available in Internet is vast, being unstructured like audio, text, video, speech presents significant research disputes. To manage such unstructured text information with its semantic meanings, some NLP methods i.e. extraction of information and informational retrieval is available. For the purpose of extracted texts, researchers are made efforts in recent days to automate and retrieve the topics in texts based on semantic structures. Topic models holds significant interests in information retrieval for several diverse formats [2]. The common sub-task required for topic models are explained in the below sections.

2.1 Data Acquisition

Due to rapid growth of various online resources, data acquirement is extremely conceptual. Data acquisition is highly subjective to the sort of media, kind of analysis need to perform, data supported with various media. Analysis is made to perform with different micro-blogging and documents data. Data are collected from various media like: twitter, question and answers collections, web searches, news groups, reuters, various scientific documents, public data from different sites and so on. These data are pre-processed with various techniques to remove unwanted information in unstructured data [22] [23] [24].

2.2 Pre-processing

The data acquired from various sites need to pre-process to remove the unnecessary irrelevant and redundant features [25]. Some preprocessing steps performed for topic modeling are: convert letters to lowercase, removal of stop words and all non-alphabetic characters, tokenization, lemmatization, stemming, parts of speech tagging, hashtags removal etc. Tokenization is the process of partitioning the text into terms (features), called tokens. In lower-case conversion, converting whole tokens into lower-case form. During punctuation removal, removing all special and non-alphabetic characters (~, `, @, #, $, %, ^, &, <, >,",', ;). Further, lemmatization is the process of grouping together the different inflected forms of word so that can be analyzed as a single item. Stemming is performed to find its root words. POS tagging is processed to identify various parts of speech present in the text data.

2.3 Feature extraction and Selection

Features extraction plays more essential role to identify the relevant patterns and obtain information. Selecting these features helps in dimensionality reduction to enhance the performance of text mining. The extracted and selected features are represented in vector form. Vector representation is transforming meaningful words with context terms represented as number, where each word is distributed with weights [26]. Various forms of extracting words are: Frequency based Embedding [27], Prediction based Embedding [28], Non-Negative Matrix Factorization [29]. TF-IDF (Term-Frequency – Inverse document frequency) [30], Count Vector, Co-occurrence vector and document frequency. Word2vector is NLP (Natural Language Processing) based technique which is the combination of CBOW (Continuous Bag of words) and skip-gram model and is used for prediction based on the sense of finding probabilities of words.

2.4 Title and Authors

The title (Helvetica 18-point bold), authors' names (Helvetica 12-point) and affiliations (Helvetica 10-point) run across the full width of the page – one column wide. We also recommend e-mail address (Helvetica 12-point). See the top of this page for three addresses. If only one address is needed, center all address text. For two addresses, use two centered tabs, and so on. For three authors, you may have to improvise.

Table 1. Distribution of short text topic modeling articles based on the representation of words

Text Representation	# No. of Articles in last decade	#Articles Reference
Frequency of Co-occurrence Words	7	[5],[27],[31-36]
Pseudo documents	6	[22],[23],[36-39]
Word weighting	6	[26],[40-44]
Word Embedding	15	[31],[22],[26],[28][45-56]
Sentence Level	1	[57]
Hash tags	5	[23],[43],[46],[58],[59]

3. RELATED WORKS

In this section, we cover the contributions of topic models used in short texts and document dataset. Many researches proposed new ideas to model the topics in the texts data. The existing various methods for documents and short texts data are discussed in section 4.1 and 4.2 respectively.

3.1 Topic model on documents

Topics are built by the user's activity. These topics are extracted from the documents based on the distributions of words. Kawamae, [60] proposed a novel approach to read each document which is bundled and viewed on author interest and discover the semantic pattern. Schneider and Vlachos [24], worked on an approach that connects each document to retrieve the topics using aspect models and word occurrence methods which builds keywords and context.

In contrast with the frequency of co-occurrence of words to infer the topics, Salerno et al., [33] proposed Word Community Allocation (WCA) model to develop a cluster of words from documents and vocabulary words are expressed with weights based on contextual behavior of co-event words. Clinchant and Perronnin [61] introduced a representation form for continuous words. Each document represented as BoEW (Bag-of-Embedded-Words). It uses a non-linear mapping technique for vector representation and then aggregates to a large corpus.

To improvise the LDA model, AlSumait [16] proposed a model called OLDA (Online Latent Dirichlet Allocation) which focus on the immediate updating of new documents and compared it with existing documents for generating a sequence of patterns by analyzing a fraction of data. Traditional LDA model was assigned with equal weights to all the words [14]. Hence ranking each word also plays a important role in the selection of meaningful words by weighting each term [42]. Reisinger et al. proposed TF-IDF (Term Frequency and Inverse –Document Frequency) for Spherical Topic Model (STM) for feature extraction [48]. Kim et al., [62] proposed a generative approach called Entity Topic Model (ETM) for large corpus collection. It uses prior knowledge to build the model entity relation indirectly using its repetitive arrangements of co-occurring words. Jadhav et al., [63] presented a novel way, Pattern Based Topic Model (PBTM) which enriches the quality representation of patterns based on user's activity and semantic representation of words. Lee et al., [40] proposed Weighted Topic Model (WTM) and Balance Weighted Topic Model (BWTM) approaches for extracting the features in the corpus using IDF method. WTM concentrated on weighing all behaving words and resulting in the low parameter. Topic distributions of words increases its iterations. Efficiency was decreased in resulting high probability of topics. To achieve the efficiency BWTM used to manage more weights for specific words. Fewer weights are assigned for unspecific words. Kai et al., [41] introduced a model TWLDA (Term Weighting LDA) using various supervised and unsupervised weighting schemes. Supervised weigh schemes use prior information from pre-defined categories. Unsupervised weighting schemes uses term occurrence factors with its terms. The model results in assigning low weights to words with low topic distribution. Wilson and Chew [42] also concentrated on weighting terms using LDA as a baseline method.

The important part of topic modeling is its semantic relations. Wang et al., [43] proposed a model SHTTM (Semantic Hashing using Tags and Topic Modeling) which coordinates data labels and the likelihood of the words in the corpus. It also dependent on comparability connection of hashing codes to bunch the semantic subjects. Jiang et al., [57] build an Associated Topic Model (ATM), in which sequential sentences are viewed as imperative. Yao et al., [19] proposed a model Sparse LDA (SLDA), which appropriates information structure that considerably improves sampling execution speed.

Smoothing parameters helps in taking care of the dispersion of subjects expand on words. Yi and Allan [18] make sense of the ground-breaking model for smoothing the appropriation of themes and its similarity between the different models. Execution of results are accumulated for various theme display like MU (Mixture of Unigrams), LDA (Latent Dirichlet Allocation) [14] [2], PAM (Pachinko Allocation Model) [64], RM (Relevance Modeling) [65]. Retrieving topics are enhanced by smoothing query in each corpus.

3.2 Topic modeling on short texts

The growth of internet popularity has resulted in increasing emerge of short text data. Short text data emerges with more sparse information compared to documents [66] [67]. Handling the short text data, which contains few words is one of the well-known challenging tasks [8] [7]. Co-occurrence of words is a common strategy used for retrieving topics from the corpus. Pedrosa et al., [36] proposed a novel method to develop pseudo-document structure by aggregating the co-occurring frequency of words which are well-organized in the corpus.

Word embedding concept created a trend for representation of words in the vector form which creates a more meaningful way of representing the words. Geeganage and Tharanga [56] proposed an advanced model based on semantic concept retrieval to obtain more appropriate topics using domain-specific information on meaningful words in the context. Xun et al., [28] proposed a novel way to find more desirable topic representation. The model worked on using multivariate Gaussian approaches by integrating training words with Continuous Bag of Words (CBOW) embedded in vector space. It uses Wikipedia source to study the relationship between topic and context using its lexical relations with situational words in the text corpus. Jiang et al., [39] introduced Biterm Pseudo Document Topic Model (BPDTM), which handles all co-occurrence words and biterms. Further aggregated long pseudo-documents are balanced with its semantic relationship. Lu et al., [31] proposed a model to work on similarity words and built GloSS-BTM (Global Semantically Similar Biterms Topic Model) recognizing the

global biterms and formulate semantic relation between global biterms and co-occurrence words in the corpus to build semantic relations using word embeddings. In contrast with BTM model, Hu et al., [68] proposed an approach for streaming short text data to classify the topics. They used OBTM (Online Bit term Topic Model) [35] for enriching the short text extensions and concept drift detection [69]. Wandabwa et al., [70] proposed Metamodel Enabled Latent Dirichlet Association (MELDA) which assumes that prior information is used for learning the domain interest topics. Seed topic parameter is used to learn uneven topics distribution using seed-topic and formal topic sets.

Li et al., [44] proposed a novel way to build a model with CEW (Combinational form of Entropy Weighting) on integrating entropy weighting schemes like BDC (Balanced Distributional Concentration) [71] and logarithm frequency [43] weighting methods. It allocates more weights to informative words and fewer weights to general specific words (one word! colloquial words) and balances the semantically meaning words with higher weights. By using of Author-Topic Model (ATM) concepts used in lengthy documents, Rajani et al., [46] proposed a model called Author-Recipient-Topic (ART) model for tweets. ART works with similar tweets on users' interest, hashtags, users' recipients, which are united in a single document. Shi et al., [72] construct a user-interest model by merging the short corpus to lengthy corpus and finds the topic relations based on each user events and their similarity. Using a baseline of LDA and Author-Topic Model [23], Tsai [58] also worked with inclined tags to extract similar topics from blogs. Here tags serve as a probability distribution over themes. Wang et al., [59] proposed a model called HGTM (Hashtag Graph-based Topic Model) to reduce sparse and noise. HGTM uses tags relational knowledge to analyze the semantic relations between users and their hashtags to develop more reliable topics even if words are unavailable in the specific tweets. Instead of using only hashtags, Sapul et al., [73] introduced a CLOPE algorithm, a logic of using hashtags with each keyword in a text which extracts more accurate feature and results with more relevant topics based on the available feature words. Qiu et al., [67] proposed a Dynamic Social Network Topic Model (DSM) to group cluster based on topics scattered by users' interest and also their connectivity with social networks. Zheng et al., [74] introduced a novel way, to conjugate definitions proportion for topics and terms. A virtual term frequency is added to words which are not appeared in micro blog texts. Once semantic smoothing is done, information points are then mapped with the original records. Another major issue with social media short text is filled with meaningless and colloquial words. Filtering out these noisy words is a challenging task. Li et al., [75] proposed a novel method to introduce a common topics scheme to gather noisy words. Common topics collects noise words like rt (Retweet) in tweets and Colloquial words like lol etc. Later, it adds global topic distribution words extended for only selected topics.

In contrast with probabilistic models (PLSA, LDA), Non-negative Matrix Factorization (NMF) model is addressed, which uses a geometrically linear algebra approach (SVD) for representation of words. Chen et al., [71] proposed one effective method, KGNF (Knowledge- guided Non-negative matrix factorization) which is designed with word-word pair wise semantic regularization via low-rank representations with the time-efficient algorithm. Shi et al., [72] developed Semantics-assisted Non-negative Matrix Factorization (SeaNMF) model to integrate semantic relations between word and context. Skip-gram model leverages to learn the context relations from the corpus. MacMillan and Wilson [76] proposed Topic Supervised Non- Negative Matrix Factorization (TS-NMF) for labeled corpus. It finds more meaningful contextual structure. Kandemir et al., [77] worked on by integrating LDA and sparse Gaussian processes. It classifies topic by joining latent cryptographic variables with each topic associated in the document.

3.3 Topic Model Approaches

Benchmark models like LDA, PLSA fails to fulfill some scenario and enervate to extract topics especially in short text data [78]. In this survey, we focused on different models that are built using a different representation form and statistical models for inferring the latent semantic topics [79].

Table 2. Distribution of short text topic modeling articles based on the strategy for the representation of words

Methods	Models	Reference
Windows-based strategy [49]	DMM(Dirichlet Multinomial mixture)	[80]
	BTM(Biterm Topic modeling)	[15],[35],
	PYNM(Pitman Yor Process Mixture)	[66]
Self-aggregation strategy [49]	PTM(Pseudo-Document-based Topic Model)	[37]
	SPTM(Sparsity Enhanced Topic Model)	[37]
	SATM(Self –Aggregate Topic Model)	[38]
	ETM(Embedding based topic model)	[22]
Word Embeddings [49]	GPU-DMM (Generalized Polya Urn -Dirichlet Multinomial Mixture)	[49]
	GPU-PDMM (Generalized Polya Urn -Poison based Dirichlet Multinomial Mixture)	[52]
	GLTM(Global and Local Word Embedding-Based Topic Model)	[45]

3.3.1 Window-based strategy

Short texts are creating more trends on internet especially in social media sites, news website and so on. Windows-based strategy scrutinizes on the co-occurrence of words in the data and cluster into its appropriated topics. Dirichlet Multinomial Mixture (DMM) model results better than LDA for short text data to analyze the topics. Nigam et al., [53] used DMM for categorizing the text documents based on the generative process estimated on the one-one correspondence of clusters [80]. Yin et al., [80] proposed a novel approach to infer the topics using Collapsed Gibbs Sampling for DMM (GSDMM)

which speed up the inference and achievement of more clustered topics. Yan et al., [81] proposed a new approach for the acquisition of knowledge for inferring and learning the topics. The key principle used for learning data is symmetric non-negative matrix factorization for every correlated word. Yan et al., [35] introduced Biterm Topic Model (BTM), which helps in grouping the correlated words. Cheng et al., [15] progressed Biterm Topic Model (BTM) for well organizing the topics and introduced two algorithms. Online Biterm Topic Model (OBTM) and Incremental Biterm Topic Model (IBTM). OBTM and IBTM are used to speed up the inference and updates the model parameters instantly as long as a new biterm is observed respectively. He et al., [82] introduced Alias Method and Metropolis-Hastings for BTM model to reduce the time and complexity. This process speeds up the inference when the topic grows more. Qiang et al., [66] built a new Pitman Yor Process Mixture Model (PYPM) for clustering the short text data. It uses the power law property to accurate the cluster and focuses on different parameters for grouping the probability of latent words.

3.3.2 Self-aggregation Strategy

To overcome sparse and noisy problem, researchers worked on aggregating the short text into long lengthy pseudo-documents in a different way like: users message aggregation [66], hashtag aggregation [83] under different context. Hong and Davison [23] introduced an aggregation method for twitter data. Quan et al., [38] proposed an automating Self-Aggregation based Topic Model (SATM) for short text. SATM deal with Sparsity by observing close similarity of original text data with hidden pseudo-document. It automatically aggregates topics during inferring the large short texts without using auxiliary information. SATM results in over fitting and expensive for computing the latent topic. To deal with this problem, Zuo et al., [37] introduced two models, Pseudo-document Topic Model [PTM] and Sparsity-enhanced Pseudo-document Topic Model (SPTM). PTM model could be beneficial for parameter estimation and enhance the efficiency for learning topics from data. Further it eliminates undesired correlations between pseudo-documents and latent topics by applying spike and slab priors. To provide an opportunity for the correlated words to cluster into a similar group, Qiang et al., [22] proposed a novel method of grouping the frequent co-occurrence words and aggregates to pseudo-document using Markov Random Field. Further uses semantic knowledge to extract topics

3.3.3 Word Embedding Strategy

Representation of words plays an important role in alleviating sparse and noise data in short text than long documents. Semantic information can be retrieved with less sparse using word embedding. Wikipedia and Word net are most commonly used embedding leverage to train the model [45], [52]. Using an auxiliary information Li et al., [49] developed a word embedding model to extract the thematic structure from the short texts. The model integrates GPU (Generalized Polya Urn) [30] and DMM (Dirichlet Multinomial Mixture). This model uses Global contextual information for retrieving the semantic relational words with frequently occurring words for vector representation. Li et al., [52] proposed a GPU-PDMM (Poison based Dirichlet Multinomial Mixture) Model to overcome the limitations of the DMM model. PDMM model helps to find more than one similar topic which are associated in short text using its semantic background knowledge. Concentrating on local and global variables, Liang et al., [45] introduced the GLTM (Global and Local Word Embedding-Based Topic Model). GLTM helps in mapping the semantically related word between local and global variables. This Model uses a continuous skip-gram with negative sampling framework [84]. It also makes uses of spike and slab priors' parameters.

4. DATASET FOR TOPIC MODELING

Table 9 presents a list of publically available dataset for topic modeling for short text as well as document data.

5. CHALLENGES

Short text emerges with more difficulties to understand the topics distributed in the corpus. Due to space limit of social media websites, text outcomes filled with more noisy and sparse words. Due to space limit of social media websites, text outcomes filled with more noisy and sparse words. Topic

Table 3. List of datasets used in topic modeling strategy

Data Type	Dataset	# No. of Topics	Related Works
Short text	Google News	-	[22],[66],[80]
	Snippets	8	[8],[29],[36],[44],[45], [49-52],[68],[74],[75]
	NIPS	-	[38]
	DBLP	6	[31],[37],[62]
	Yahoo! Answers	11	[38],[45],[77],[82]
	Online News	7	[26-29],[31],[32],[37], [47],[62],[68],[81]
	Baidu Q & A	35	[49],[52],[37],[81][35] [15],[39],[75]
	Satck Overflow Q & A	-	[29],[44],[47]
	Tweets {August to Oct 2008} {Microblog terc 2011 and 2012}	- 10	[15],[19],[22],[23],[27] ,[32],[35],[44-46],[51], [60][66],[67],[72],[73], [75],[80],[82],[83]
	Open Directory Project (ODP)	-	[37]
	Tweets related to Apple, Google, Microsoft, Twitter	-	[36],[70]
	Weibo Collections	-	[15],[35]
	Amazon Reviews (May 1994 - June 2014 contains product reviews)	- 7	[41] [45]
Documents	Wikipedia documents	12	[8]
	Finweek news articles	-	[17]

	RCV1	126	[56]
	20Newsgroups	20	[24],[36],[40],[43],[44],[48],[55],[57],[61]
	WebKB	7	[43],[44],[61]
	ohsumed	23	[24],[44]
	Reuters	65	[16],[24] [44],[51],[57]
	NIPS	50	[16],[18],[19],[57]
	ACM CIKM, SIGIR, KDD, and WWW (2001-2008) Documents	-	[60]

modeling solve the challenging issues like: sparse words, short form words, colloquial words, meaningless words etc.

- **Sparsity:** The words or terms in the corpus is not sufficient or filled with more short forms will fail to model the language accurately. Due to the limitation of words, users are frequently used to streamline the way to publish the short text. It will lead to the generation of highly sparse feature matrix, which are filled with more zeros.
- **Noise:** Non-relevant information, colloquial words, abbreviation, cyber languages are frequently implanted in short content. It will prompt result in noisy content.
- **Scalability:** For building up the inference algorithms which can scale to much massive stream data and non-trivial problem.

6. EVALUATION METRICS

Quantification of the degree of topic proportion learned by different topic models are examined by using various evolutions measures. These metrics illustrate to assesses the model's quality to find out the best model. Accuracy, Precision, Recall and f-measure are the evaluation measure used for classification of text data.

PMI (Point Mutual Information) Score, NMI (Normalized Mutual Information), NPMI (Normalized PMI), Purity, Perplexity, ARI (Adjusted Rand Index) are the coherence measures used in text data.

7. APPLICATIONS

Place Topic modeling is one hot trending method used in many text mining. It is used in various felids to retrieve information's in short or documents data, which is framed by mixture of words. Few major applications are listed below:

- **Industrial Applications**: Used in many search engines, online advertising system and to predict real-time topics of news, social media blogs or unseen corpus.
- **Bioinformatics:** To study patient related texts constructed from clinical, analyze cellular endpoints from microarray data, to study multidimensional genomic measurements and sequence classification.
- **Recommendation System:** In real time system like job recommendation, mapping right job for the candidate based on the content information, restaurant reviews based on users review interest, to study scientific records, history, comparative literature, political science law, cognitive science, sociology, media theory, linguistics, biology.
- **Financial Analysis:** Structure of the stock marketing exchange, utilizing stock value information to induce subjects over different traded on an open market organization.

8. CONCLUSION

Topic modeling has gained a trend in extracting the hidden themes in the large unseen data from the last decade. This paper looks over into the detail description and evolution of topic modeling. Further, focused on understanding the importance of topic modeling in the digital era. We also concentrated on the methodologies used for extraction of topic from documents and short text data. The different strategies used for analysis of latent structure from the data and its representation form in short texts. One of the main contributions of this paper is dataset used so far for topic modeling. Furthermore, we listed various measures used for topic coherence and classification of topic modeling. Finally, we mentioned some major applications and challenges of topic models. The work presented in this paper can be significant source for the researchers in the topic modeling domain. From the overview of the state-of-the-art, topic modeling over short texts is an increasingly challenging compared with normal texts.

9. REFERENCES

[1] Ghanshyambhai, C.U., and Shah, A., 2018. Optimizing topic coherence in the Gujarati text topic modeling: a relevant words-based approach. Ph.D. thesis.

[2] Blei, D.M., 2012. Probabilistic topic models. Communications of the ACM, 55(4), pp.77-84.

[3] Das, R., Zaheer, M. and Dyer, C., 2015. Gaussian lda for topic models with word embeddings. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 1, pp. 795-804.

[4] Revanasiddappa, M.B., Harish, B.S. and Kumar, S.A., 2018. Meta-cognitive Neural Network based Sequential Learning Framework for Text Categorization. *Procedia computer science*, *132*, pp.1503-1511.

[5] Revanasiddappa, M. B., and Harish, B. S. 2019. A Novel Text Representation Model to Categorize Text Documents using Convolution Neural Network. International Journal of Intelligent Systems and Applications, 5, 36-45.

[6] Gupta,V. and Lehal, G.S., 2009. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), pp.60-76.

[7] Gangemi, A., Presutti, V. and Recupero, D.R., 2014. Frame-based detection of opinion holders and topics: a model and a tool. IEEE Computational Intelligence Magazine, 9(1), pp.20-30.

[8] Phan, X.H., Nguyen, C.T., Le, D.T., Nguyen, L.M., Horiguchi, S. and Ha, Q.T., 2011. A hidden topic-based framework toward building applications with short web documents. IEEE Transactions on Knowledge and Data Engineering, 23(7), pp.961-976.

[9] Papadimitriou, C.H., Raghavan, P., Tamaki, H. and

Vempala, S., 2000. Latent semantic indexing: A probabilistic analysis. Journal of Computer and System Sciences, 61(2), pp.217-235.

[10] Landauer, T.K., Foltz, P.W. and Laham, D., 1998. An introduction to latent semantic analysis. Discourse processes, 25(2-3), pp.259-284.

[11] Hofmann, T., 1999, July. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc.

[12] Bassiou, N.K. and Kotropoulos, C.L., 2014. Online PLSA: Batch updating techniques including out-of-vocabulary words. IEEE transactions on neural networks and learning systems, 25(11), pp.1953-1966.

[13] Liu, S., Xia, C. and Jiang, X., 2010, December. Efficient probabilistic latent semantic analysis with sparsity control. In 2010 IEEE International Conference on Data Mining (pp. 905-910). IEEE.

[14] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

[15] Cheng, X., Yan, X., Lan, Y. and Guo, J., 2014. Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12), pp.2928-2941.

[16] AlSumait, L., Barbará, D. and Domeniconi, C., 2008, December. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 3-12). IEEE.

[17] Mazarura, J. and de Waal, A., 2016. A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. In Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2016 (pp. 1-6). IEEE.

[18] Yi, X. and Allan, J., 2008. Evaluating topic models for information retrieval. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 1431-1432). ACM.

[19] Yao, L., Mimno, D. and McCallum, A., 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 (pp. 937-946). ACM.

[20] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L., 2017. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Multimedia Tools and Applications, pp.1-43.

[21] Alghamdi, R. and Alfalqi, K., 2015. A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 6(1).

[22] Qiang, J., Chen, P., Wang, T. and Wu, X., 2017. Topic modeling over short texts by incorporating word embeddings. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 363-374). Springer.

[23] Hong, L. and Davison, B.D., 2010. Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80-88). ACM.

[24] Schneider, J. and Vlachos, M., 2018. Topic Modeling based on Keywords and Context. In Proceedings of the 2018 SIAM International Conference on Data Mining (pp. 369-377).

[25] Revanasiddappa, M. B., & Harish, B. S. (2018). A New Feature Selection Method based on Intuitionistic Fuzzy Entropy to Categorize Text Documents. International Journal of Interactive Multimedia & Artificial Intelligence, 5(3).

[26] Li, L., Sun, Y., Han, X. and Wang, C., 2018, June. Research on Improve Topic Representation over Short Text. In 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC) (pp. 848-853). IEEE.

[27] Chen, G.B. and Kao, H.Y., 2017. Word co-occurrence augmented topic model in short text. Intelligent Data Analysis, 21(S1), pp.S55-S70.

[28] Xun, G., Gopalakrishnan, V., Ma, F., Li, Y., Gao, J. and Zhang, A., 2016. Topic discovery for short texts using word embeddings. In Data Mining (ICDM), 2016 IEEE 16th International Conference on (pp. 1299-1304). IEEE.

[29] Chen, Y., Zhang, H., Liu, R., Ye, Z. and Lin, J., 2019. Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems, 163, pp.1-13.

[30] Harish, B.S. and Revanasiddappa, M.B., 2017. A comprehensive survey on various feature selection methods to categorize text documents. International Journal of Computer Applications, 164(8), pp.1-7.

[31] Lu, H.Y., Ge, G.J., Li, Y., Wang, C.J. and Xie, J.Y., 2018, November. Exploiting Global Semantic Similarity Biterms for Short-Text Topic Discovery. In 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 975-982). IEEE.

[32] Chen, G.B. and Kao, H.Y., 2015. Word co-occurrence augmented topic model in short text. International Journal of Computational Linguistics & Chinese Language Processing, 20(2).

[33] Salerno, M.D., Tataru, C.A. and Mallory, M.R., 2015. Word Community Allocation: Discovering Latent Topics via Word Co-Occurrence Network Structure.

[34] Chen, B., 2009. Latent topic modelling of word co-occurence information for spoken document retrieval.

[35] Yan, X., Guo, J., Lan, Y. and Cheng, X., 2013. A biterm topic model for short texts. In Proceedings of the 22nd international conference on World Wide Web (pp. 1445-1456). ACM.

[36] Pedrosa, G., Pita, M., Bicalho, P., Lacerda, A. and Pappa, G.L., 2016. Topic modeling for short texts with co-occurrence frequency-based expansion. In 2016 5th Brazilian Conference on Intelligent Systems (BRACIS) (pp. 277-282).

[37] Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K. and Xiong, H., 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 (pp. 2105-2114). ACM.

[38] Quan, X., Kit, C., Ge, Y. and Pan, S.J., 2015. Short and sparse text topic modeling via self-aggregation. In 24th International Joint Conference on Artificial Intelligence.

[39] Jiang, L., Lu, H., Xu, M. and Wang, C., 2016. Biterm Pseudo Document Topic Model for Short Text. In Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on (pp. 865-872). IEEE.

[40] Lee, S., Kim, J. and Myaeng, S.H., 2015. An extension of topic models for text classification: A term weighting approach. In Big Data and Smart Computing (BigComp), 2015 International Conference on (pp. 217-224). IEEE.

[41] Kai, Y., Yi, C., Zhenhong, C., Ho-fung, L. and Raymond, L.A.U., 2016. Exploring topic discriminating power of words in latent dirichlet allocation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2238-2247).

[42] Wilson, A.T. and Chew, P.A., 2010. Term weighting schemes for latent dirichlet allocation. In human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics (pp. 465-473).

[43] Wang, Q., Zhang, D. and Si, L., 2013. Semantic hashing using tags and topic modeling. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 213-222).

[44] Li, X., Zhang, A., Li, C., Ouyang, J. and Cai, Y., 2018. Exploring coherent topics by topic modeling with term weighting. Information Processing & Management.

[45] Liang, W., Feng, R., Liu, X., Li, Y. and Zhang, X., 2018. GLTM: A Global and Local Word Embedding-Based Topic Model for Short Texts. IEEE, 6, pp.43612-43621.

[46] Rajani, N.F.N., McArdle, K. and Baldridge, J., 2014. Extracting topics based on authors, recipients and content in microblogs. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 1171-1174). ACM.

[47] Lu, H.Y., Ge, G.J., Li, Y., Wang, C.J. and Xie, J.Y., 2018. Exploiting Global Semantic Similarity Biterms for Short-Text Topic Discovery. 30th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 975-982).

[48] Reisinger, J., Waters, A., Silverthorn, B. and Mooney, R.J., 2010. Spherical topic models. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 903-910).

[49] Li, C., Wang, H., Zhang, Z., Sun, A. and Ma, Z., 2016, July. 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 (pp. 165-174). ACM.

[50] Mimno, D., Wallach, H.M., Talley, E., Leenders, M. and McCallum, A., 2011. Optimizing semantic coherence in topic models. In Proceedings of the conference on empirical methods in natural language processing (pp. 262-272). Association for Computational Linguistics.

[51] Zhao, H., Du, L. and Buntine, W., 2017, November. A word embeddings informed focused topic model. In Asian Conference on Machine Learning (pp. 423-438).

[52] Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A. and Ma, Z., 2017. Enhancing Topic Modeling for Short Texts with Auxiliary Word Embeddings. ACM Transactions on Information Systems (TOIS), 36(2), p.11.

[53] Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T., 2000. Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2-3), pp.103-134.

[54] Liu, Y., Liu, Z., Chua, T.S. and Sun, M., 2015, January. Topical Word Embeddings. In AAAI (pp. 2418-2424).

[55] Nguyen, D.Q., Billingsley, R., Du, L. and Johnson, M., 2015. Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 3, pp.299-313.

[56] Geeganage, K. and Tharanga, D., 2018. Concept Embedded Topic Modeling Technique. International World Wide Web Conferences Steering Committee, (pp. 831-835).

[57] Jiang, H., Zhou, R., Zhang, L., Wang, H. and Zhang, Y., 2018. Sentence level topic models for associated topics extraction. World Wide Web, pp.1-16.

[58] Tsai, F.S., 2011. A tag-topic model for blog mining. Expert Systems with Applications, 38(5), pp.5330-5335.

[59] Wang, Y., Liu, J., Huang, Y. and Feng, X., 2016. Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs. IEEE Transactions on Knowledge and Data Engineering, 28(7), pp.1919-1933.

[60] Kawamae, N., 2010, July. Author interest topic model. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 887-888). ACM.

[61] Clinchant, S. and Perronnin, F., 2013. Aggregating continuous word embeddings for information retrieval. In Proceedings of the workshop on continuous vector space models and their compositionality (pp. 100-109).

[62] Kim, H., Sun, Y., Hockenmaier, J. and Han, J., 2012. Etm: Entity topic models for mining documents associated with entities. In Data Mining (ICDM), 2012 IEEE 12th International Conference on (pp. 349-358).

[63] Jadhav, B.S., Bhosale, D.S. and Jadhav, D.S., 2016, August. Pattern based topic model for data mining. In Inventive Computation Technologies (ICICT), International Conference on (Vol. 2, pp. 1-6). IEEE.

[64] Li, W. and McCallum, A., 2006, June. Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577-584). ACM.

[65] V. Lavrenko and W. B. Croft. Relevance-based language models. In Proc. of ACM SIGIR, pp. 120-127, 2001.

[66] Qiang, J., Li, Y., Yuan, Y. and Wu, X., 2018. Short text clustering based on Pitman-Yor process mixture model. *Applied Intelligence*, *48*(7), pp.1802-1812.

[67] Qiu, Z. and Shen, H., 2017. User clustering in a dynamic social network topic model for short text streams. Information Sciences, 414, pp.102-116.

[68] Hu, X., Wang, H. and Li, P., 2018. Online Biterm Topic Model based short text stream classification using short text expansion and concept drifting detection. Pattern

Recognition Letters, 116, pp.187-194.

[69] Demšar, J. and Bosnić, Z., 2018. Detecting concept drift in data streams using model explanation. Expert Systems with Applications, 92, pp.546-559.

[70] Wandabwa, H., Naeem, M.A., Pears, R. and Mirza, F., 2018. A Metamodel Enabled Approach for Discovery of Coherent Topics in Short Text Microblogs. IEEE Access, 6, pp.65582-65593.

[71] Wang, T., Cai, Y., Leung, H.F., Cai, Z. and Min, H., 2015. Entropy-based term weighting schemes for text categorization in VSM. In Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on (pp. 325-332). IEEE.

[72] Shi, L.L., Liu, L., Wu, Y., Jiang, L. and Hardy, J., 2017. Event detection and user interest discovering in social media data streams. IEEE Access, 5, pp.20953-20964.

[73] Sapul, M.S.C., Aung, T.H. and Jiamthapthaksin, R., 2017. Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms. In 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-6). IEEE.

[74] Zheng, C.T., Liu, C. and San Wong, H., 2018. Corpus-based topic diffusion for short text clustering. Neurocomputing, 275, pp.2444-2458.

[75] Li, X., Wang, Y., Zhang, A., Li, C., Chi, J. and Ouyang, J., 2018. Filtering out the noise in short text topic modeling. Information Sciences, 456, pp.83-96.

[76] MacMillan, K. and Wilson, J.D., 2017. Topic supervised non-negative matrix factorization. arXiv:1706.05084

[77] Kandemir, M., Kekeç, T. and Yeniterzi, R., 2018. Supervising topic models with Gaussian processes. Pattern Recognition, 77, pp.226-236.

[78] Phan, X.H., Nguyen, L.M. and Horiguchi, S., 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. 17th international conference on WWW (pp. 91-100).

[79] Qiang, J., Li, Y., Yuan, Y., Liu, W. and Wu, X., 2018. STTM: A Tool for Short Text Topic Modeling. arXiv preprint arXiv:1808.02215.

[80] Yin, J. and Wang, J., 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 233-242). ACM.

[81] Yan, X., Guo, J., Liu, S., Cheng, X. and Wang, Y., 2013. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In Proceedings of the 2013 SIAM International Conference on Data Mining (pp. 749-757).

[82] He, X., Xu, H., Li, J., He, L. and Yu, L., 2017. FastBTM: Reducing the sampling time for biterm topic model. Knowledge-Based Systems, 132, pp.11-20.

[83] Mehrotra, R., Sanner, S., Buntine, W. and Xie, L., 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 889-892).

[84] Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.