Multimedia Tools and Applications
https://doi.org/10.1007/s11042-018-6894-4

Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey

Hamed Jelodar[1] · **Yongli Wang**[1,2] · **Chi Yuan**[1] · **Xia Feng**[1] · **Xiahui Jiang**[1] · **Yanchao Li**[1] · **Liang Zhao**[1]

Received: 5 June 2018 / Revised: 28 October 2018 / Accepted: 13 November 2018 /
Published online: 28 November 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Topic modeling is one of the most powerful techniques in text mining for data mining, latent data discovery, and finding relationships among data and text documents. Researchers have published many articles in the field of topic modeling and applied in various fields such as software engineering, political science, medical and linguistic science, etc. There are various methods for topic modelling; Latent Dirichlet Allocation (LDA) is one of the most popular in this field. Researchers have proposed various models based on the LDA in topic modeling. According to previous work, this paper will be very useful and valuable for introducing LDA approaches in topic modeling. In this paper, we investigated highly scholarly articles (between 2003 to 2016) related to topic modeling based on LDA to discover the research development, current trends and intellectual structure of topic modeling. In addition, we summarize challenges and introduce famous tools and datasets in topic modeling based on LDA.

Keywords Topic modeling · Latent Dirichlet allocation · Tag recommendation · Semantic web · Gibbs sampling

✉ Yongli Wang
YongliWang@njust.edu.cn

Hamed Jelodar
Jelodar@njust.edu.cn

Chi Yuan
yuanchi@njust.edu.cn

Xia Feng
779477284@qq.com

Xiahui Jiang
jxhchina@gmail.com

[1] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China

[2] China Electronics Technology Cyber Security Co., Ltd, Chengdu, China

1 Introduction

Natural language processing (NLP) is a challenging research in computer science to information management, semantic mining, and enabling computers to obtain meaning from human language processing in text-documents. Topic modeling methods are powerful smart techniques that widely applied in natural language processing to topic discovery and semantic mining from unordered documents [12]. In a wide perspective, Topic modeling methods based on LDA have been applied to natural language processing, text mining, and social media analysis, information retrieval. For example, topic modeling based on social media analytics facilitates understanding the reactions and conversations between people in online communities, As well as extracting useful patterns and understandable from their interactions in addition to what they share on social media websites such as twiiter facebook [145, 176]. Topic models are prominent for demonstrating discrete data; also, give a productive approach to find hidden structures(semantics) in gigantic information. There are many papers for in this field and definitely cannot mention to all of them, so we selected more signification papers. Topic models are applied in various fields including medical sciences [62, 116, 170, 197] , software engineering [6, 46, 89, 148, 149], geography [34, 39, 139, 147, 184], political science [18, 30, 49], etc.

For example in political science, In [49] proposed a new two-layer matrix factorization methodology for identifying topics in large political speech corpora over time and identify both niche topics related to events at a particular point in time and broad, long-running topics. This paper has focused on European Parliament speeches, the proposed topic modeling method has a number of potential applications in the study of politics, including the analysis of speeches in other parliaments, political manifestos, and other more traditional forms of political texts. In [42] suggested a new unsupervised topic model based on LDA for contrastive opinion modeling which purpose to find the opinions from multiple views, according to a given topic and their difference on the topic with qualifying criteria, the model called Cross-Perspective Topic (CPT) model. They performed experiments with both qualitative and quantitative measures on two datasets in the political area that include: first dataset is statement records of U.S. senators that show political stances of senators. Also for the second dataset, extracted of world News Medias from three representative media in U.S (New York Times), China (Xinhua News) and India (Hindu). To evaluate their approach with other models, used corrIDA and LDA as two baselines.

Another group of researchers focused on topic modeling in software engineering, in [89] for the first time, they used LDA, to extract topics in source code and perform visualization of software similarity, In other words, LDA is used as an intuitive approach for calculation of similarity between source files and obtain their respective distributions of each document over topics. They utilized their method on 1,555 software projects from Apache and SourceForge that includes 19 million source lines of code (SLOC). The authors demonstrated this approach, can be effective for project organization, software refactoring. In [150] introduced a method based on LDA for automatically categorizing software systems, called LACT. For evaluation of LACT, used 43 open-source software systems in different programming languages and showed LACT can categorize software systems based on type of programming language. In [98, 99] proposed an approach topic modeling based on LDA model for the purpose of bug localization. Their idea, applied to analysis of same bugs in Mozilla and Eclipse and result showed that their LDA-based approach is better than LSI for evaluate and analyze bugs in these source codes.

An analysis of geographic information is another issue that can be referred to [139]. They introduced a novel method based on multi-modal Bayesian models to describe social media by merging text features and spatial knowledge that called GeoFolk. As a general outlook, this method can be considered as an extension of Latent Dirichlet Allocation (LDA). They used the available standard CoPhIR dataset that contains an abundance of over 54 million Flickr. The GeoFolk model has the ability to be used in quality-oriented applications and can be merged with some models from Web 2.0 social. In [184], this article examines the issue of topic modeling to extract the topics from geographic information and GPS-related documents. They suggested a new location text method that is a combination of topic modeling and geographical clustering called LGTA (Latent Geographical Topic Analysis). To test their approaches, they collected a set of data from the website Flickr, according to various topics.

In other view, Most of the papers that were studied, had goals for this topic modeling, such as: Source code analysis [19, 46, 89, 90, 99, 135, 150] , Opinion and aspect Mining [7, 18, 23, 64, 115, 151, 163, 172, 192, 201], Event detection [57, 87, 122, 167], Image classification [34, 157], recommendation system [25, 65, 96, 162, 178, 199, 205] and emotion classification [127, 128, 132], etc. For example in recommendation system, in [199] proposed a personalized hashtag recommendation approach based LDA model that can discover latent topics in microblogs, called Hashtag-LDA and applied experiments on "UDI-TwitterCrawl-Aug2012-Tweets" as a real-world Twitter dataset.

1.1 Literature review and related works

Topic models have many applications in natural processing languages. Many articles have been published based on topic modeling approaches in various subject such as Social Network, software engineering, Linguistic science and etc. There are some works that have focused on survey in Topic modeling.

In [22], the authors presented a survey on topic modeling in software engineering field to specify how topic models have thus far been applied to one or more software repositories. They focused on articles written between Dec 1999 to Dec 2014 and surveyed 167 article that using topic modeling in software engineering area. They identified and demonstrate the research trends in mining unstructured repositories by topic models. They found that most of studies focused on only a limited number of software engineering task and also most studies use only basic topic models. In [35], the authors focused on survey in Topic Models with soft clustering abilities in text corpora and investigated basic concepts and existing models classification in various categories with parameter estimation (such as Gibbs Sampling) and performance evaluation measures. In addition, the authors presented some applications of topic models for modeling text corpora and discussed several open issues and future directions.

In [93], introduced and surveyed the field of opinion mining and sentiment analysis, which helps us to observe a elements from the intimidating unstructured text. The authors discussed the most extensively studied subject of subjectivity classification and sentiment which specifies whether a document is opinionated. Also, they described aspect-based sentiment analysis which exploits the full power of the abstract model and discussed about aspect extraction based on topic modeling approaches. In [36], they discussed challenges of text mining techniques in information systems research and indigested the practical application of topic modeling in combination with explanatory regression analysis, using online customer reviews as an exemplary data source.

In [144], the authors presented a survey of how topic models have thus far been applied in Software Engineering tasks from four SE journals and eleven conference proceedings. They considered 38 selected publications from 2003 to 2015 and found that topic models are widely used in various SE tasks in an increasing tendency, such as social software engineering, developer recommendation and etc. our research difference with other works is that, we had a deep study on topic modeling approaches based on LDA with the coverage of various aspects such as applications, tools , dataset and models.

1.2 Motivations and contributions

Since, topic modeling techniques can be very useful and effective in natural language processing to semantic mining and latent discovery in documents and datasets. Hence, our motivation is to investigate the topic modeling approaches in different subjects with the coverage of various aspects such as models, tools, dataset and applications. The main goal of this work is to provide an overview of the methods of topic modeling based on LDA. In summary, this paper makes four main contributions:

- We investigate scholarly articles (from 2003 to 2016) which are related to Topic Modeling based on LDA to discover the research development, current trends and intellectual structure of topic modeling based on LDA.
- We investigate topic modeling applications in various sciences.
- We summarize challenges in topic modeling, such as image processing, Visualizing topic models, Group discovery, User Behavior Modeling, and etc.
- We introduce some of the most famous data and tools in topic modeling.

2 Computer science and topic modeling

Topic models have an important role in computer science for text mining and natural language processing. In Topic modeling, a topic is a list of words that occur in statistically significant methods. A text can be an email, a book chapter, a blog posts, a journal article and any kind of unstructured text. Topic models cannot understand the means and concepts of words in text documents for topic modeling. Instead, they suppose that any part of the text is combined by selecting words from probable baskets of words where each basket corresponds to a topic. The tool goes via this process over and over again until it stays on the most probable distribution of words into baskets which call topics. Topic modeling can provide a useful view of a large collection in terms of the collection as a whole, the individual documents, and relationships between documents. In Fig. 1, we provided a taxonomy of topic modeling methods based on LDA, from some of the impressive works.

2.1 Latent Dirichlet allocation

LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words. Latent Dirichlet allocation (LDA), first introduced by Blei, Ng and Jordan in 2003 [12], is one of the most popular methods in topic modeling. LDA represents topics by word probabilities. The words with highest probabilities in each topic usually give a good idea of what the topic is can word probabilities from LDA.

LDA, an unsupervised generative probabilistic method for modeling a corpus, is the most commonly used topic modeling method. LDA assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is also represented as a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well. Given a corpus D consisting of M documents, with

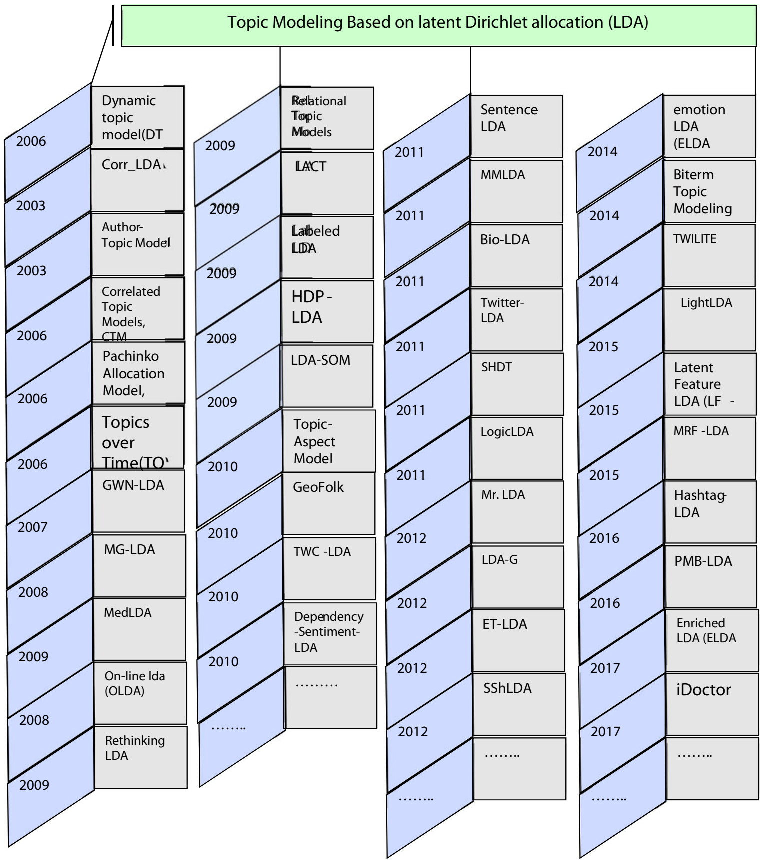


Fig. 1 A taxonomy of methods based on extension LDA, considered some of the impressive works

document d having N_d words ($d \in 1, \ldots, M$), LDA models D according to the following generative process [12]:

(a) Choose a multinomial distribution $\varphi_t$for topic $t(t \in \{1, \ldots, T\})$ from a Dirichlet distribution with parameter β

(b) Choose a multinomial distribution θ_d for document $d(d \in \{1, \ldots, M\})$ from a Dirichlet distribution with parameter α.

(c) For a word w_n $(n \in \{1, \ldots, N_d\})$ in document d,

1. (a) i Select a topic z_n from θ_d.
   ii Select a word w_n from φ_{zn}.

In above generative process, words in documents are only observed variables while others are latent variables (φ and θ) and hyper parameters (α and β). The probability of observed data D is computed and obtained of a corpus as follows:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \tag{1}$$

Defined α parameters of topic Dirichlet prior and the distribution of words over topics, which, drawn from the Dirichlet distribution, given β. Defined, T is the number of topics, M as number of documents; N is the size of the vocabulary. The Dirichlet-multinomial pair for the corpus-level topic distributions, considered as (α, θ). The Dirichlet-multinomial pair for topic-word distributions, given (β, φ). The variables θ_d are document-level variables, sampled when per document. z_{dn} , w_{d_n} variables are word-level variables and are sampled when for each word in each text-document.

LDA is a distinguished tool for latent topic distribution for a large corpus. Therefore, it has the ability to identify sub-topics for a technology area composed of many patents, and represent each of the patents in an array of topic distributions. With LDA, the terms in the set of documents, generate a vocabulary that is then applied to discover hidden topics. Documents are treated as a mixture of topics, where a topic is a probability distribution over this set of terms. Each document is then seen as a probability distribution over set of topics. We can think of the data as coming from a generative process that is defined by the joint probability distribution over what is observed and what is hidden.

2.2 Parameter estimation, inference, training for LDA

Various methods have been proposed to estimate LDA parameters, such as variational method [12], expectation propagation [110] and Gibbs sampling [51].

Gibbs sampling, is a Monte Carlo Markov-chain algorithm, powerful technique in statistical inference, and a method of generating a sample from a joint distribution when only conditional distributions of each variable can be efficiently computed. According to our knowledge, researchers have widely used this method for the LDA. Some of works related based on LDA and Gibbs, such as [61, 97, 108, 114, 124, 127, 150, 154, 175, 183, 198].

Expectation-maximization (EM), is a powerful method to obtain parameter estimation of graphical models and can use for unsupervised learning. In fact, the algorithm relies on discovering the maximum likelihood estimates of parameters when the data model depends on certain latent variables EM algorithm contains two steps, the E-step (expectation) and the M-step (maximization). Some researchers have applied this model to LDA training, such as [13, 16, 52, 79, 203].

Variational Bayes inference (VB), VB can be considered as a type of EM extension that uses a parametric approximation to the posterior distribution of both parameters and other latent variables and attempts to optimize the fit (e.g. using KL-divergence) to the observed data. Some researchers have applied this model to LDA training, such as [26, 200].

2.3 A brief look at past work: research between 2003 to 2009

The LDA was first presented in 2003, and researchers have been tried to provide extended approaches based on LDA, as shown in Table 1. Undeniably, this period (2003 to 2009) is very important because key and baseline approaches were introduced, such as: Corr_LDA, Author-Topic Model, DTM, Rethinking LDA and RTM etc.

Author-Topic model [133], is a popular and simple probabilistic model in topic modeling for finding relationships among authors, topics, words and documents. This model provides a distribution of different topics for each author and also a distribution of words for each topic. For evaluation, the authors used 1700 papers of NIPS conference and also 160,000 CiteSeer abstracts of CiteSeer dataset. To estimate the topic and author distributions applied Gibbs sampling. According to their result, showed this approach can provide a significantly predictive for interests of authors in perplexity measure. This model has attracted much attention from researchers and many approaches proposed based on ATM, such as [105, 149].

DTM, Dynamic Topic Model (DTM) is introduced by Blei and Lafferty as an extension of LDA that this model can obtain evolution of topics over time in a sequentially arranged corpus of documents and exhibits evolution of word-topic distribution which makes it easy to vision the topic trend [14]. As an advantage, DTM is very impressible for extracting topics from collections that change slowly over a period of time.

Labeled LDA (LLDA) is another LDA extension which suppose that each document has a set of known labels [124]. This model can be trained with labeled documents and even supports documents with more than one label. Topics are learned from the co-occurring terms in places from the same category, with topics approximately capturing different place categories. A separate L-LDA model is trained for each place category, and can be used to infer the category of new, previously unseen places. LLDA is a supervised algorithm that makes topics applying the Labels assigned manually. Therefore, LLDA can obtain meaningful topics, with words that map well to the labels applied. As a disadvantage, Labeled LDA has limitation to support latent subtopics within a determined label or any global topics. For overcome this problem, proposed partially labeled LDA (PLDA) [125].

MedLDA, proposed the maximum entropy discrimination latent Dirichlet allocation (MedLDA) model, which incorporates the mechanism behind the hierarchical Bayesian models (such as, LDA) with the max margin learning (such as SVMs) according to a unified restricted optimization framework. In fact each data sample is considered to a point in a finite dimensional latent space, of which each feature corresponds to a topic, i.e., unigram distribution over terms in a vocabulary [203]. MEDLDA model can be applied in both classification and regression. However, the quality of the topical space learned using MedLDA is heavily affected by the quality of the classifications which are learned at per iteration of MedLDA, as a disadvantage.

Relational Topic Models (RTM), is another extension, RTM is a hierarchical model of networks and per-node attribute data. First, each document was created from topics in LDA. Then, modelling the connections between documents and considered as binary variables, one for each pair from documents. These are distributed based on a distribution that depends on topics used to generate each of the constituent documents. So in this way, the content of the documents are statistically linked to the link structure between them and we can say that this model can be used to summarize a network of documents [16]. In fact, the advantage of RTM is that it considers both links and document context between documents. The disadvantage of RTM is its scalability. Since RTM can only make predictions for a document couple, this means that the action of recommending related questions for a recently created one would bring in computing couple RTM responses among every available question in the query and the collection.

2.4 A brief look at past work: research between 2010 to 2011

Twenty seven approaches are introduced in Table 1, where eight were published in 2010 and six in 2011. According to the Table 1, used LDA model for variety subjects, such as: Scientific topic discovery [115], Source code analysis [135], Opinion Mining [192], Event detection [87], Image Classification [157].

Sizov et al. in [139] introduced a novel method based on multi-modal Bayesian models to describe social media by merging text features and spatial knowledge that called GeoFolk. As a general outlook, this method can be considered as an extension of Latent

Table 1 Some impressive articles based on LDA: between 2003 - 2011

Author-study	Model	Years	Parameter estimation / inference	Methods	Problem domain
[13]	Corr_LDA	2003	Variational EM	LDA	Image annotation and retrieval
[133]	Author-topic model	2004	Gibbs sampling	LDA	Find the relationships between authors, documents, words, and topics
[105]	AuthorRecipient-topic (ART)	2005	Gibbs sampling	LDA	Social network analysis and role discovery
				Author-topic (AT)	
[14]	Dynamic topic model (DTM)	2006	Kalman variational algorithm	LDA	Provide a dynamic model for evolution of topics
				Galton Watson process	
[155]	Topics over time (TOT)	2006	Gibbs sampling	LDA	Capture word co-occurrences and localization in continuous time
[74]	Pachinko allocation model, PAM	2006	Gibbs sampling	LDA	Capture arbitrary topic correlations
				A directed acyclic graph method	
[193]	GWN-LDA	2007	Gibbs sampling	LDA	Probabilistic community profile Discovery in social network
				Hierarchical Bayesian algorithm	
[4]	On-line lda (OLDA)	2008	Gibbs sampling	LDA	Tracking and topic detection
				Empirical Bayes method	
[151]	MG-LDA	2008	Gibbs sampling	LDA	Sentiment analysis in multi-aspect
[124]	Labeled LDA	2009	Gibbs sampling	LDA	Producing a labeled document collection
[16]	Relational topic models	2009	Expectation-maximization (EM)	LDA	Make predictions between nodes, attributes, links structure
[156]	HDP-LDA	2009	Gibbs sampling	LDA	
[67]	DiscLDA	2009	Gibbs sampling	LDA	Classification and dimensionality reduction in documents

Table 1 (continued)

Author-study	Model	Years	Parameter estimation / inference	Methods	Problem domain
[139]	GeoFolk	2010	Gibbs sampling	LDA	Content management and retrieval of spatial information
[61]	JointLDA	2010	Gibbs sampling	LDA	Mining multilingual topics
				Bag-of-word model	
[115]	Topic-aspect model (TAM)	2010	Gibbs sampling	LDA SVM	Scientific topic discovery
[75]	Dependency-sentiment-LDA	2010	Gibbs sampling	LDA	Sentiment classification
[135]	TopicXP	2010		LDA	Source code analysis
[192]	Constrained-LDA	2010	Gibbs sampling	LDA	Opinion mining and grouping product features
[87]	PET - popular events tracking	2010	Gibbs sampling	LDA	Event analysis in social network
[87, 167]	EDCoW	2010		LDA	Event analysis in twitter
				Wavelet transformation	
[158]	Bio-LDA.	2011	Gibbs sampling	LDA	Extract biological terminology
[198]	Twitter-LDA	2011	Gibbs sampling	LDA	Extracting topical keyphrases and analyzing twitter content
				Author-topic model PageRank	
[157]	Max-margin latent Dirichlet allocation (MMLDA)	2011	Variational inference	LDA SVM	Image classification and annotation
[64]	Sentence-LDA	2011	Gibbs sampling	LDA	Aspects and sentiment discovery for web review
[92]	PLDA+	2011	Gibbs sampling	LDA	Reduce inter-computer communication time
				Weighted round-robin	
[26]	Dirichlet class language model, (DCLM)	2011	Variational Bayesian EM (VB-EM) algorithm	Speech recognition and exploitation of language models	Dirichlet class language model, (DCLM)

Dirichlet Allocation (LDA). They used the available standard CoPhIR dataset that it contains an abundance of over 54 million Flickr. The GeoFolk model has ability to be used in quality-oriented applications and can be merged with some models from Web 2.0 social.

Zhai et al. in [192] used prior knowledge as a constraint in the LDA models to improve grouping of features by LDA. They extract must link and cannot-link constraint from the corpus. Must link indicates that two features must be in the same group while cannot-link restricts that two features cannot be in the same group. These constraints are extracted automatically. If at least one of the terms of two product features are same, they are considered to be in the same group as must link. On the other hand, if two features are expressed in the same sentence without conjunction “and”, they are considered as a different feature and should be in different groups as cannot-link.

Wang et al. in [158] suggested an approach based on LDA that called Bio-LDA that can identify biological terminology to obtain latent topics. The authors have shown that this approach can be applied in different studies such as association search, association predication, and connectivity map generation. And they showed that Bio-LDA can be applied to increase the application of molecular bonding techniques as heat maps.

2.5 A brief look at past work: research between 2012 to 2013

According to Table 2, some of the popular works published between 2012 and 2013 focused on a variety of topics, such as music retrieve [186], opinion and aspect mining [77], Event analysis [57].

ET-LDA, in this work, the authors developed a joint Bayesian model that performs event segmentation and topic modeling in one unified framework. In fact, they proposed an LDA model to obtain event’s topics and analysis tweeting behaviors on Twitter that called Event and Tweets LDA (ET-LDA). They employed Gibbs Sampling method to estimate the topic distribution [57]. The advantage of ET-LDA is that can extract top general topic for entire tweet collection, which means that is very relevant and influenced with the event.

Mr. LDA, the authors introduced a novel model and parallelized LDA algorithm in the MapReduce framework that called Mr. LDA. In contrast other approaches which use Gibbs sampling for LDA, this model uses variational inference [200]. LDA-GA, The authors focused on the issue of Textual Analysis in Software Engineering. They proposed an LDA model based on Genetic Algorithm to determine a near-optimal configuration for LDA (LDA-GA), This approach is considered by three scenarios that include: (a) traceability link recovery, (b) feature location, and (c) labeling. They applied the Euclidean distance for measuring distance between documents and used Fast collapsed Gibbs sampling to approximate the posterior distributions of parameters [114].

TopicSpam, they proposed a generative LDA based topic modeling methods for detecting fake review their model can obtain differences among truthful and deceptive reviews. Also, their approach can provide a clear probabilistic prediction about how probable a review is truthful or deceptive. For evaluation, the authors used 800 reviews of 20 Chicago hotels and showed that TopicSpam slightly outperforms TopicTDB in accuracy.

SSHLDA, the authors proposed a semi supervised hierarchical approach based on topic model which goals for exploring new topics in the data space whenever incorporating the information from viewed labels of hierarchical into the modeling process, called SemiSupervised Hierarchical LDA (SSHLDA). They applied labels hierarchy in as basic hierarchy, called Base Tree (BT); then they used hLDA to make topic hierarchy automatically for each

Table 2 Some impressive articles based on LDA: between 2012-2016

Author-study	Model	Years	Parameter estimation / inference	Methods	Problem domain
[171]	Locally discriminative topic model (LDTM)	2012	Expectation-maximization (EM)	LDA	Document semantic analysis
[57]	ET-LDA	2012	Gibbs sampling	LDA	Event segmentation twitter
[186]	Infinite latent harmonic allocation (iLHA)	2012	Expectation-maximization (EM) algorithm	LDA	Multipitch analysis and music information retrieval
			Variational Bayes (VB)	HDP(Hierarchical Dirichlet processes)	
[200]	Mr. LDA	2012	Variational Bayes inference	LDA	Exploring document collections from large scale
				Newton-raphson method MapReduce algorithm	
[146]	FB-LDA, RCB-LDA	2012	Gibbs sampling	LDA	Analyze and track public sentiment variations (on twitter)
[117]	Factorial LDA	2012	Gibbs sampling	LDA	Analysis text in a multi-dimensional structure
[103]	SShLDA	2012	Gibbs sampling	LDA	Topic discovery in data space
				hLDA	
[28]	Utopian	2013	Gibbs sampling	LDA	Visual text analytics

Table 2 (continued)

Author-study	Model	Years	Parameter estimation / inference	Methods	Problem domain
[114]	LDA-GA	2013	Gibbs sampling	LDA Genetic algorithm	Software textual retrieval and analysis
[172]	Multi-aspect sentiment analysis for chinese online social reviews (MSA-COSRs	2013	Gibbs sampling?	LDA	Sentiment analysis and aspect mining of Chinese social reviews
[77]	TopicSpam	2013	Gibbs sampling	LDA	Opinion spam detection
[20]	WT-LDA	2013	Gibbs sampling	LDA	Web service clustering
[128]	Emotion-LDA(ELDA)	2014	Gibbs sampling	LDA	Social emotion classification of online news
[65]	TWILITE	2014	EM algorithm	LDA	Recommendation system for twitter
[31]	Red-LDA	2014	Gibbs-Samplin	LDA	Extract information and and data modeling in patient record notes
[24, 177]	Biterm-topic-modeling(BTM)	2014	Gibbs sampling	LDA	Document clustering for short text
[178]	Trend sensitive-latent Dirichlet allocation (TS-LDA)	2014	Gibbs sampling	LDA Normalized discounted cumulative gain (nDCG) Amazon mechanical turk (AMT)2 platform	Interesting tweets discover for users, recommendation system
[163]	Fine-grained labeled LDA (FL-LDA), Unified fine-grained labeled LDA (UFL-LDA)	2014	Gibbs sampling	LDA	Aspect extraction and review mining

Table 2 (continued)

Author-study	Model	Years	Parameter estimation / inference	Methods	Problem domain
[96, 162]	Regularized latent Dirichlet allocation (rLDA)	2014	Variational Bayes inference	LDA	Automatic image tagging or tag recommendation
[23]	Generative probabilistic aspect mining model (PAMM)	2014	Expectation-maximization (EM)	LDA	Opinion mining and groupings of drug reviews, aspect mining
[201]	AEP-based latent Dirichlet allocation (AEP-LDA)	2014	Gibbs sampling	LDA	Opinion /aspect mining and sentiment word identification
[7]	ADM-LDA	2014	Gibbs sampling	LDA Markov chain	Aspect mining and sentiment analysis
[174]	MRF-LDA	2015	EM algorithm	LDA Markov random field	Exploiting word correlation knowledge
[191]	LightLDA	2015	Gibbs Sampling	LDA	Topic modeling for very large data sizes
[113]	LFLDA,LF-DMM	2015	Gibbs Sampling	LDA	Document clustering for short text
[189]	LFT	2015	Gibbs Sampling	LDA	Semantic community detection
[79]	ATC	2015	EM	LDA	Author community discovery
[80]	FLDA, DFLDA	2015	Gibbs Sampling	LDA	Multi-label document categorization
[157]	Hashtag-LDA	2016	Gibbs sampling	LDA	Hashtag recommendation and find relationships between topics and hashtags
[199]	PMB-LDA	2016	Expectation-maximization (EM)	LDA	Extract the population mobility behaviors for large scale
[70]	Automatic rule generation (LARGen	2016	Gibbs sampling	LDA	Malware analysis and automatic signature generation
[174]	PT-LDA	2016	Gibbs-EM algorithm	LDA	Personality recognition in social network

Table 2 (continued)

Author-study	Model	Years	Parameter estimation / inference	Methods	Problem domain
[82]	Corr-wddCRF	2016	Gibbs sampling	LDA	Knowledge discovery in electronic medical record
[205]	Multi-idiomatic LDA model (MiLDA)	2016	Gibbs sampling	LDA	Content-based recommendation and automatic linking
[25]	Location-aware topic model (LTM)	2016	Gibbs sampling	LDA	Music recommendation
[108]	TopPRF	2016	Gibbs sampling	LDA	Evaluate the relevancy between feedback documents
[127]	Contextual sentiment topic model (CSTM)	2016	Expectation-maximization (EM)	LDA	Emotion classification in social network
[183]	Conceptual dynamic latent Dirichlet allocation (CDLDA)	2016	Gibbs sampling	LDA	Topic detection in conversations
[97]	Multiple-channel latent Dirichlet allocation (MCLDA)	2016	Gibbs sampling	LDA	Find the relations between diagnoses and medications from healthcare data
[122]	Multi-modal event topic model (mmETM)	2016	Gibbs sampling	LDA	Tracking and social event analysis
[43, 44]	Dynamic online hierarchical Dirichlet process model (DOHDP)	2016	Gibbs samplin	LDA	Dynamic topic evolutionary discovery for Chinese social media
[175]	Topicsketch	2016	Gibbs sampling	LDA Ttensor decomposition algorithm Count-Min algorithm	Realtime detection and bursty topics dicovery from twitter
[202]	Fast online EM (FOEM)	2016	Expectation-maximization (Batch EM)	LDA	Big topic modeling
[2]	Joint multi-grain topic sentiment(JMTS)	2016	Gibbs sampling	LDA	Extracting semantic aspects from online reviews
[123]	Character word topic model (CWTM)	2016	Gibbs sampling	LDA	Capture the semantic contents in text documents (Chinese language).

leaf node in BT, called Leaf Topic Hierarchy (LTH). One of the benefits from SSHLDA is that, it can incorporate labeled topics into the generative process of documents. In addition, SSHLDA can automatically explore latent topic in data space, and extend existing hierarchy of showed topics.

2.6 A brief look at past work: research between 2014 to 2015

According to Table 2, some of the popular works published between 2014 and 2015 focused on a variety of topics, such as: Hash/tag discovery [96, 162], opinion mining and aspect mining [7, 23, 163, 201], recommendation system [65, 96, 178].

Biterm Topic Modeling (BTM), Topic modeling over short texts is an increasingly important task due to the prevalence of short texts on the Web. Short texts are popular on today's Web, especially with emergence of social media. Inferring topics from large scale short texts becomes critical. They proposed a novel topic model for short texts, namely the biterm topic model (BTM). This model can well capture topics within short texts by explicitly modeling word co-occurrence patterns in the whole corpus. BTM obtains underlying topics in a set of text-documents and a distribution of global from per topic in each of them with an analysis of the generation of biterms. Their results showed that BTM generates discriminative topic representations as well as rather coherent topics in short texts. The main advantage of BTM prevents the data sparsity issue with learning a global topic distribution [24, 177].

TOT-MMM, introduced a hashtag recommendation that called TOT-MMM, This approach is a hybrid model that combines a temporal clustering component similar to that of the Topics-over-Time (TOT) Model with the Mixed Membership Model (MMM) that was originally proposed for word-citation co-occurrence. This model can capture the temporal clustering effect in latent topics, thereby improving hashtag modeling and recommendations. They developed a collapsed Gibbs sampling (CGS) to approximate the posterior modes of the remaining random variables [96]. The posterior distribution of latent topic equaling k for the nth hashtag in tweet d is given by:

$$
\begin{aligned}
P(z^{h}_{(d_n)} &= k | z^{(m)}_{..}, z^{(h)}_{-dn}, w^{(m)}_{..}, w^{(h)}_{..}, t^{(.)}_{..} \\
&\propto \frac{\beta_h + c^{w^{(h)}_{dn}}_{-dn,k}}{vh\beta h + c^{(h)}_{-dn,k}} \; \frac{\alpha + c^{(d_b)}_{-dn,k} + c^{(d_m)}_{.,k}}{K\alpha + N_{dm} + N_{db} - 1} \\
&\times \frac{t_d^{\psi_{k1}-1}(1-t_d)^{\psi_{k2}-1}}{B(\psi_{k1}, \psi_{k2})}
\end{aligned}
$$

(h) where denotes the number of hashtags type $w^{(b)}_{dn}$ assigned to latent topic k, excluding the hashtag currently undergoing processing; $c^{(h)}_{-dn,k}$ denotes the number of hashtags assigned to latent topic k, excluding the assignment at position d_n; $c^{(dh)}_{-dn,k}$ denotes the number of hashtags assigned to latent topic k in tweet d, excluding the hashtag currently undergoing processing; $c^{(d_m)}_{.,k}$ denotes the number of words assigned to latent topic k in tweet d; V_b is the number of unique hashtags; t_d is the tweet time stamp omitting position subscripts and superscripts (all words and hashtags share the same time stamp); ψ_{k1}, ψ_{k2} are the parameters of the beta distribution for latent topic k.

The probability for a hashtag given the observed words and time stamps is:

$$p\left(w^{h}_{vn}|w^{(m)}_{v.},\mathrm{t}_{\mathrm{v}}\right) = \int p\left(w^{(h)}_{vn}|\theta^{(v)}\right) p\left(\theta^{(v)}|w^{|(m)}_{v.},\ t_{v}\right) d\theta^{(v)}$$

$$s \approx \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K} \varnothing^{(s)}_{h,k,w^{(h)}_{vn}}\theta^{(v)(s)}_{k},$$

where S is the total number of recorded sweeps, and the superscript s marks the parameters computed based on a specific recorded sweep. To provide the top N predictions, they ranked from largest to smallest and output the first N hashtags.

rLDA , the authors introduced a novel probabilistic formulation to obtain the relevance of a tag with considering all the other images and their tags and also they proposed a novel model called regularized latent Dirichlet allocation (rLDA). This model can estimates the latent topics for each document, with making use of other documents. They used a collective inference scheme to estimate the distribution of latent topics and applied a deep network structure to analyze the benefit of regularized LDA [96, 162].

2.7 A brief look from some impressive past works: research in 2016

According to Table 2, some of the popular works published for this year focused on a variety of topics, such as recommendation system [25, 199, 205], opinion mining and aspect mining [7, 23, 163, 192, 201].

A bursty topic on Twitter is one that triggers a surge of relevant tweets within a short period of time, which often reflects important events of mass interest. How to leverage Twitter for early detection of bursty topics has, therefore, become an important research problem with immense practical value. In TopicSketch [175], proposed a sketch-based topic model together with a set of techniques to achieve real-time bursty topic detection from the perspective of topic modeling, that called in this paper TopicSketch.

The bursty topics are often triggered by some events such as some breaking news or a compelling basketball game, which get a lot of attention from people, and "**force**" people to tweet about them intensely. For example, in physics, this "**force**" can be expressed by "**acceleration**", which in our setting describes change of "**velocity**", i.e., arriving rate of tweets. Bursty topics can get significant acceleration when they are bursting, while the general topics usually get nearly zero acceleration. So the "**acceleration**" trick can be used to preserve information of bursty topics but filter out the others. Equation (2) shows how we calculate the "**velocity**" $\hat{v}(t)$ and "acceleration" $\hat{a}(t)$ of words.

$$\hat{v}_{\Delta T} d = \sum_{t_i \leq t} X_i . \frac{\exp((t_i - t)/\Delta T)}{\Delta T} \tag{2}$$

In equation (1), X_i is the frequency of a word (or a pair of words, or a triple of words) in the i-th tweet, t_i is its timestamp. The exponential part in $\hat{v}_{\Delta T}(t)$ works like a soft moving window, which gives the recent terms high weight, but gives low weight to the ones far away, and the smoothing parameter ΔT is the window size. In fact, the authors proposed a novel data sketch which efficiently maintains at a cost of low-level computational of three quantities: the total number of every tweets, the occurrence of per word and also the occurrence of per word pair. Therefore, low calculation costs is one of the advantages from this approach.

Hashtag-LDA, the authors a personalized hashtag recommendation approach is introduced according to the latent topical information in untagged microblogs. This model can enhance influence of hashtags on latent topics' generation by jointly modeling hashtags and words in microblogs. This approach inferred by Gibbs sampling to latent topics and considered a real-world Twitter dataset to evaluation their approach [199]. CDLDA proposed a conceptual dynamic latent Dirichlet allocation model for tracking and topic detection for conversational communication, particularly for spoken interactions. This model can extract dependencies between topics and speech acts. The CDLDA applied hypernym information and speech acts for topic detection and tracking in conversations, and it captures contextual information from transitions, incorporated concept features and speech acts [183].

mmETM, the authors proposed a novel multi modal social event tracking to capture the evolutionary trends from social events and also for generating effective event summary details over time. The mmETM can model the multimodal property of social event and learn correlations among visual modalities and textual to apart the non-visual-representative topics and visual representative topics. This model can work in an online mode with the event consisting of many stories over time and this is a great advantage for getting the evolutionary trends in event tracking and evolution [122].

3 Topic modeling for which the area is used?

With the passage of time, the importance of Topic modeling in different disciplines will be increase. According to previous studies, we present a taxonomy of current approaches topic models based on LDA model and in different subject such as Social Network [54, 94, 95, 105, 161, 188], Software Engineering [19, 46, 89, 90], Crime Science [21, 45, 160] and also in areas of Geographical [34, 39, 56, 63, 139, 147, 184], Political Science [30, 49, 120], Medical/Biomedical [60, 91, 173, 197] and Linguistic science [9, 38, 106, 153, 169] as illustrated by Fig. 2.

3.1 Topic modeling in linguistic science

LDA is an advanced textual analysis technique grounded in computational linguistics research that calculates the statistical correlations among words in a large set of documents to identify and quantify the underlying topics in these documents. In this subsection, we examine some of topic modeling methodology from computational linguistic research, which showed some significant research in Table 3. In [153], employed the distributional hypothesis in various direction and it efforts to cancel the requirement of a seed lexicon

Fig. 2 A vision of the application of topic modeling in various sciences (based on previous works)

Table 3 Impressive works LDA-based in linguistic science

Study	Year	Purpose	Dataset
[153]	2011	Introduce various ways to identify the translation of words among languages [BiLDA].	A wikipedia dataset (Arabic, Spanish, French, Russian and English)
[169]	2010	Obtain term weighting based on LDA	A multilingual dataset
[106]	2013	Present a diversity of new visualization techniques to make concept of topic-solutions	- Dissertation abstracts (1980 to 2010) - 1 million abstracts
[9]	2012	A topic modeling approach, that it consider geographic information	Foursquare dataset
[100]	2014	An approach that is capable to find a document with different language	ALTW2010
[53]	2013	A method for linguistic discovery and conceptual metaphors resources	Wikipedia

as an essential prerequisite for use of bilingual vocabulary and introduce various ways to identify the translation of words among languages. In [38] introduced a method that leads the machine translation systems to relevant translations based on topic-specific contexts and used the topic distributions to obtain topic-dependent lexical weighting probabilities. They considered a topic model for training data, and adapt translation model. To evaluate their approach, they performed experiments on Chinese to English machine translation and show the approach can be an effective strategy for dynamically biasing a statistical machine translation towards relevant translations.

In [106] presented a diversity of new visualization techniques to make the concept of topic-solutions and introduce new forms of supervised LDA, to evaluate they considered a corpus of dissertation abstracts from 1980 to 2010 that belongs to 240 universities in the United States. In [9] developed a standard topic modeling approach, that consider geographic and temporal information and this approach used to Foursquare data and discover the dominant topics in the proximity of a city. Also, the researchers have shown that the abundance of data available in location-based social network (LBSN) enables such models to obtain the topical dynamics in urbanite environments. In [53] have introduced a method for discovery of linguistic and conceptual metaphors resources and built an LDA model on Wikipedia; align its topics to possibly source and aim concepts, they used from both target and source domains to identify sentences as potentially metaphorical. In [100] presented an approach that is capable to find a document with a different language and identify the current language in a document and next step calculate their relative proportions, this approach is based on LDA and used from ALTW2010 as a dataset to evaluation their method.

3.2 Topic modeling in political science

Some topic modeling methods have been adopted in the political science literature to analyze political attention. In settings where politicians have limited time-resources to express their views, such as the plenary sessions in parliaments, politicians must decide what topics to address. Analyzing such speeches can thus provide insight into the political priorities of the politician under consideration. Single membership topic models that assume each speech relates to one topic; have successfully been applied to plenary speeches made in the

105th to the 108th U.S. Senate in order to trace political attention of the Senators within this context over time [18]. In addition, in [49] proposed a new two-layer matrix factorization methodology for identifying topics in large political speech corpora over time and identify both niche topics related to events at a particular point in time and broad, long-running topics. This paper has focused on European Parliament speeches, the proposed topic modeling method has a number of potential applications in the study of politics, including the analysis of speeches in other parliaments, political manifestos, and other more traditional forms of political texts. In [30], the purpose of the study is to examine the various effects of dataset selection with consideration of policy orientation classifiers and built three datasets that each data set include of a collection of Twitter users who have a political orientation.In this approach, the output of an LDA has been used as one of many features as a feed to apply SVM classifier and another part of this method used an LLDA that Considered as a stand-alone classifier. Their assessment showed that there are some limitations to building labels for non-political user categories. Shown some significant research based on LDA to political science in Table 4.

Fang et al. in [42] suggested a new unsupervised topic model based on LDA for contrastive opinion modeling which purpose to find the opinions from multiple views, according to a given topic and their difference on the topic with qualifying criteria, the model called Cross-Perspective Topic (CPT) model. They performed experiments with both qualitative and quantitative measures on two datasets in the political area that include: first dataset is statement records of U.S. senators that show political stances of senators by these

Table 4 Impressive works LDA-based in political science

Study	Year	Purpose	Dataset
[30]	2013	Evaluate the behavioral effects of different databases from political orientation classifiers	-Political Figures Dataset -Politically Active Dataset -Politically Modest Dataset -Conover 2011 Dataset (C2D)
[42]		Introduce a topic model for contrastive opinion modeling	Statement records of U.S. senators
[8]	2012	Detection topics that evoke different reactions from communities that lie on the political spectrum	A collection of blog posts from five blogs: 1. Carpetbagger(CB) thecarpetbaggerreport.com 2. Daily Kos(DK) dailykos.com 3. Matthew Yglesias(MY) yglesias.thinkprogress.org 4. Red State(RS) redstate.com 5. Right wing news(RWN) rightwingnews.com
[18]	2010	Discover the hidden relationships between opinion word and topics words	The statement records of senators through the project vote smart (http://www.votesmart.org)
[140]	2014	Analyze issues related to Korea's presidential election	Project vote smart website (https://votesmart.org/)
[71]	2014	Examine political contention in the U.S. trucking industry	Regulations.gov online portal
[204]	2015	Presented a method for multi-dimensional analysis of political documents	Three Germannational elections (2002, 2005 and 2009)

records, also for the second dataset, extracts of world News Medias from three representative media in U.S (New York Times), China (Xinhua News) and India (Hindu). To evaluate their approach with other models used corrIDA and LDA as two baselines. Yano et al. in [181, 182] applied several probabilistic models based on LDA to predict responses from political blog posts. In more detail, they used topic models LinkLDA and CommentLDA to generate blog data(topics, words of post) in their method and with this model can found a relationship between the post, the commentators and their responses. To evaluate their model, gathered comments and blog posts with focusing on American politics from 40 blog sites.

In [101] introduced a new application of universal sensing based on using mobile phone sensors and used an LDA topic model to discover pattern and analysis of behaviors of people who changed their political opinions, also evaluated to various political opinions for residents of individual , with consider a measure of dynamic homophily that reveals patterns for external political events. To collect data and apply their approach, they provided a mobile sensing platform to capture social interactions and dependent variables of American Presidential campaigns of John McCain and President Barack Obama in last three months of 2008. In [8] analyzed reactions of emotions and suggested a novel model Multi Community Response LDA (MCR-LDA) which in fact is a multi-target and for predicting comment polarity from post content used sLDA and support vector machine classification. To evaluate their approach, provided a dataset of blog posts from five blogs that focus on US politics that was made by [181].

In [18], the authors suggested a generative model to auto discover of the latent associations between opinion words and topics that can be useful for extraction of political standpoints and used an LDA model to reduce the size of adjective words, the authors successfully get that sentences extracted by their model and they shown this model can effectively in different opinions. They were focused on statement records of senators that includes 15, 512 statements from 88 senators from Project Vote Smart WebSite. In [140], examined how social and political issues related to South Korean presidential elections in 2012 on Twitter and used an LDA method to evaluate the relationship between topics extracted from events and tweets. In [204], proposed a method for evaluating and comparing documents, based on an extension of LDA, and used LogicLDA and Labeled LDA approaches for topic modeling in their method. They are considered German National Elections since 1990 as a dataset to apply their method and shown that the use of their method consistently better than a baseline method that simulates manual annotation based on text and keywords evaluation.

3.3 Topic modeling in medical and biomedical

Topic models applied to pure biomedical or medical text mining, researchers have introduced this approach into the fields of medical science. Topic modeling could be advantageously applied to the large datasets of biomedical/medical research, shown some significant research based on LDA to analyze and evaluate content in medical and biomedical in Table 5. In [173] introduced three LDA-like models and found that this model has higher accuracy than the state-of-the-art alternatives. Authors demonstrated that this approach based on LDA could successfully uncover the probabilistic patterns between Adverse drug reaction (ADR) topics and used ADRS database for evaluating their approach. The aim of the authors to predict ADR from a large number of ADR candidates to obtain a drug target. In [197], focused on the issue of professionalized medical recommendations and introduced a new healthcare recommendation system that called iDoctor, that used Hybrid

Table 5 Impressive works LDA-based in medical and biomedical

Study	Year	Purpose/problem domain	Dataset
[60]	2017	Presented three LDA-based models adverse drug reaction prediction	ADReCS database
[109]	2011	Extract biological terminology	-MEDLINE and Bio-Terms Extraction -Chem2Bio2Rdf
[197]	2017	User preference distribution discovery and identity distribution of doctor feature	-Yelp Dataset (Yelp.com)
[170]	2012	-Ranking GENE-DRUG -Detecting relationship between gene and drug	National library of medicine
[116]	2011	Analyzing public health information on twetter	20 disease articles of twitter data
[161]	2013	Analysis of Generated content by user from social networking sites	One million English posted from Facebook's server logs
[60]	2013	Pattern discovery and extraction for clinical processes	A data-set from Zhejiang Huzhou central hospital of China
[91]	2011	Identifying miRNA-mRNA in functional miRNA regulatory modules	Mouse mammary dataset
[194]	2011	Extract common relationship	T2DM clinical dataset
[62]	2012	Extract the latent topic in traditional Chinese medicine document	T2DM clinical dataset

matrix factorization methods for professionalized doctor recommendation. In fact, They adopted an LDA topic model to extract the topics of doctor features and analyzing document similarity. The dataset this article is college from a crowd sourced website that called Yelp. Their result showed that iDoctor can increase the accuracy of health recommendations and it can has higher prediction in users ratings.

In [158], the authors suggested an approach based on LDA that called Bio-LDA that can identify biological terminology to obtain latent topics. The authors have shown that this approach can be applied in different studies such as association search, association predication, and connectivity map generation. And they showed that Bio-LDA can be applied to increase the application of molecular bonding techniques as heat maps. In [170] proposed a topic modeling for rating gene-drug relations by using probabilistic KL distance and LDA that called LDA-PKL and showed that the suggested model achieved better than Mean Average Precision (MAP). They found that the presented method achieved a high Mean Average Precision (MAP) to rating and detecting pharmacogenomics(PGx) relations. To analyze and apply their approach used a dataset from National Library of Medicine. In [116], Presented Ailment Topic Aspect Model (ATAM) to the analysis of more than one and a half million tweets in public health and they were focused on a specific question and specific models; "What public health information can be learned from Twitter?".

In [60] introduced an LDA based method to discover patterns of internal treatment for Clinical processes (CPs), and currently, detect these hidden patterns is one of the most serious elements of clinical process evaluation. Their main approach is to obtain care flow logs and also estimate hidden patterns for the gathered logs based on LDA. Patterns identified can apply for classification and discover clinical activities with the same medical treatment. To experiment the potentials of their approach, used a data-set that collected from

Zhejiang Huzhou Central Hospital of China. In [91] introduced a model for the discovery of functional miRNA regulatory modules (FMRMs) that merge heterogeneous datasets and it including expression profiles of both miRNAs and mRNAs, using or even without using exploit the previous goal binding information. This model used a topic model based on Correspondence Latent Dirichlet Allocation (Corr-LDA). As an evaluation dataset, they perform their method to mouse model expression datasets to study the issue of human breast cancer. The authors found that their model is mighty to obtain different biologically meaningful models. In [194], the authors had a study on Chinese medical (CM) diagnosis by topic modeling and introduced a model based on Author-Topic model to detect CM diagnosis from Clinical Information of Diabetes Patients, and called Symptom-Herb-Diagnosis topic (SHDT) model. Evaluation dataset has been collected from 328 diabetes patients. The results indicated that the SHDT model can discover herb prescription topics and typical symptom for a bunch of important medical-related diseases in comorbidity diseases (such as; heart disease and diabetic kidney) [194].

3.4 Topic modeling in geographical and locations

There is a significant body of research on geographical topic modeling. According to past work, researchers have shown that topic modeling based on location information and textual information can be effective to discover geographical topics and Geographical Topic Analysis. Table 6 shown some significant research based on LDA to topic modeling and content analysis in geographical issues. In [184], this article examines the issue of topic modeling to extract topics from geographic information and GPS-related documents. They suggested a new location text method that is a combination of topic modeling and geographical clustering called LGTA (Latent Geographical Topic Analysis). To test their approaches, they collected a set of data from the website Flickr, according to various topics. In [139] introduced a novel method based on multi-modal Bayesian models to describe social media by merging text features and spatial knowledge that called GeoFolk. As a general outlook, this method can be considered as an extension of Latent Dirichlet Allocation (LDA). They

Table 6 Impressive works LDA-based in geographical and locations

Study	Year	Purpose	Dataset
[139]	2010	Discovering multi-faceted summaries of documents	CoPhIR dataset
[184]	2011	Content management and retrieval	Flicker dataset
[147]	2013	Semantic clustering in very high resolution panchromatic satellite images	A QUICKBIRD image of a suburban area
[39]	2010	Data discovery, evaluation of geographically coherent linguistic regions and find the relationship between topic variation and regional	A twitter dataset
[34]	2008	Geo-located image categorization and georecognition	3013 images Panoramio in France
[196]	2015	Cluster discovery in geo-locations	Reuters-21578
[107]	2014	Discovering newsworthy information from twitter	A small twitter dataset

used the available standard CoPhIR dataset that it contains an abundance of over 54 million Flickr. The GeoFolk model has ability to be used in quality-oriented applications and can be merged with some models from Web 2.0 social. In [147] proposed a multiscale LDA model that is a combination of multiscale image representation and probabilistic topic model to obtain effective clustering VHR satellite images.

In [39], the authors introduced a model that includes two sources of lexical variation: geographical area and topic, in another word, this model can discover words with geographical coherence in different linguistic regions, and find a relationship between regional and variety of topics. To test their model, they gathered a dataset from the website Twitter and also we can say that, also can show from an author's geographic location from raw text [139, 147, 184]. In [34] suggested a statistical model for classification of geo-located images based on latent representation. In this model, the content of a geo-located database able be visualized by means of some few selected images for each geo-category. This model can be considered as an extension of probabilistic Latent Semantic Analysis (pLSA). They built a database of the geo-located image which contains 3013 images (Panoramio), that is related to southeastern France.

In addition, in [78] designed a browsing system (GeoVisNews) and proposed an extended matrix factorization method based on geographic information for ranking locations, location relevance analysis, and obtain intra-relations between documents and locations, called Ordinal Correlation Consistent Matrix Factorization (OCCMF). For evaluation, the authors used a large multimedia news dataset and showed that their system better than Story Picturing Engine and Yahoo News Map.

In this work [196], the authors focused on the issue of identifying textual topics of clusters including spatial objects with descriptions of text. They presented combined methods based on cluster method and topic model to discover textual object clusters from documents with geo-locations. In fact, they used a probabilistic generative model (LDA) and the DBSCAN algorithm to find topics from documents. In this paper, they utilized dataset Reuters-21578 as a dataset for Analysis of their methods. In [107], the authors presented a study on characterizing significant reports from Twitter. The authors introduced a probabilistic model to topic discovery in the geographical topic area and this model can find hidden significant events on Twitter and also considered stochastic variational inference (SVI) to apply gradient ascent on the variable objective with LDA. They collected 2,535 geo-tagged tweets from the Upper Manhattan area of New York. that the KL divergence is a good metric for identifying significant tweet event, but for a large dataset of news articles, the result will be negative.

3.5 Software engineering and topic modeling

Software evolution and source code analysis can be effective in solving current and future software engineering problems. Topic modeling has been used in information retrieval and text mining where it has been applied to the problem of briefing large text corpora. Recently, many articles have been published for evaluating / mining software using topic modeling based on LDA, Table 7 shown some significant research based on LDA to source code analysis and an other related issues in software engineering. In [89], for the first time, the authors used LDA, to extract topics in source code and perform to visualization of software similarity, In other words, LDA uses an intuitive approach for calculation of similarity between source files with obtain their respective distributions of each document over topics. They utilized their method on 1,555 software projects from Apache and SourceForge that includes 19 million source lines of code (SLOC). The authors demonstrated this

Table 7 Impressive works LDA-based in software engineering

Study	Year	Purpose	Dataset
[89]	2007	Mining software and extracted concepts from code	SourceForge and apache (1,555 projects)
[46]	2010	Identifying latent topics and find their relationships in source code	Thirteen open source software systems
[6]	2010	Generating traceability links	ArchStudio software project
[19]	2012	Find relationship between the conceptual concerns in source code.	Source code entities
[90]	2008	Analyzing software evolution	Open source Java projects, Eclipse and ArgoUML
[150]	2009	Automatic categorization of software systems	43 open-source software systems
[98]	2008	Source code retrieval for bug localization	Mozila, Eclipse source code
[99]	2010	Automatic bug localization and evaluate its effectiveness	Open source software such as (Rhino, and Eclipse)
[180]	2017	Detection of malicious android apps	1612 malicious application

approach, can be effective for project organization, software refactoring. In addition, in [46], introduced a new coupling metric based on Relational Topic Models (RTM) that called Relational Topic based Coupling (RTC), that can identifying latent topics and analyze the relationships between latent topic distributions software data. Also, can say that the RTM is an extension of LDA. The authors used thirteen open source software systems for evaluation this metric and demonstrated that RTC has a useful and valuable impact on the analysis of large software systems.

The work in [6] focused on software traceability by topic modeling and proposed a combining approach based on LDA model and automated link capture. They utilized their method to several data sets and demonstrated how topic modeling increase software traceability, and found this approach, able to scale for carried larger numbers from artifacts. In [148] studied about the challenges use of topic models to mine software repositories and detect the evolution of topics in the source code, and suggested the apply of statistical topic models (LDA) for the discovery of textual repositories. Statistical topic models can have different applications in software engineering such as bug prediction, traceability link recovery and software evolution. Chen et al. in [19] used a generative statistical model(LDA) for analyzing source code and find relationships between software defects and software development. They showed LDA can easily scale to large documents and utilized their approach on three large dataset that includes: Mozilla Firefox, and Mylyn, Eclipse. Linsteadet al. in [90] used and utilized Author-Topic models(AT) to analysis in source codes. AT modeling is an extension of LDA model that evaluation and obtain the relationship of authors to topics and applied their method on Eclipse 3.0 source code including of 700,000 code lines and 2,119 source files with considering of 59 developers. They demonstrated that topic models provided the effective and statistical basis for evaluation of developer similarity.

The work in [150] introduced a method based on LDA for automatically categorizing software systems, called LACT. For evaluation of LACT, used 43 open-source software systems in different programming languages and showed LACT can categorization of software systems based on the type of programming language. In addition, in [64, 98], Proposed an approach topic modeling based on LDA model for the purpose of bug localization. Their idea, applied to the analysis of same bugs in Mozilla and Eclipse and result showed that their LDA-based approach is better than LSI to evaluate and analyze of bugs in these source codes. In [180], introduced a topic-specific approach by considering the combination of description and sensitive data flow information and used an advanced topic model based on LDA with GA, for understanding malicious apps, cluster apps according to their descriptions. They utilized their approach on 3691 benign and 1612 malicious application. The authors found Topic-specific, data flow signatures are very efficient and useful in highlighting the malicious behavior.

3.6 Topic modeling in social network and microblogs

Social networks are a rich source for knowledge discovery and behavior analysis. For example, Twitter is one of the most popular social networks that its evaluation and analysis can be very effective for analyzing user behavior and etc. Figure 3, shows a simple idea based on LDA algorithm to building a tag recordation system on Twitter. Recently, researchers have proposed many LDA approaches for analyzing user tweets on Twitter. Weng et al. the authors were concentrated on identifying influential twitterers on Twitter and proposed an approach based on an extension of PageRank algorithm to rate users, called TwitterRank, and used an LDA model to find latent topic information from a large collection of documentation. For evaluation this approach, they prepared a dataset from Top-1000 Singapore-based twitterers, showed that their approach is better than other related algorithms [166]. Hong et al. This paper examines the issue of identifying the Message popularity as measured based on the count of future retweets and sheds. The authors utilized TF-IDF scores and considered it as a baseline, also used Latent Dirichlet Allocation (LDA) to calculate the

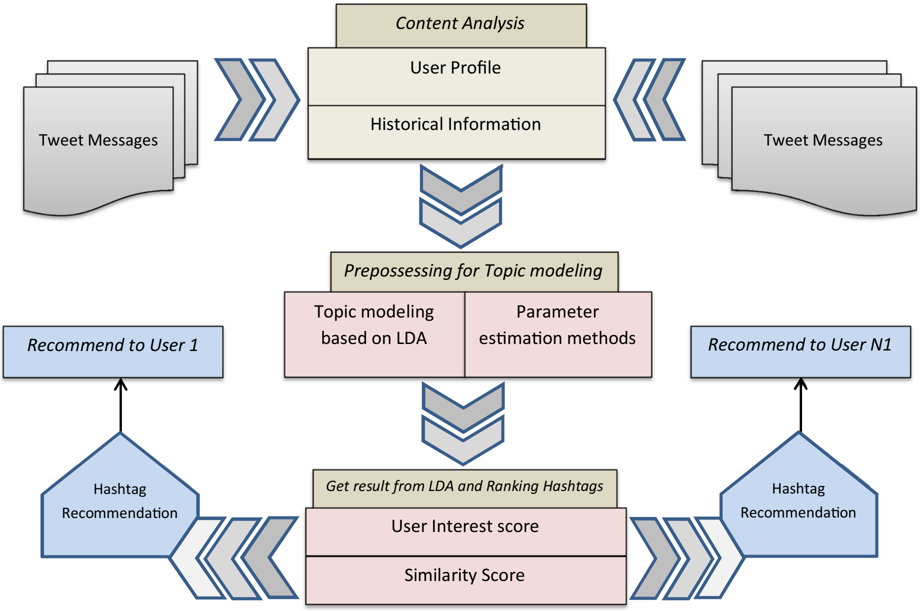


Fig. 3 A simple framework based on LAD to generate tag as a recommendation system on twitter

topic distributions for messages. They collected a dataset that includes 2,541,178 users and 10,612,601 messages and demonstrated that this method can identify messages which will attract thousands of retweets [55]. Table 8 shown some significant research based on LDA to topic modeling and user behavior analysis in social network and microblogs.

The authors in [10] focused on topical recommendations on tweeter and presented a novel methodology for topic discovery of interests of a user on Twitter. In fact, they used a Labeled Latent Dirichlet Allocation (L-LDA) model to discover latent topics between two tweet-sets. The authors found that their method could be better than content based methods for discovery of user-interest. In addition, in [65] suggested, a recommendation system based on LDA for obtaining the behaviors of users on Twitter, called TWILITE. In more detail, TWILTW can calculate the topic distributions of users to tweet messages and also they introduced ranking algorithms in order to recommend top-K followers for users on Twitter.

In [160] investigated in the context of a criminal incident prediction on Twitter. They suggested an approach for analysis and understanding of Twitter posts based a probabilistic language model and also considered a generalized linear regression model. Their evaluation showed that this approach is the capability of predict hit-and-run crimes, only using information that exists in the content of the training set of tweets. This paper [48], they introduced a novel method based LDA model to hashtag recommendation on Twitter that can categories posts with them (hashtags). Lin et al. in [88] investigated the cold-start issue with useing the social information for App recommendation on Twitter and used an LDA model to discovering latent group from “Twitter personalities” to recommendations discovery. For test and experiment, they considered Apple’s iTunes App Store and Twitter as a dataset. Experimental results show, their approach significantly better than other state-of-the-art recommendation techniques.

The authors in [33] presented a technique to analysis and discovered events by an LDA model.The authors found that this method can detect events in inferred topics from tweets

Table 8 Impressive works LDA-based in social network and microblogs

Study	Year	Purpose	Dataset
[166]	2010	Finding influential twitterers on social network(twitter)	Top-1000 Singapore-based twitterers
[10]	2014	Building a topical recommendation systems	A twitter dataset
[65]	2014	A recommendation system for twitter	A twitter dataset
[33]	2012	Analysis and discovered events on twitter	A twitter dataset
[146]	2014	Analyze public sentiment variations regarding a certain tar on twitter	A twitter dataset
[132]	2012	Analysis of the emotional and stylistic distributions on twitter	A twitter dataset
[130]	2016	A topic-enhanced word embedding for twitter sentiment classification	SemEval-2014
[81]	2016	Categorize emotion tendency on Sina Wibo	A Sina Wibo dataset
[83]	2016	Ranking items and summarize news information based on hLDA and maximum spanning tree	A large news Dataset(CNN,BBC,ABC and Google News)

by wavelet analysis. For test and evaluation, they collected 13.6 million tweets from Twitter as a dataset and showed the use of both hashtag names and inferred topics is a beneficial effect in description information for events. In addition, in [121] investigated the issue of how to effectively discover and find health-related topics on Twitter and presented an LDA model for identifies latent topic information from a dataset and it includes 2,231,712 messages from 155,508 users. They found that this method may be a valuable tool for detect public health on Twitter. Tan and et al. in [146] focused on tracking public sentiment and modeling on Twitter. They suggest a topic model approach based on LDA, Foreground and Background LDA to distill topics of the foreground. Also proposed another method for ranking a set of reason candidates in natural language, called Reason Candidate and Background LDA (RCB-LDA). Their results showed that their models can be used to identify special topics and find different aspects. The authors in [132] collected a large corpus from Twitter in seven emotions that includes; disgust, Anger, Fear, Love, Joy, sadness, and surprise. They used a probabilistic topic model, based on LDA, which considered for discovery of emotions in a corpus of Twitter conversations. Srijith et al. in [141] proposed a probabilistic topic model based on hierarchical Dirichlet processes (HDP)) for detection of sub-story. They compared HDP with spectral clustering (SC) and locality sensitive hashing (LSH) and showed that HDP is very effective for story detection data sets, and has an improvement of up to 60% in the F-score.

In [130] proposed a method based on Twitter sentiment classification using topic-enhanced word embedding and also used an LDA model to generate a topic distribution of tweets, considered SVM for classifying tasks in sentiment classification. They used the dataset on SemEval-2014 from Twitter Sentiment Analysis Track. Experiments show that their model can obtain 81.02% in macro F-measure. In [165] focused on examining of demographic characteristics in Trump Followers on Twitter. They considered a negative binomial regression model for modeling the “likes” and used LDA to extract tweets of Trump. They provided evaluations on the dataset US2016 (Twitter) that include a number of followers for all the candidates in the United States presidential election of 2016. The authors demonstrated that topic-enhanced word embedding is very impressive for classification of sentiment on Twitter.

3.7 Crime prediction/evaluation

Over time; definitely, provides further applications for modeling in various sciences. According to recent work, some researchers have applied the topic modeling methods to crime prediction and analysis. Table 9 shown some significant research based on LDA to topic discovery and crime activity analysis. In [21] introduced an early warning system to find the crime activity intention base on an LDA model and collaborative representation classifier (CRC).The system includes two steps: They utilized LDA for learning features and extract the features that can represent from article sources. For the next step, used from achieved features of LDA to classify a new document by collaborative representation classifier (CRC). In [45] used a statistical topic modeling based on LDA to identify discussion topics among a big city in the United States and used kernel density estimation (KDE) techniques for a standard crime prediction . Sharma et al. the authors introduced an approach based on the geographical model of crime intensities to detect the safest path between two locations and used a simple Naive Bayes classifier based on features derived from an LDA model [136].

Table 9 Impressive works LDA-based to crime prediction

Study	Year	Purpose	Dataset
[160]	2012	Automatic semantic analysis on twitter posts	A corpus of tweets from twitter(manual)
[45]	2014	Crime prediction using tagged tweets	City of Chicago data (data.cityofchicago.org)
[21]	2015	Detect the crime activity intention	800 news articles from yahoo Chinese news

4 Open source library and tools / datasets / software packages and tools for the analysis

We need new tools to help us organize, search, and understand these vast amounts of information. In this section, we introduce some impressive datasets and tools for Topic Modeling.

4.1 Library and tools

Many tools for Topic modeling and analysis are available, including professional and amateur software, commercial software, and open source software and also, there are many popular datasets that are considered as a standard source for testing and evaluation. Table 10, show some well-known tools for topic modeling and Table 11, show some well-known datasets for topic modeling. For example; Mallet tools,The MALLET topic model package incorporates an extremely quick and highly scalable implementation of Gibbs sampling, proficient methods for tools and document-topic hyperparameter optimization for inferring topics for new documents given trained models. Topic models provide a simple approach

Table 10 Some well-known tools for topic modeling

Tools	Implementation/ language	Inference/parameter	Source code availability
Mallet [104]	Java	Gibbs sampling	http://mallet.cs.umass.edu/topics.php
TMT [126]	Java	Gibbs sampling	https://nlp.stanford.edu/software/tmt/tmt-0.4/
Mr.LDA [200]	Java	Variational Bayesian inference	https://github.com/lintool/Mr.LDA
JGibbLDA [118]	Java	Gibbs sampling	http://jgibblda.sourceforge.net/
Gensim [129]	Python	Gibbs sampling	https://radimrehurek.com/gensim
TopicXP	Java(Eclipse plugin)		http://www.cs.wm.edu/semeru/TopicXP/
Matlab topic modeling [143]	Matlab	Gibbs sampling	http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
Yahoo_LDA [1]	C++	Gibbsampling	https://github.com/shravanmn/Yahoo_LDA
Lda in R [17]	R	Gibbsampling	https://cran.r-project.org/web/packages/lda/

Table 11 Some well-known Dataset for topic modeling

Dataset	Language	Date of publish	Short-detail	Availability address
Reuters (Reuters21578) [72]	English	1997	Newsletters in various categories	http://kdd.ics.uci.edu/databases/reuters21578/reuters21578
ReutersV 1 (Reuters-Volume I) [73]	English	2004	Newsletters in various categories	
UDI-TwitterCrawl-Aug2012 [76]	English	2012	A twitter dataset from millions of tweets	https://wiki.illinois.edu//wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012
SemEval-2013 Dataset [102]	English	2013	A twitter dataset from millions of tweets	
Wiki10 [134]	English	2009	A wikipedia document in various category	http://nlp.uned.es/social-tagging/wiki10+/
Weibo dataset [195]	Chinese	2013	A popular Chinese microblogging network	
Bag of words	English	2008	A multi dataset (PubMed abstracts, KOS blog, NYTimes news, NIPS full papers, Enron Emails)	https://archive.ics.uci.edu/ml/datasets/Bag+of+Words
CiteUlike [159]	English	2011	A bibliography sharing service of academic papers	http://www.citeulike.org/faq/data.adp
DBLP Dataset [68]	English		A bibliographic database about computer science journals	https://hpi.de/naumann/projects/repeatability/datasets/dblp-dataset.html
HowNet lexicon	Chinese	2000–2013	A Chinese machine-readable dictionary / lexical knowledge	http://www.keenage.com/html/e_index.html
Virastyar , Persian lexicon [5]	Persian	2013	Persian poems electronic lexica	http://ganjoor.net/ http://www.virastyar.ir/data/
NIPS abstracts	English	2016	The distribution of words in the full text of the NIPS conference (1987 to 2015)	https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015
Ch-wikipedia [32, 123]	Chinese		A Chinese corpus from Chinese Wikipedia	
Pascal VOC 2007 [40, 41]	English	2007	A dataset of natural images	http://host.robots.ox.ac.uk/pascal/VOC/voc2007/
AFP_ARB corpus [69]	Arabic	2001	A collection of newspaper articless in Arabic from Agence France Presse	
20Newsgroups4 corpus [131]	English	2008	Newsletters in various categories	http://qwone.com/~jason/20Newsgroups/
New York Times (NYT) dataset [134]	English	2008	Newsletters in various categories	

to analyze huge volumes of unlabeled text. The role of these tools, as mentioned, A “topic” consists of a group of words that habitually happen together. Topic models can associate words with distinguish and similar meanings among uses of words with various meanings and considering contextual clues [142].

For evaluation and testing, according to previous work, researchers have released many dataset in various subjects, size, and dimensions for public access and other future work. So, due to the importance of this research, we examined the well-known dataset from previous work. Table 11, shows lists of some famous and popular datasets in various languages.

5 Discussion and seven important issues in challenges

There are challenges and discussions that can be considered as future work in topic modeling. According to our studies, some issues require further research, which can be very effective and attractive for the future. In this section, we discussed seven important issues and we found that the following issues have not been sufficiently solved. These are the gaps in the reviewed work that would prove to be directions for future work.

5.1 Topics modeling in image processing, image classification and annotation

Image classification and annotation are important problems in computer vision, but rarely considered together and need some intelligent approach for predict classes. For example, an image of a class highway is more likely annotated with words “road” and “traffic”, “car” than words “fish” “scuba” and “boat”. In [27] developed a new probabilistic model for jointly modeling the image, its annotations, and its class label. Their model behaves the class label as a global description of the image and behaves annotation terms as local descriptions of parts of the image. Its underlying probabilistic hypotheses naturally integrate these sources of information. They derive an approximate inference and obtain algorithms based on variational ways as well as impressive approximations for annotating and classifying new images and extended supervised topic modeling (sLDA) to classification problems.

Lienou and et al. in [86] focused on the problem of an image semantic interpretation of large satellite images and used a topic modeling, that each word in a document considering as a segment of image and a document is as an image. For evaluation, they performed experiments on panchromatic QuickBird images. Philbin and et al. in [119] proposed a geometrically consistent latent topic model to detect significant objects, called Latent Dirichlet Allocation (gLDA) and then introduced methods for effectiveness of calculations a matching graph, that images are the nodes and the edge strength in visual content. The gLDA method is able to group images of a specific object despite large imaging variations and can also pick out different views of a single object. In [152] introduced a semi-automatic approach to latent information retrieval. According to the hierarchical structure from the images. They considered a combined investigation using LDA model and invariant descriptors of image region for a visual scene modeling. Wick and et al. in [168] They presented an error correction algorithm using topic modeling based on LDA to Optical character recognition (OCR) error correction. This algorithm including two models: a topic model to calculate the word probabilities and an OCR model for obtaining the probability of character errors. In addition, we can combine Topic models with matrix factorization methods to image understanding, tag assignment and semantic discovery from social image datasets [84, 85].

5.2 Audio, music information retrieval and processing

According to our knowledge, few research works have been done in music information analysis using topic modeling. For example; in [58] proposed a modified version of LDA to process continuous data and audio retrieval. In this model, each audio document includes various latent topics and considered each topic as a Gaussian distribution on the audio feature data. To evaluate the efficiency of their model, used 1214 audio documents in various categories (such as rain, bell, river, laugh, gun, dog and so on). The authors in [112], focused on estimation and estimation of singing characteristics from signals of audio. This paper introduces a topic modeling to the vocal timbre analysis, that each song is considered as a weighted mixture of multiple topics. In this approach, first extracted features of vocal timbre of polyphonic music and then used an LDA model to estimate merging weights of multiple topics. For evaluation, they applied 36 songs that consist of 12 Japanese singers.

5.3 Drug safety evaluation and approaches to improving it

Understanding safety of drug and performance continue to be critical and challenging issues for academia and also it is an important issue in new drug discovery. Topic modeling holds potential for mining the biological documents and given the importance and magnitude of this issue, researchers can consider it as a future work. Bisgin and et al. in [11] introduced an 'in silico' framework to drug repositioning guided through a probabilistic graphical model, that defined a drug as a 'document' and a phenotype form a drug as a 'word'. They applied their approach on the SIDER database to estimate phenome distribution from drugs and identified 908 drugs from SIDER with new capacity indications and demonstrated that the model can be effective for further investigations. Yu and et al. in [187] investigated the issue of drug-induced acute liver failure (ALF) with considering the role of topic modeling to drug safety evaluation, they explored the LiverTox database for drugs discovery with a capacity to cause ALF. Yang and et al. introduced an automatic approach based on keyphrase extraction to detect expressions of consumer health, according to adverse drug reaction (ADRs) in social media. They used an LDA model as a Feature space modeling to build a topic space on the consumer corpus and consumer health expressions mining [11, 179, 187].

5.4 Analysis of comments of famous personalities, social demographics

Public social media and micro-blogging services, most notably Twitter, the people have found a venue to hear and be heard by their peers without an intermediary. As a consequence and helped by the public nature of twitter political scientists now potentially have the means to evaluate and understand narratives that organically form, decline among and spread the public in a political campaign. For this field we can refer to some impressive recent works, for example; Wang and et al. they introduced a framework to derive the topic preferences of Donald Trump's followers on Twitter and used LDA to infer the weighted mixture for each Trump tweet from topics. In addition, we can refer to [3, 59, 137, 165].

5.5 Group discovery and topic modeling

Graph mining and social network analysis in large graphs is a challenging problem. Group discovery has many applications, such as understanding the social structure of organizations, uncovering criminal organizations, and modeling large scale social networks in the Internet community. LDA Models can be an efficient method for discovering latent group structure in large networks. In [54], the authors proposed a scalable Bayesian alternative based on LDA and graph to group discovery in a big real-world graph. For evaluation, they collected three datasets from PubMed. In [188], the authors introduced a generative approach using a hierarchical Bayes model for group discovery in Social Media Analysis that called Group Latent Anomaly Detection (GLAD) model. This model merged the ideas from both the LDA model and Mixture Membership Stochastic Block (MMSB) model.

5.6 User behavior modeling

Social media provides valuable resources to analyze user behaviors and capture user preferences. Since the user generated data (such as users activities, user interests) in social media is a challenge [37, 185], using topic modeling techniques(such as LDA) can contribute to an important role for the discovery of hidden structures related to user behavior in social media. Although some topic modeling approaches have been proposed in user behavior modeling, there are still many open questions and challenges to be addressed. For example; Giri et al. in [47] introduced a novel way using an unsupervised topic model for hidden interests discovery of users and analyzing browsing behavior of users in a cellular network that can be very effective for mobile advertisements and online recommendation systems. In addition, we can refer to [138, 164, 190].

5.7 Visualizing topic models

Although different approaches have been investigated to support the visualization of text in large sets of documents such as machine learning, but it is an open challenge in text mining and visualizing data in big data source. Some of the few studies that have been done, such as [15, 29, 50, 66, 111]. Chuang and et al. in [29] proposed a topic tool based on a novel visualization technique to the evaluation of textual topical in topic modeling, called Termite. The tool can visualize the collection from the distribution of topic term in LDA with considering a matrix layout. The authors used two measures for understanding a topic model of the Useful terms that including: “saliency” and “distinctiveness”. They used the Kullback-Liebler divergence between the topics distribution determined the term for obtain these measures. This tools can increase the interpretations of topical results and make a legible result.

6 Conclusion

Topic models have an important role in computer science for text mining. In Topic modeling, a topic is a list of words that occur in statistically significant methods. A text can be an email, a book chapter, a blog posts, a journal article and any kind of unstructured text. Topic models cannot understand the means and concepts of words in text documents for topic modeling. Instead, they suppose that any part of the text is combined by selecting words from probable baskets of words where each basket corresponds to a topic. The tool goes via this process over and over again until it stays on the most probable distribution of words into baskets which call topics. Topic modeling can provide a useful view of a large collection in terms of the collection as a whole, the individual documents, and the relationships between the documents. In this paper, we investigated scholarly articles highly (between 2003 to 2016) related to Topic Modeling based on LDA in various science. Given the importance of research, we believe this paper can be a significant source and good opportunities for text mining with topic modeling based on LDA for researchers and future works.

Acknowledgements This article has been awarded by the National Natural Science Foundation of China (61170035, 61272420, 81674099, 61502233), the Fundamental Research Fund for the Central Universities (30916011328, 30918015103), and Nanjing Science and Technology Development Plan Project (201805036).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Ahmed A et al (2012) Scalable inference in latent variable models. In: Proceedings of the fifth ACM international conference on web search and data mining. ACM
2. Alam MH, Ryu W-J, Lee S (2016) Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Inf Sci 339:206–223
3. Alashri S et al (2016) An analysis of sentiments on facebook during the 2016 US presidential election. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016. IEEE
4. AlSumait L, Barbara D, Domeniconi C (2008) On-line lda: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Eighth IEEE International Conference on Data Mining, 2008. ICDM'08. IEEE
5. Asgari E, Chappelier J-C (2013) Linguistic Resources and Topic Models for the Analysis of Persian Poems in CLfL@ NAACL-HLT
6. Asuncion HU, Asuncion AU, Taylor RN (2010) Software traceability with topic modeling. In: Proceedings of the 32nd ACM/IEEE international conference on software engineering, vol 1. ACM
7. Bagheri A, Saraee M, De Jong F (2014) ADM-LDA: an aspect detection model based on topic modelling using the structure of review sentences. J Inf Sci 40(5):621–636
8. Balasubramanyan R et al (2012) Modeling polarizing topics: When do different political communities respond differently to the same news? in ICWSM
9. Bauer S et al (2012) Talking places: Modelling and analysing linguistic content in foursquare. In: Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (SocialCom). IEEE
10. Bhattacharya P et al (2014) Inferring user interests in the twitter social network. In: Proceedings of the 8th ACM conference on recommender systems. ACM
11. Bisgin H et al (2014) A phenome-guided drug repositioning through a latent variable model. BMC Bioinforma 15(1):267
12. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
13. Blei DM, Jordan MI (2003) Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. ACM
14. Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning. ACM
15. Chaney AJ-B, Blei DM (2012) Visualizing Topic Models in ICWSM
16. Chang J, Blei DM (2009) Relational topic models for document networks in international conference on artificial intelligence and statistics
17. Chang J (2011) lda: collapsed Gibbs sampling methods for topic models. R
18. Chen B et al (2010) What is an opinion about? Exploring political standpoints using opinion scoring model. In: AAAI
19. Chen T-H et al (2012) Explaining software defects using topic models. In: 2012 9th IEEE working conference on mining software repositories (MSR), IEEE
20. Chen L et al (2013) WT-LDA: user tagging augmented LDA for web service clustering. In: International conference on service-oriented computing. Springer
21. Chen S-H et al (2015) Latent dirichlet allocation based blog analysis for criminal intention detection system. In: 2015 International Carnahan Conference on Security Technology (ICCST). IEEE

22. Chen T-H, Thomas SW, Hassan AE (2016) A survey on the use of topic models when mining software repositories. Empir Softw Eng 21(5):1843–1919
23. Cheng VC et al (2014) Probabilistic aspect mining model for drug reviews. IEEE Trans Knowl Data Eng 26(8):2002–2013
24. Cheng X et al (2014) Btm: topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering 26(1):2928–2941
25. Cheng Z, Shen J (2016) On effective location-aware music recommendation. ACM Transactions on Information Systems (TOIS) 34(2):13
26. Chien J-T, Chueh C-H (2011) Dirichlet class language models for speech recognition. IEEE Transactions on Audio Speech, and Language Processing 19(3):482–495
27. Chong W, Blei D, Li F-F (2009) Simultaneous image classification and annotation. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE
28. Choo J et al (2013) Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. IEEE transactions on visualization and computer graphics 19(12):1992–2001
29. Chuang J, Manning CD, Heer J (2012) Termite: Visualization techniques for assessing textual topic models. In: Proceedings of the international working conference on advanced visual interfaces. ACM
30. Cohen R, Ruths D (2013) Classifying political orientation on twitter: it's not easy! In: ICWSM
31. Cohen R et al (2014) Redundancy-aware topic modeling for patient record notes. PloS one 9(2):e87555
32. Cong Y et al (2012) Cross-modal information retrieval-a case study on Chinese wikipedia. In: International conference on advanced data mining and applications. Springer, Berlin
33. Cordeiro M (2012) Twitter event detection: combining wavelet analysis and topic inference summarization in doctoral symposium on informatics engineering
34. Cristani M et al (2008) Geo-located image analysis using latent representations. in Computer Vision and Pattern Recognition, 2008. CVPR, vol 2008. IEEE, IEEE Conference on
35. Daud A et al (2010) Knowledge discovery through directed probabilistic topic models: a survey. Frontiers of Computer Science in China 4(2):280–301
36. Debortoli S et al (2016) Text mining for information systems researchers: an annotated topic modeling tutorial. CAIS 39:7
37. Diao Q et al (2012) Finding bursty topics from microblogs. In: Proceedings of the 50th annual meeting of the association for computational linguistics: long papers-volume 1. Association for Computational Linguistics
38. Eidelman V, Boyd-Graber J, Resnik P (2012) Topic models for dynamic translation model adaptation. In: Proceedings of the 50th annual meeting of the association for computational linguistics: short papers-volume 2. Association for computational linguistics
39. Eisenstein J et al (2010) A latent variable model for geographic lexical variation. In: Proceedings of the 2010 conference on empirical methods in natural language processings. Association for computational linguistics
40. Everingham M et al (2008) The pascal visual object classes challenge 2007 (voc 2007) results (2007)
41. Everingham M et al (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
42. Fang Y et al (2012) Mining contrastive opinions on political texts using cross-perspective topic model. In: Proceedings of the fifth ACM international conference on web search and data mining. ACM
43. Fu X et al (2015) Dynamic non-parametric joint sentiment topic mixture model. Knowl-Based Syst 82:102–114
44. Fu X et al (2016) Dynamic online HDP model for discovering evolutionary topics from Chinese social texts. Neurocomputing 171:412–424
45. Gerber MS (2014) Predicting crime using Twitter and kernel density estimation. Decis Support Syst 61:115–125
46. Gethers M, Poshyvanyk D (2010) Using relational topic models to capture coupling among classes in object-oriented software systems. In: 2010 IEEE international conference on software maintenance (ICSM). IEEE
47. Giri R et al (2014) User behavior modeling in a cellular network using latent dirichlet allocation. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin
48. Godin F et al (2013) Using topic models for twitter hashtag recommendation. In: Proceedings of the 22nd international conference on world wide web. ACM
49. Greene D, Cross JP (2015) Unveiling the political agenda of the european parliament plenary: a topical analysis. In: Proceedings of the ACM web science conference. ACM
50. Gretarsson B et al (2012) Topicnets: Visual analysis of large text corpora with topic modeling. ACM Transactions on Intelligent Systems and Technology (TIST) 3(2):23
51. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235

52. Guo J et al (2009) Named entity recognition in query. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. ACM
53. Heintz I et al (2013) Automatic extraction of linguistic metaphor with lda topic modeling Inproceedings of the First Workshop on Metaphor in NLP
54. Henderson K, Eliassi-Rad T (2009) Applying latent dirichlet allocation to group discovery in large graphs. In: 2009 Proceedings of the ACM symposium on applied computing. ACM
55. Hong L, Dan O, Davison BD (2011) Predicting popular messages in twitter. In: Proceedings of the 20th international conference companion on world wide web. ACM
56. Hong L, Frias-Martinez E, Frias-Martinez V (2016) Topic models to infer socio-economic maps in AAAI
57. Hu Y et al (2012) ET-LDA: joint topic modeling for aligning events and their twitter feedback. In: AAAI
58. Hu P et al (2014) Latent topic model for audio retrieval. Pattern Recogn 47(3):1138–1143
59. Hou L et al (2015) Newsminer: Multifaceted news analysis for event search. Knowl-Based Syst 76:17–29
60. Huang Z, Lu X, Duan H (2013) Latent treatment pattern discovery for clinical processes. Journal of medical systems 37(2):9915
61. Jagarlamudi J, Daume HIII (2010) Extracting multilingual topics from unaligned comparable corpora. In: ECIR. Springer
62. Jiang Z et al (2012) Using link topic model to analyze traditional chinese medicine clinical symptom-herb regularities. In: 2012 IEEE 14th international conference on e-health networking, applications and services (Healthcom). IEEE
63. Jiang D et al (2015) SG-WSTD: a framework for scalable geographic web search topic discovery. Knowl-Based Syst 84:18–33
64. Jo Y, Oh AH (2011) Aspect and sentiment unification model for online review analysis. In: Proceedings of the fourth ACM international conference on web search and data mining. ACM
65. Kim Y, Shim K (2014) TWILITE: a recommendation system for twitter using a probabilistic model based on latent Dirichlet allocation. Inf Syst 42:59–77
66. Kim M et al (2017) Topiclens: efficient multi-level visual topic exploration of large-scale document collections. IEEE Trans Vis Comput Graph 23(1):151–160
67. Lacoste-Julien S, Sha F, Jordan MI (2009) DiscLDA: discriminative learning for dimensionality reduction and classification. In: Advances in neural information processing systems
68. Lange D, Naumann F (2011) Frequency-aware similarity measures: why Arnold Schwarzenegger is always a duplicate. In: Proceedings of the 20th ACM international conference on Information and knowledge management. ACM
69. Larkey LS, Connell ME (2001) Arabic information retrieval at UMass in TREC-10 in TREC
70. Lee S et al (2016) LARGen: automatic signature generation for Malwares using latent Dirichlet allocation IEEE Transactions on Dependable and Secure Computing
71. Levy KE, Franklin M (2014) Driving regulation: using topic models to examine political contention in the US trucking industry. Soc Sci Comput Rev 32(2):182–194
72. Lewis DD (1997) Reuters-21578 text categorization collection
73. Lewis DD et al (2004) Rcv1: a new benchmark collection for text categorization research. J Mach Learn Res 5(Apr):361–397
74. Li W, McCallum A (2006) Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on machine learning. ACM
75. Li F, Huang M, Zhu X (2010) Sentiment Analysis with Global Topics and Local Dependency in AAAI
76. Li R (2012) Towards social user profiling: unified and discriminative influence model for inferring home locations. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM
77. Li J, Cardie C, Li S (2013) TopicSpam: a topic-model based approach for spam detection in ACL (2)
78. Li Z et al (2013) Enhancing news organization for convenient retrieval and browsing. ACM Transactions on Multimedia Computing. Communications, and Applications (TOMM) 10(1):1
79. Li C et al (2015) The author-topic-community model for author interest profiling and community discovery. Knowl Inf Syst 44(2):359–383
80. Li X, Ouyang J, Zhou X (2015) Supervised topic models for multi-label classification. Neurocomputing 149:811–819
81. Li Y et al (2016) Design and implementation of Weibo sentiment analysis based on LDA and dependency parsing. China Communications 13(11):91–105
82. Li C et al (2016) Hierarchical Bayesian nonparametric models for knowledge discovery from electronic medical records. Knowl-Based Syst 99:168–182

83. Li Z et al (2016) Multimedia news summarization in search. ACM Transactions on Intelligent Systems and Technology (TIST) 7(3):33
84. Li Z, Tang J (2017) Weakly supervised deep matrix factorization for social image understanding. IEEE Trans Image Process 26(1):276–288
85. Li Z, Tang J, Mei T (2018) Deep collaborative embedding for social image understanding. IEEE transactions on pattern analysis and machine intelligence
86. Lienou M, Maitre H, Datcu M (2010) Semantic annotation of satellite images using latent Dirichlet allocation. IEEE Geosci Remote Sens Lett 7(1):28–32
87. Lin CX et al (2010) PET: a statistical model for popular events tracking in social communities. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. ACM
88. Lin J et al, Addressing cold-start in app recommendation: latent user models constructed from twitter followers (2013). In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM
89. Linstead E et al (2007) Mining concepts from code with probabilistic topic models. ACM, Inproceedings of the twenty-second IEEE/ACM international conference on automated software engineering
90. Linstead E, Lopes C, Baldi P (2008) An application of latent Dirichlet allocation to analyzing software evolution. In: 7th international conference on machine learning and applications, 2008. ICMLA'08. IEEE
91. Liu B et al (2010) Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. Bioinformatics 26(24):3105–3111
92. Liu Z et al (2011) Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3):26
93. Liu B, Zhang L (2012) A survey of opinion mining and sentiment analysis. In: Mining text data. Springer, pp 415–463
94. Liu Y, Wang J, Jiang Y (2016) PT-LDA: a latent variable model to predict personality traits of social network users. Neurocomputing 210:155–163
95. Liu Y et al (2016). In: AAAI, Fortune teller: predicting Your Career Path
96. Lu H-M, Lee C-H (2015) The topic-over-time mixed membership model (TOT-MMM): a twitter hashtag recommendation model that accommodates for temporal clustering effects. IEEE Intell Sys 30(1):18–25
97. Lu H-M, Wei C-P, Hsiao F-Y (2016) Modeling healthcare data using multiple-channel latent Dirichlet allocation. J Biomed Inform 60:210–223
98. Lukins SK, Kraft NA, Etzkorn LH (2008) Source code retrieval for bug localization using latent dirichlet allocation. In: 15th working conference on reverse engineering, 2008. WCRE'08. IEEE
99. Lukins SK, Kraft NA, Etzkorn LH (2010) Bug localization using latent dirichlet allocation. Inf Softw Technol 52(9):972–990
100. Lui M, Lau JH, Baldwin T (2014) Automatic detection and language identification of multilingual documents. Transactions of the Association for Computational Linguistics 2:27–40
101. Madan A et al (2011) Pervasive sensing to model political opinions in face-to-face networks. In: International conference on pervasive computing. Springer
102. Manandhar S, Yuret D (2013) Second joint conference on lexical and computational semantics (* sem), volume 2: Proceedings of the seventh international workshop on semantic evaluation (semeval 2013). In: 2nd joint conference on lexical and computational semantics (* SEM), volume 2: proceedings of the 7th international workshop on semantic evaluation (SemEval 2013)
103. Mao X-L et al, SSHLDA: a semi-supervised hierarchical topic model (2012). In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for computational linguistics
104. McCallum AK (2002), A machine learning for language toolkit, Mallet
105. McCallum A, Corrada-Emmanuel A, Wang X (2005) Topic and role discovery in social networks. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp 786–791
106. McFarland DA et al (2013) Differentiating language usage through topic models. Poetics 41(6):607–625
107. McInerney J, Blei DM (2014) Discovering newsworthy tweets with a geographical topic model in NewsKDD: Data Science for News Publishing workshop Workshop in conjunction with KDD2014 the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
108. Miao J, Huang JX, Zhao J (2016) TopPRF: a probabilistic framework for integrating topic space into pseudo relevance feedback. ACM Transactions on Information Systems (TOIS) 34(4):22
109. Millar JR, Peterson GL, Mendenhall MJ (2009) Document clustering and visualization with latent Dirichlet allocation and self-organizing maps in FLAIRS Conference

110. Minka T, Lafferty J (2002) Expectation-propagation for the generative aspect model. In: Proceedings of the eighteenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc
111. Murdock J, Allen C (2015) Visualization Techniques for Topic Model Checking. In: AAAI
112. Nakano T, Yoshii K, Goto M (2014) Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014. IEEE
113. Nguyen DQ et al (2015) Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics 3:299–313
114. Panichella A et al (2013) How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms. In: Proceedings of the 2013 international conference on software engineering. IEEE Press
115. Paul M, Girju R (2010) A two-dimensional topic-aspect model for discovering multi-faceted topics. Urbana 51(61801):36
116. Paul MJ, Dredze M (2011) You are what you tweet: analyzing twitter for public health. Icwsm 20:265–272
117. Paul M, Factorial M. Dredze. (2012) LDA: Sparse multi-dimensional text models in advances in neural information processing systems
118. Phan X-H, Nguyen C-T (2006) Jgibblda: a java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference
119. Philbin J, Sivic J, Zisserman A (2011) Geometric latent dirichlet allocation on a matching graph for large-scale image datasets. Int J Comput Vis 95(2):138–153
120. Preotiuc-Pietro D et al (2017) Beyond binary labels: political ideology prediction of twitter users Inproceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
121. Prier KW et al (2011) Identifying health-related topics on twitter. in International Conference on Social Computing. Springer, Behavioral-Cultural Modeling, and Prediction
122. Qian S et al (2016) Multi-modal event topic model for social event analysis. IEEE Trans Multimedia 18(2):233–246
123. Qin Z, Cong Y, Wan T (2016) Topic modeling of Chinese language beyond a bag-of-words. Computer Speech and Language 40:60–78
124. Ramage D et al (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing: volume 1-volume 1. Association for computational linguistics
125. Ramage D, Manning CD, Dumais S (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
126. Ramage D, Rosen E (2011) Stanford topic modeling toolbox
127. Rao Y (2016) Contextual sentiment topic model for adaptive social emotion classification. IEEE Intell Syst 31(1):41–47
128. Rao Y et al (2014) Building emotional dictionary for sentiment analysis of online news. World Wide Web 17(4):723–742
129. Rehurek R, Sojka P (2011) Gensim-statistical semantics in python
130. Ren Y, Wang R, Ji D (2016) A topic-enhanced word embedding for Twitter sentiment classification. Inf Sci 369:188–198
131. Rennie J (2017) The 20 Newsgroups data set. http
132. Roberts K et al (2012) EmpaTweet: annotating and detecting emotions on twitter. In: LREC
133. Rosen-Zvi M et al (2004) The author-topic model for authors and documents. In: Proceedings of the 20th conference on uncertainty in artificial intelligence. AUAI Press
134. Sandhaus E (2008) The New York times annotated corpus. Linguistic Data Consortium, Philadelphia
135. Savage T et al (2010) Topic XP: exploring topics in source code using latent Dirichlet allocation. In: 2010 IEEE International Conference on software maintenance (ICSM). IEEE
136. Sharma V et al (2015) Analyzing Newspaper Crime Reports for Identification of Safe Transit Paths in HLT-NAACL
137. Shi B et al (2016) Detecting common discussion topics across culture from news reader comments in ACL (1)
138. Siersdorfer S et al (2014) Analyzing and mining comments and comment ratings on the social web. ACM Trans Web (TWEB) 8(3):17
139. Sizov S (2010) Geofolk latent spatial semantics in web 2.0 social media. In: Proceedings of the third ACM international conference on web search and data mining. ACM

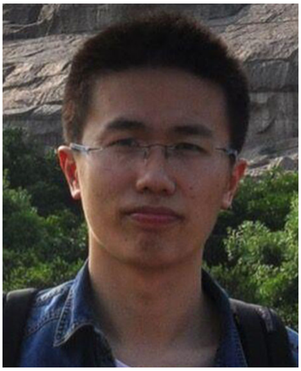
140. Song M, Kim MC, Jeong YK (2014) Analyzing the political landscape of 2012 korean presidential election in twitter. IEEE Intell Syst 29(2):18–26
141. Srijith P et al (2017) Sub-story detection in Twitter with hierarchical Dirichlet processes. Inf Process Manag 53(4):989–1003
142. Steyvers M, Griffiths T (2007) Probabilistic topic models. Handbook of latent semantic analysis 427(7):424–440
143. Steyvers M, Griffiths T (2011) Matlab topic modeling toolbox 1.4. http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
144. Sun X et al (2016) Exploring topic models in software engineering data analysis: a survey. In: 2016 17th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD). IEEE
145. Sun S, Luo C, Chen J (2017) A review of natural language processing techniques for opinion mining systems. Information Fusion 36:10–25
146. Tan S et al (2014) Interpreting the public sentiment variations on twitter. IEEE transactions on knowledge and data engineering 26(5):1158–1170
147. Tang H et al (2013) A multiscale latent Dirichlet allocation model for object-oriented clustering of VHR panchromatic satellite images. IEEE Trans Geosci Remote Sens 51(3):1680–1692
148. Thomas SW (2011) Mining software repositories using topic models. In: Proceedings of the 33rd international conference on software engineering. ACM
149. Thomas SW et al (2011) Modeling the evolution of topics in source code histories. In: Proceedings of the 8th working conference on mining software repositories. ACM
150. Tian K, Revelle M, Poshyvanyk D (2009) Using latent dirichlet allocation for automatic categorization of software. In: 6th IEEE International working conference on mining software repositories, 2009. MSR'09. IEEE
151. Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th international conference on world wide web. ACM
152. Vaduva C, Gavat I, Datcu M (2013) Latent Dirichlet allocation for spatial analysis of satellite images. IEEE Trans Geosci Remote Sens 51(5):2770–2786
153. Vulic I, De Smet W, Moens M-F (2011) Identifying word translations from comparable corpora using latent topic models. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers-volume 2. Association for computational linguistics
154. Wallach HM, Mimno DM, McCallum A (2009) Rethinking LDA: why priors matter. In: Advances in neural information processing systems
155. Wang X, McCallum A (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
156. Wang C, Blei DM (2009) Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In: Advances in neural information processing systems
157. Wang Y, Mori G (2011) Max-margin latent Dirichlet allocation for image classification and annotation. In: BMVC
158. Wang H et al (2011) Finding complex biological relationships in recent PubMed articles using Bio-LDA. PloS one 6(3):e17243
159. Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM
160. Wang X, Gerber MS, Brown DE (2012) Automatic Crime Prediction Using Events Extracted from Twitter Posts. SBP 12:231–238
161. Wang Y-C, Burke M, Kraut RE (2013) Gender, topic, and audience response: an analysis of user-generated content on facebook. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM
162. Wang J et al (2014) Image tag refinement by regularized latent Dirichlet allocation. Comput Vis Image Underst 124:61–70
163. Wang T et al (2014) Product aspect extraction supervised with online domain knowledge. Knowl-Based Syst 71:86–100
164. Wang S et al (2014) Cross media topic analytics based on synergetic content and user behavior modeling. In: IEEE International Conference on Multimedia and Expo (ICME), 2014. IEEE
165. Wang Y et al (2016) Catching fire via" Likes": inferring topic preferences of trump followers on twitter. In: ICWSM

166. Weng J et al (2010) Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. ACM
167. Weng J, Lee B-S (2011) Event detection in twitter. ICWSM 11:401–408
168. Wick M, Ross M, Learned-Miller E (2007) Context-sensitive error correction: using topic models to improve OCR. In: 9th international conference on document analysis and recognition, 2007. ICDAR 2007. IEEE
169. Wilson AT, Chew PA (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. Association for Computational Linguistics
170. Wu Y et al (2012) Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. In: Pacific symposium on biocomputing. NIH Public Access
171. Wu H et al (2012) Locally discriminative topic modeling. Pattern Recogn 45(1):617–625
172. Xianghua F et al (2013) Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowl-Based Syst 37:186–195
173. Xiao C et al (2017) Adverse drug reaction prediction with symbolic latent dirichlet allocation in AAAI
174. Xie P, Yang D, Xing EP (2015) Incorporating word correlation knowledge into topic modeling in HLT-NAACL
175. Xie W et al (2016) Topicsketch: real-time bursty topic detection from twitter. IEEE Trans Knowl Data Eng 28(8):2216–2229
176. Xu Z et al (2017) Crowdsourcing based social media data analysis of urban emergency events. Multimedia Tools and Applications 76(9):11567–11584
177. Yan X et al (2013) A biterm topic model for short texts. In: Proceedings of the 22nd international conference on world wide web. ACM
178. Yang M-C, Rim H-C (2014) Identifying interesting Twitter contents using topical analysis. Expert Syst Appl 41(9):4330–4336
179. Yang M, Kiang M (2015) Extracting Consumer Health Expressions of Drug Safety from Web Forum. In: 2015 48th Hawaii international conference on system sciences (HICSS). IEEE
180. Yang X et al (2017) Characterizing malicious Android apps by mining topic-specific data flow signatures Information and Software Technology
181. Yano T, Cohen WW, Smith NA (2009) Predicting response to political blog posts with topic models. In: Proceedings of human language technologies: the 2009 annual conference of the north american chapter of the association for computational linguistics. Association for computational linguistics
182. Yano T, Smith NA (2010) What's worthy of comment? content and comment volume in political blogs in ICWSM
183. Yeh J-F, Tan Y-S, Lee C-H (2016) Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation. Neurocomputing 216:310–318
184. Yin Z et al (2011) Geographical topic discovery and comparison. In: Proceedings of the 20th international conference on world wide web. ACM
185. Yin H et al (2014) A temporal context-aware model for user behavior modeling in social media systems. In: Proceedings of the ACM SIGMOD international conference on Management of data, 2014. ACM
186. Yoshii K, Goto M (2012) A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation. IEEE Transactions on Audio. Speech, and Language Processing 20(3):717–730
187. Yu K et al (2014) Mining hidden knowledge for drug safety assessment: topic modeling of LiverTox as a case study. BMC Bioinforma 15(17):S6
188. Yu R, He X, Liu Y (2015) Glad: group anomaly detection in social media analysis. ACM Transactions on Knowledge Discovery from Data (TKDD) 10(2):18
189. Yu X, Yang J, Xie Z-Q (2015) A semantic overlapping community detection algorithm based on field sampling. Expert Syst Appl 42(1):366–375
190. Yuan B et al (2014). In: International conference on web information systems engineering. Springer, Berlin
191. Yuan J et al (2015) Lightlda: big topic models on modest computer clusters. In: Proceedings of the 24th international conference on world wide web. International world wide web conferences steering committee
192. Zhai Z, Liu B, Xu H, Jia P (2011) Constrained LDA for grouping product features in opinion mining. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, pp 448–459
193. Zhang H et al (2007) Probabilistic community discovery using hierarchical latent gaussian mixture model. In: AAAI
194. Zhang X-P et al (2011) Topic model for chinese medicine diagnosis and prescription regularities analysis: case on diabetes. Chinese Journal Of Integrative Medicine 17(4):307–313

195. Zhang J et al (2013) Social Influence Locality for Modeling Retweeting Behaviors in IJCAI
196. Zhang L, Sun X, Zhuge H (2015) Topic discovery of clusters from documents with geographical location. Concurrency and Computation: Practice and Experience 27(15):4015–4038
197. Zhang Y et al (2017) iDoctor: personalized and professionalized medical recommendations based on hybrid matrix factorization. Futur Gener Comput Syst 66:30–35
198. Zhao WX et al (2011) Comparing twitter and traditional media using topic models. In: European conference on information retrieval. Springer
199. Zhao F et al (2016) A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. Futur Gener Comput Syst 65:196–206
200. Zhai K et al (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
201. Zheng X et al (2014) Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowl-Based Syst 61:29–47
202. Zeng J, Liu Z-Q, Cao X-Q (2016) Fast online EM for big topic modeling. IEEE Trans Knowl Data Eng 28(3):675–688
203. Zhu J, Ahmed A, Xing EP (2009) MedLDA: maximum margin supervised topic models for regression and classification. In: Proceedings of the 26th annual international conference on machine learning. ACM
204. Zirn C, Stuckenschmidt H (2014) Multidimensional topic analysis in political texts. Data and Knowledge Engineering 90:38–53
205. Zoghbi S, Vulic I, Moens M-F (2016) Latent Dirichlet allocation for linking user-generated content and e-commerce data. Inf Sci 367:573–599

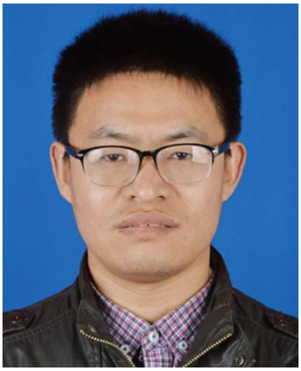
Hamed Jelodar born in 1990, is a Ph.D. candidate in Nanjing University of Science and Technology. His main research interests include web information detecting and natural language processing.

Yongli Wang born in 1974, received the Ph.D. degree from the South-east University in 2006. He is currently a professor in the Department of Computer Science and Engineering at Nanjing University of Science and Technology, China. He is the author or coauthor of more than 70 papers in database and leading pattern recognition journals and conferences. His research includes pattern recognition, machine learning, data streams management, mobile computing, cyber physical system, and health care monitoring, and is supported by the NCSF and other agencies. For his research activities, he also spent extended periods of time at Drexel University, USA, Brunel University, UK and Fraunhofer Institutes, Germany. He is a member of the ACM and the IEEE.



Chi Yuan born in 1990, is a Ph.D. candidate in Nanjing University of Science and Technology. His main research interests include bioinformatics, machine learning, and massive data analysis.

Xia Feng born in 1993, is a master candidate in Nanjing University of Science and Technology. Her main research include machine learning and data analysis.



Xiahui Jiang born in 1978, is a Ph.D. candidate in Nanjing University of Science and Technology. His main research interests include machine learning and pattern recognition.

Yanchao Li born in 1990, is a Ph.D. candidate in Nanjing University of Science and Technology. His main research interests include machine learning, data quality, and massive data analysis.

Liang Zhao is a master student in Nanjing University of Science and Technology. His main research interests include web information detecting and data mining.