

Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey

Hamed Jelodar¹, Yongli Wang¹, Chi Yuan¹, Xia Feng²
Department of Computer Science and Engineering
Nanjing University of Science and Technology,
Nanjing - 210094, China
{Jelodar, Yongliwang, Yuanchi}@njjust.edu.cn¹, 779477284@qq.com²

Abstract:

Topic modeling is one of the most powerful techniques in text mining for data mining, latent data discovery, and finding relationships among data, 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 modeling, which Latent Dirichlet allocation (LDA) is one of the most popular methods in this field. Researchers have proposed various models based on the LDA in topic modeling. According to previous work, this paper can be very useful and valuable for introducing LDA approaches in topic modeling. In this paper, we investigated scholarly articles highly (between 2003 to 2016) related to Topic Modeling based on LDA to discover the research development, current trends and intellectual structure of topic modeling. Also, we summarize challenges and introduce famous tools and datasets in topic modeling based on LDA.

Keywords: Topic modelling, Gibbs Sampling, Latent Dirichlet Allocation, Expectation Maximization, LDA

1. Introduction

Topic models (TM) are a well-know and significant modern machine learning technology that has been widely used in text mining, network analysis and genetics, and more other domains. 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 (Zhang et al., 2017) (Jiang et al., 2012) (Paul and Dredze, 2011) (Wu et al., 2012b) , software engineering (Linstead et al., 2007) (Gethers and Poshyvanyk, 2010) (Asuncion et al., 2010) (Thomas, 2011) (Thomas et al., 2011), geography (Cristani et al., 2008) (Eisenstein et al., 2010) (Tang et al., 2013) (Yin et al., 2011) (Sizov, 2010), political science (Chen et al., 2010) (Cohen and Ruths, 2013) (Greene and Cross, 2015) .

From an applied perspective in the field of political science, **Greene et al.** 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 (Greene and Cross, 2015). Other researchers have also 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 [17]. **Fang et al.**

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 records, 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 (Fang et al., 2012).

Another group of researchers focused on topic modeling in software Engineering, **Linstead et al.** For the first time, they used LDA, to extract topics in source code and perform to visualization of software similarity, In other words, LDA use 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 approach, can be effective for project organization, software refactoring (Linstead et al., 2007). **Tian et al.** 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 type of programming language (Tian et al., 2009). **Lukinet al.** 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 of bugs in these source codes (Lukins et al., 2008, Lukins et al., 2010).

An analysis of geographic information is another issue that can be referred to Sizov **et al.** 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 it 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 (Sizov, 2010). **Yin et al.** This article examines the issue of topic modeling to extract the topics from geographic information and GPS-related documents. They proposed three strategies of modeling geographical topics including , text-driven model, location-driven model and a novel joint model called LGTA (Latent Geographical Topic Analysis) that is a combination of topic modeling and geographical clustering. To test their approaches, they collected a set of data from the website Flickr, according to various topics (Yin et al., 2011).

In other view, According to our knowledge, most of the previous studies had various goals, such as: Source code analysis (Linstead et al., 2007) (Lukins et al., 2010) (Linstead et al., 2008) (Tian et al., 2009) (Chen et al., 2012) (Gethers and Poshyvanyk, 2010) (Savage et al., 2010), Opinion and aspect Mining (Chen et al., 2010) (Zheng et al., 2014) (Cheng et al., 2014) (Zhai et al., 2011) (Bagheri et al., 2014) (Wang et al., 2014c) (Xianghua et al., 2013, Jo and Oh, 2011) (Paul and Girju, 2010) (Titov and McDonald, 2008), Event detection (Qian et al., 2016) (Hu et al., 2012, Weng and Lee, 2011) (Lin et al., 2010), image classification (Cristani et al., 2008) (Wang and Mori, 2011), system recommendation (Zoghbi et al., 2016) (Cheng and Shen, 2016) (Zhao et al., 2016) (Lu and Lee, 2015) (Wang et al., 2014a) (Yang and Rim, 2014) (Kim and Shim, 2014) and emotion classification (Roberts et al., 2012) (Rao,

2016) (Rao et al., 2014), etc. For example in gforecommendation system, **Zhao and et al.** 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(Zhao et al., 2016). **Jin and et al.** The authors focused on the issue of tag recommendation. They proposed hybrids approach based on a combination of Language Model (LM) and LDA for tag recommendation in terms of topic knowledge. The authors used a subset from Bibsonomy dataset that including; 14,443 resources, 33,256 words 1,185 users, 13,276 tags, and 262,445 bookmarks in total. Finally, they found that combination of keyword and topic layer based approaches can be significantly effective to recommend new tags.

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. 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.

2.3 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(Blei et al., 2003), 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 document d having N_d words ($d \in \{1, \dots, M\}$), LDA models D according to the following generative process [4]:

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

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

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

(i) Select a topic z_n from θ_d .

(ii) Select a word w_n from φ_{z_n} .

In above generative process, words in documents are the only observed variables while others are latent variables (φ and θ) and hyper parameters (α and β). In order to infer the latent variables and hyper parameters, the probability of observed data D is computed and maximized as follows:

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

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 collection of documents produce a vocabulary that is then used to generate the latent 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 the 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.

2.4 Parameter estimation, Inference, Training for LDA

Various methods have been proposed to estimate LDA parameters, such as variational method (Blei et al., 2003), expectation propagation (Minka and Lafferty, 2002) and Gibbs sampling (Griffiths and Steyvers, 2004).

- **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 (Xie et al., 2016) (Lu et al., 2016) (Yeh et al., 2016) (Rao, 2016) (Miao et al., 2016) (Panichella et al., 2013, Zhao et al., 2011) (Jagarlamudi and Daumé III, 2010) (Tian et al., 2009) (Ramage et al., 2009).
- **Expectation-Maximization (EM) algorithm** 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 (Zhu et al., 2009) (Guo et al., 2009) (Chang and Blei, 2009) (Blei and Jordan, 2003).
- **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 (Zhai et al., 2012) (Asuncion et al., 2010) (Chien and Chueh, 2011).



Fig1. Taxonomy of methods based on extension LDA, considered some of the impressive works

2.2.1. 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. Undeniably, this period (2003 to 2009) is very important because key and baseline approaches were introduced, such as: Corr_LDA, Author-Topic Model , DTM and , RTM etc .

DTM, Dynamic Topic Model (DTM) is introduced by Blei and Lafferty as an extension of LDA that this model can obtain the evolution of topics over time in a sequentially arranged corpus of documents and exhibits the evolution of word-topic distribution which causes it easy to vision the topic trend(Blei and Lafferty, 2006). **Lable LDA**, Labeled LDA (LLDA) is another of LDA extension which suppose that each document has a set of known labels (Ramage et al., 2009). 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 .

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 (Zhu et al., 2009). **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 the 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 (Chang and Blei, 2009).

Table 1. Some impressive articles based on LDA: between 2003- 2009

Author-study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Blei and Jordan, 2003)	Corr_LDA	2003	Variational EM	LDA	Image annotation and retrieval
(Rosen-Zvi et al., 2004)	Author-Topic Model	2004	Gibbs Sampling	LDA	Find the relationships between authors, documents, words, and topics
(McCallum et al., 2005)	AuthorRecipient-Topic (ART)	2005	Gibbs Sampling	-LDA -Author-Topic (AT)	Social network analysis and role discovery
(Blei and Lafferty, 2006)	Dynamic topic model(DTM)	2006	Kalman variational algorithm	LDA -Galton-Watson process	Provide a dynamic model for evolution of topics
(Wang and McCallum, 2006)	Topics over Time(TOT)	2006	Gibbs Sampling	LDA	Capture word co-occurrences and localization in continuous time.
(Li and McCallum, 2006)	Pachinko Allocation Model, PAM	2006	Gibbs Sampling	-LDA - a directed acyclic graph method	Capture arbitrary topic correlations
(Zhang et al., 2007)	GWN-LDA	2007	Gibbs sampling	LDA hierarchical Bayesian algorithm	Probabilistic community profile Discovery in social network
(AlSumait et al., 2008)	On-line lda (OLDA)	2008	Gibbs Sampling	-LDA -Empirical Bayes method	Tracking and Topic Detection
(Titov and McDonald, 2008)	MG-LDA	2008	Gibbs sampling	LDA	Sentiment analysis in multi-aspect
(Ramage et al., 2009)	Labeled LDA	2009	Gibbs Sampling	LDA	Producing a labeled document collection.
(Chang and Blei, 2009)	Relational Topic Models	2009	Expectation-maximization (EM)	LDA	Make predictions between nodes , attributes, links structure
(Wang and Blei, 2009)	HDP-LDA	2009	Gibbs Sampling	LDA	
(Lacoste-Julien et al., 2009)	DiscLDA	2009	Gibbs Sampling	LDA	Classification and

					dimensionality reduction in documents
(Tian et al., 2009)	LACT	2009	Gibbs sampling	LDA	Automatic Categorization of Software systems
(Wallach et al., 2009)	Rethinking LDA	2009	Gibbs Sampling	LDA	Data Discovery
(Guo et al., 2009)	WS-LDA	2009	Expectation-maximization (EM)	-LDA -query log	Query log mining
(Millar et al., 2009)	LDA-SOM	2009	Gibbs sampling	-LDA -self-organizing maps	Clustering and visualization of large documents
(Zhu et al., 2009)	MedLDA	2009	Expectation-maximization (EM)	-LDA -max-margin learning - support vector regression (SVR)	Regression and Classification

2.2.2. A brief look at past work: Research between 2010 to 2011

Eighteenth approaches are summarized in this subsection, where tenth are published in 2010 and Eighth in 2011. According to the Table 2, used LDA model for variety subjects, such as: Scientific topic discovery (Paul and Girju, 2010), Source code analysis (Savage et al., 2010), Opinion Mining (Zhai et al., 2011), Event detection (Lin et al., 2010), Image Classification (Wang and Mori, 2011) .

Sizov et al. 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 it 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 (Sizov, 2010).

Z. Zhai et al. use 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 (Zhai et al., 2011).

Wang et al. they 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 (Wang et al., 2011).

Table 2. Some impressive articles based on LDA: between 2010- 2011

Author-study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Sizov, 2010)	GeoFolk	2010	Gibbs sampling	LDA	Content management and retrieval of spatial information
(Jagarlamudi and Daumé III, 2010)	JointLDA	2010	Gibbs Sampling	-LDA -bag-of-word model	Mining multilingual topics
(Paul and Girju, 2010)	Topic-Aspect Model (TAM)	2010	Gibbs Sampling	-LDA -SVM	Scientific topic discovery
(Li et al., 2010)	Dependency-Sentiment-LDA	2010	Gibbs Sampling	LDA	Sentiment classification
(Savage et al., 2010)	TopicXP	2010		LDA	Source code analysis
(Zhai et al., 2011)	constrained-LDA	2010	Gibbs Sampling	LDA	Opinion Mining and Grouping Product Features
(Lin et al., 2010)	PET - Popular Events Tracking	2010	Gibbs Sampling	LDA	Event analysis in social network
(Weng and Lee, 2011) (Lin et al., 2010)	EDCoW	2010		-LDA - Wavelet Transformation	Event analysis in Twitter
(Wang et al., 2011)	Bio-LDA.	2011	Gibbs Sampling	-LDA	Extract biological terminology
(Zhao et al., 2011)	Twitter-LDA	2011	Gibbs Sampling	- LDA - author-topic model -PageRank	Extracting topical keyphrases and analyzing Twitter content
(Wang and Mori, 2011)	max-margin latent Dirichlet allocation (MMLDA)	2011	variational inference	LDA - SVM	Image Classification and Annotation
(Jo and Oh, 2011)	Sentence-LDA	2011	Gibbs sampling	LDA	Aspects and sentiment discovery for web review
(Liu et al., 2011)	PLDA+	2011	Gibbs sampling	-LDA -weighted round-robin	Reduce inter-computer communication time
(Chien and Chueh, 2011)	Dirichlet class language model (DCLM),	2011	variational Bayesian EM (VB-EM) algorithm	speech recognition and exploitation of language models	Dirichlet class language model (DCLM),

2.2.3. A brief look at past work: Research between 2012 to 2013

According to Table 3, some of the popular works published between 2012 and 2013 focused on a variety of topics, such as music retrieve (Yoshii and Goto, 2012), opinion and aspect mining(Li et al., 2013), Event analysis(Hu et al., 2012).

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(Hu et al., 2012). **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 (Zhai et al., 2012). **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 to measuring the distance between documents and used Fast collapsed Gibbs sampling to approximate the posterior distributions of parameters(Panichella et al., 2013).

Table1 3. Some impressive articles based on LDA: between 2012- 2013

Author-study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Wu et al., 2012a)	locally discriminative topic model (LDTM)	2012	Expectation-maximization (EM)	LDA	document semantic analysis
(Wu et al., 2012a)	locally discriminative topic model (LDTM)	2012	Expectation-maximization (EM)	LDA	document semantic analysis
(Hu et al., 2012)	ET-LDA	2012	Gibbs sampling	LDA	event segmentation Twitter
(Yoshii and Goto, 2012)	infinite latent harmonic allocation (iLHA)	2012	expectation-maximization (EM) algorithm /variational Bayes (VB)	-LDA -variational Bayes(VB) - HDP(Hierarchical Dirichlet processes)	multipitch analysis and music information retrieval
(Zhai et al., 2012)	Mr. LDA	2012	Variational Bayes inference	-LDA -Newton-Raphson method - MapReduce Algorithm	exploring document collections from large scale
(Tan et al., 2014)	FB-LDA , RCB-LDA	2012	Gibbs Sampling	LDA	analyze and track public sentiment variations (on

					twitter)
(Paul and Dredze, 2012)	factorial LDA	2012	Gibbs sampling	LDA	analysis text in a Multi-Dimensional multi-dimensional structure
(Mao et al., 2012)	SshLDA	2012	Gibbs sampling	LDA hLDA	Topic discovery in data space
(Choo et al., 2013)	Utopian	2013	Gibbs sampling	LDA	visual text analytics
(Panichella et al., 2013)	LDA-GA	2013	Gibbs sampling	LDA -Genetic Algorithm	software textual retrieval and analysis
(Xianghua et al., 2013)	Multi-aspect Sentiment Analysis for Chinese Online Social Reviews (MSA-COSRs)	2013	Gibbs Sampling	LDA	sentiment analysis And aspect mining of of Chinese social reviews
(Li et al., 2013)	TopicSpam	2013	Gibbs sampling	LDA	opinion spam detection
(Chen et al., 2013)	WT-LDA	2013	Gibbs sampling	LDA	Web Service Clustering

2.2.4. A brief look at past work: Research between 2014 to 2015

According to Table 4, some of the popular works published between 2014 and 2015 focused on a variety of topics, such as: Hash/tag discovery (Wang et al., 2014a) (Lu and Lee, 2015), opinion mining and aspect mining (Bagheri et al., 2014) (Zheng et al., 2014) (Cheng et al., 2014) (Wang et al., 2014c), system recommendation(Lu and Lee, 2015) (Yang and Rim, 2014) (Kim and Shim, 2014).

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 the 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 the topics within short texts by explicitly modeling word co-occurrence patterns in the whole corpus(Cohen et al., 2014).

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) for approximate the posterior modes of the remaining random variables(Lu and Lee, 2015). The posterior distribution of latent topic equaling k for the n th hashtag in tweet d is given by:

$$P(z_{dn}^h = k | z_{..}^{(m)}, z_{-dn}^{(h)}, w_{..}^{(m)}, w_{..}^{(h)}, t_{..}^{(\cdot)}) \propto \frac{\beta_{h+c_{-dn,k}^{(h)}}}{vh\beta_{h+c_{-dn,k}^{(h)}}} \frac{\alpha + c_{-dn,k}^{(db)} + c_{.,k}^{(dm)}}{K\alpha + N_{dm} + N_{db} - 1}$$

$$\times \frac{t_d^{\psi_{k1}-1} (1-t_d)^{\psi_{k2}-1}}{B(\psi_{k1}, \psi_{k2})} \quad (1)$$

Type equation here.

(h) where denotes the number of hashtags type $w_{dn}^{(b)}$ assigned to latent topic k , excluding the hashtag currently undergoing processing; $c_{-dn,k}^{(h)}$ denotes the number of hashtags assigned to latent topic k , excluding the assignment at position d_n ; $c_{-dn,k}^{(dh)}$ denotes the number of hashtags assigned to latent topic k in tweet d , excluding the hashtag currently undergoing processing; $c_{.,k}^{(dm)}$ 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:

$$p(w_{vn}^h | w_{v.}^{(m)}, t_v) = \int p(w_{vn}^{(h)} | \theta^{(v)}) p(\theta^{(v)} | w_{v.}^{(m)}, t_v) d\theta^{(v)}$$

$$s \approx \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} \sum_{k=1}^K \phi_{h,k,w_{vn}^{(h)}}^{(s)} \theta_k^{(v)(s)}, \quad (2)$$

where \mathcal{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 $p(w_{vn}^{(h)} | \cdot)$ 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 (Wang et al., 2014a) (Lu and Lee, 2015).

Table1 4. Some impressive articles based on LDA: between 2014-2015

Author-study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Rao et al., 2014)	emotion LDA (ELDA)	2014	Gibbs sampling	LDA	Social emotion classification of online news
(Kim and Shim, 2014)	TWILITE	2014	EM algorithm	LDA	Recommendation system for Twitter
(Cohen et al., 2014)	Red-LDA	2014	Gibbs-Samplin	-LDA	Extract information and and data modeling... in Patient Record Notes
(Cohen et al., 2014)	Biterm Topic Modeling (BTM)	2014	Gibbs sampling	LDA	Document clustering for short text
(Yang and Rim, 2014)	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, system recommendation
(Wang et al., 2014c)	Fine-grained Labeled LDA (FL-LDA), Unified Fine-grained Labeled LDA (UFL-LDA)	2014	Gibbs sampling	LDA	Aspect extraction and review mining
(Wang et al., 2014a) (Lu and Lee, 2015)	regularized latent Dirichlet allocation (rLDA)	2014	Variational Bayes inference	LDA	Automatic image tagging or tag recommendation
(Cheng et al., 2014)	generative probabilistic aspect mining model (PAMM)	2014	Expectation-maximization (EM)	LDA	Opinion mining and groupings of drug reviews, aspect mining
(Zheng et al., 2014)	AEP-based Latent Dirichlet Allocation (AEP-LDA)	2014	Gibbs sampling	LDA	Opinion /aspect mining and sentiment word identification
(Bagheri et al., 2014)	ADM-LDA	2014	Gibbs sampling	LDA Markov chain	Aspect mining and sentiment analysis
(Xie et al., 2015)	MRF-LDA	2015	EM algorithm	-LDA -Markov Random Field	Exploiting word correlation knowledge
	Hawkes-LDA	2015	Variational Bayes inference	LDA	Analyzing text content and modeling scientific influence
(Yuan et al.,	LightLDA	2015	Gibbs	LDA	Topic modeling

2015)			sampling	Metropolis-Hastings sampling algorithm	for very large data sizes
(Nguyen et al., 2015)	Latent Feature LDA (LF-LDA), LF-DMM	2015	Gibbs sampling	LDA	Document clustering for short text
	TH Rank	2015	Gibbs sampling	-LDA -Author-Conference-Topic (ACT) model - PageRank	Topic sensitive ranking and find the relevant papers in journals
(Li et al., 2015a)	author-topic-community (ATC)	2015	Expectation-maximization (EM)	LDA	Author community discovery and Author interest profiling
(Lu and Lee, 2015)	TOT-MMM	2015	Gibbs sampling	LDA	Twitter Hashtag Recommendation
(Yu et al., 2015b)	link-field-topic (LFT),	2015	Gibbs sampling	LDA - semantic link weight (SLW)	Semantic community detection and dynamic topic discovery
(Jiang et al., 2015)	Scalable Geographic Web Search Topic Discovery (SG-WSTD)	2015	Gibbs sampling	LDA k-means algorithm	Geographic web search topic discovery and extracting geographic information
(Fu et al., 2015)	dynamic NJST (dNJST)	2015	Gibbs sampling algorithms	-LDA -hierarchical Dirichlet process (HDP)	Dynamic sentiment topic discovery in Chinese social media
(Li et al., 2015b)	Frequency-LDA (FLDA) and Dependency-Frequency-LDA (DFLDA)	2015	Gibbs sampling algorithms	-LDA	Multi-label document categorization

2.2.4. A brief look from some impressive past works: Research in 2016

According to Table 5, some of the popular works published for this year focused on a variety of topics, such as recommendation system(Cheng and Shen, 2016) (Zoghbi et al., 2016) (Zhao et al., 2016) , opinion mining and aspect mining (Bagheri et al., 2014) (Zheng et al., 2014) (Cheng et al., 2014) (Wang et al., 2014c).

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**(Xie et al., 2016), 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 the 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 the information of bursty topics but filter out the others. Equation (3) shows how we calculate the “**velocity**” $\hat{v}(t)$ and “**acceleration**” $\hat{\alpha}(t)$ of words.

$$\hat{v}_{\Delta T} = \sum_{t_i \leq t} X_i \cdot \frac{\exp((t_i - t)/\Delta T)}{\Delta T}, \quad \hat{\alpha}(t) = \frac{\hat{v}_{\Delta T_2}(t) - \hat{v}_{\Delta T_1}(t)}{\Delta T_1 - \Delta T_2}. \quad (3)$$

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.

Hashtag-LDA, the authors a personalized hashtag recommendation approach is introduced according to the latent topical information in untagged microblogs. This model can enhance the 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 (Zhao et al., 2016). **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 the 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 (Yeh et al., 2016).

Table1 5. Some impressive articles based on LDA for 2016

Author-Study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Zhao et al., 2016)	Hashtag-LDA	2016	Gibbs sampling	LDA	Hashtag recommendation

					and Find relationships between topics and hashtags
(Hong et al., 2016)	PMB-LDA	2016	Expectation-maximization (EM)	LDA	Extract the population mobility behaviors for large scale
(Lee et al., 2016)	Automatic Rule Generation (LARGen)	2016	Gibbs Sampling	LDA	Malware analysis and Automatic Signature Generation
(Liu et al., 2016)	PT-LDA	2016	Gibbs-EM algorithm	LDA	Personality recognition in social network
(Li et al., 2016a)	Corr-wddCRF	2016	Gibbs sampling	LDA	Knowledge Discovery in Electronic Medical Record
(Zoghbi et al., 2016)	multi-idiomatic LDA model (MiLDA)	2016	Gibbs sampling	LDA bilingual LDA (BiLDA)	Content-based recommendation and automatic linking
(Cheng and Shen, 2016)	Location-aware Topic Model (LTM)	2016	Gibbs sampling	LDA	Music Recommendation
(Miao et al., 2016)	TopPRF	2016	Gibbs sampling	LDA	Evaluate the relevancy between feedback documents
(Rao, 2016)	contextual sentiment topic model (CSTM)	2016	Expectation-maximization (EM)	LDA	Emotion classification in social network
(Yeh et al., 2016)	conceptual dynamic latent Dirichlet allocation (CDLDA)	2016	Gibbs sampling	LDA	Topic detection in conversations
(Lu et al., 2016)	multiple-channel latent Dirichlet allocation (MCLDA)	2016	Gibbs sampling	LDA	Find the relations between diagnoses and medications from healthcare data
(Qian et al., 2016)	multi-modal event topic model (mmETM)	2016	Gibbs sampling	LDA	Tracking and social event analysis
(Fu et al., 2016)	Dynamic Online Hierarchical Dirichlet Process model (DOHDP)	2016	Gibbs samplin	LDA	Dynamic topic evolutionary discovery for Chinese social media
(Xie et al., 2016)	Topicsketch	2016	Gibbs sampling	LDA - tensor decomposition algorithm - Count-Min algorithm	Realtime detection and bursty topics dicoverly from Twitter

(Zeng et al., 2016)	fast online EM (FOEM)	2016	Expectation-maximization (Batch EM)	LDA	Big topic modeling
(Alam et al., 2016)	Joint Multi-grain Topic Sentiment (JMST)	2016	Gibbs sampling	LDA	Extracting semantic aspects from online reviews
(Qin et al., 2016)	character–word topic model (CWTM)	2016	Gibbs sampling	LDA	Capture the semantic contents in text documents(Chinese language).

2.2 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(McCallum et al., 2005) (Wang et al., 2013) (Henderson and Eliassi-Rad, 2009, Yu et al., 2015a), Software Engineering(Linstead et al., 2008) (Chen et al., 2012) (Gethers and Poshyvanyk, 2010) (Linstead et al., 2007), Crime Science(Chen et al., 2015) (Gerber, 2014) (Wang et al., 2012) and also in areas of Geographical(Cristani et al., 2008) (Yin et al., 2011) (Sizov, 2010) (Tang et al., 2013), Political Science(Greene and Cross, 2015) (Cohen and Ruths, 2013) , Medical/Biomedical (Liu et al., 2010) (Huang et al., 2013) (Wang et al., 2011) (Zhang et al., 2017) (Xiao et al., 2017) and Linguistic science (Bauer et al., 2012) (McFarland et al., 2013) (Eidelman et al., 2012) (Wilson and Chew, 2010) (Vulić et al., 2011) as illustrated by Fig. 2.



Fig2. A clear vision of the application of Topic modeling in various sciences (Based on previous work).

A. 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. **Vulic et al.** employed the distributional hypothesis in various direction and it efforts to cancel the

requirement of a seed lexicon as an essential prerequisite for use of bilingual vocabulary and introduce various ways to identify the translation of words among languages (Vulić et al., 2011). **Eidelman et al.** 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 the 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 (Eidelman et al., 2012).

Table6. Impressive works LDA-based in Linguistic science

Study- Author	Year	Purpose	Dataset
(Vulić et al., 2011)	2011	Introduce various ways to identify the translation of words among languages [BiLDA].	A Wikipedia dataset (Arabic, Spanish, French, Russian and English)
(Wilson and Chew, 2010)	2010	Obtain term weighting based on LDA	A multilingual dataset
(McFarland et al., 2013)	2013	Present a diversity of new visualization techniques to make concept of topic-solutions	Dissertation abstracts_1980–2010 - 1 million abstracts
(Bauer et al., 2012)	2012	A topic modeling approach, that it consider geographic information	Foursquare Dataset
(Lui et al., 2014)	2014	An approach that is capable to find a document with different language	ALTW2010
(Heintz et al., 2013)	2013	A method for linguistic discovery and conceptual metaphors resources	Wikipedia

McFarland et al. presented a diversity of new visualization techniques to make the concept of topic-solutions and introduce new forms of supervised LDA, to evaluation they considered a corpus of dissertation abstracts from 1980–2010 that belongs to 240 universities in the United States(McFarland et al., 2013). **Bauer et al.** developed a standard topic modeling approach, that it 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(Bauer et al., 2012). **Heintz et al** 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(Heintz et al., 2013). **Lui et al.** 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 (Lui et al., 2014).

B. 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]. Also, some researchers 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 (Greene and Cross, 2015). **Cohen et al.** 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 fed for 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 (Cohen and Ruths, 2013).

Table7. Impressive works LDA-based in political science

Study- Author	Year	Purpose	Dataset
(Cohen and Ruths, 2013)	2013	Evaluate the behavioral effects of different databases from political orientation classifiers	-Political Figures Dataset -Politically Active Dataset (PAD) -Politically Modest Dataset (PMD) -Conover 2011 Dataset (C2D)
(Fang et al., 2012)		Introduce a topic model for contrastive opinion modeling	-Statement records of U.S. senators
(Balasubramany an et al., 2012)	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
(Chen et al., 2010)	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)
(Song et al., 2014)	2014	Analyze issues related to Korea's presidential election	Project Vote Smart WebSite (https://votesmart.org/)
(Levy and Franklin, 2014)	2014	Examine Political Contention in the U.S. Trucking Industry	Regulations.gov online portal
(Zirn and Stuckenschmidt , 2014)	2015	presented a method for multi-dimensional analysis of political documents	three German national elections (2002, 2005 and 2009)

Fang et al. They 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 records, 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 (Fang et al., 2012).

Yano et al. 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 evaluation, their model gathered comments and blog posts with focusing on American politics from 40 blog sites (Yano et al., 2009, Yano and Smith, 2010).

Madan et al. 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 (Madan et al., 2011). **Balasubramanyan et al.** they analyzed reactions of emotional 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 (Balasubramanyan et al., 2012). To evaluation, their approach they provided a dataset of blog posts from five blogs that focus on US politics that was made by (Yano et al., 2009).

Chen et al. 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 (Chen et al., 2010). **Song and et al.** It was 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 (Song et al., 2014). **Zirn et al.** 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 (Zirn and Stuckenschmidt, 2014).

C. Topic modeling in Medical/Biomedical

Topic models applied to text mining in Medical/biomedical domain, according to previous studies, LDA can be very effective and functional in this field. Topic modeling could be advantageously applied to the large datasets of biomedical/medical research. For example, a group of researchers, introduced three LDA-like models and found that this model cans higher accuracy than the state-of-the-art alternatives. Authors demonstrated that this approach based on LDA could successfully recognize the probabilistic patterns between Adverse drug reaction (ADR) topics and used ADRS database for evaluation their approach. The aim of the authors to predict ADR from a large number of ADR candidates to obtain a drug target(Xiao et al., 2017). **Zhang et al.** They focused on the issue of professionalized medical recommendations and proposed a new healthcare recommendation system that called iDoctor, that used Hybrid matrix factorization methods for the 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(Zhang et al., 2017).

Table8. Impressive works LDA-based in medical/biomedical

Study- Author	Year	Purpose/problem domain	Dataset

(Xiao et al., 2017)	2017	Presented three LDA-based models Adverse Drug Reaction Prediction	ADReCS database
(Wang et al., 2011)	2011	Extract biological terminology	-MEDLINE and Bio-Terms Extraction -Chem2Bio2Rdf.
(Zhang et al., 2017)	2017	User preference distribution discovery and identity distribution of doctor feature	-Yelp Dataset (Yelp.com)
(Wu et al., 2012b)	2012	-Ranking GENE-DRUG -Detecting relationship between gene and drug	National Library of Medicine
(Paul and Dredze, 2011)	2011	Analyzing public health information on Twitter	20 disease articles of twitter data
(Wang et al., 2013)	2013	Analysis of Generated Content by User from social networking sites	one million English posted from Facebook's server logs
(Huang et al., 2013)	2013	Pattern discovery and extraction for Clinical Processes	a data-set from Zhejiang Huzhou Central Hospital of China
(Liu et al., 2010)	2011	Identifying miRNA-mRNA in functional miRNA regulatory modules	mouse mammary dataset (Zhu et al., 2010)
(Zhang et al., 2011)	2011	Extract common relationship	T2DM Clinical Dataset
	2012	Extract the latent topic in	T2DM Clinical Dataset

(Jiang et al., 2012)		Traditional Chinese Medicine document	
----------------------	--	---------------------------------------	--

Wang et al. they 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(Wang et al., 2011). **Wu et al.** 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(Wu et al., 2012b). **Paul et al.** 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?(Paul and Dredze, 2011)”.

Huang et al. they 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(Huang et al., 2013). **Liu et al.** They 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 (Liu et al., 2010). **Zhang et al.** 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)(Zhang et al., 2011).

D. Topic modeling in Geographical/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. **Yin et al.** 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(Yin et al., 2011). **Sizov et al.** 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 it 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 (Sizov, 2010). **Tang et al.** they proposed a multiscale LDA model that is a combination of multiscale image representation and probabilistic topic model to obtain effective clustering VHR satellite images (Tang et al., 2013).

Table 9. Impressive works LDA-based in geographical/locations

Study-Author	Year	Purpose	Dataset
(Sizov, 2010)	2010	Discovering multi-faceted summaries of documents	CoPhIR dataset
(Yin et al., 2011)	2011	Content management and retrieval	Flicker Dataset
(Tang et al., 2013)	2013	Semantic clustering in very high resolution panchromatic satellite images	A QUICKBIRD image of a suburban area
(Eisenstein et al., 2010)	2010	Data Discovery, Evaluation of geographically coherent linguistic regions and find the relationship between topic variation and regional.	A Twitter Dataset
(Cristani et al., 2008)	2008	Geo-located image categorization and georecognition	3013 images Panoramio in France
(Zhang et al., 2015)	2015	Cluster discovery in geo-locations	Reuters-21578
(McInerney and Blei, 2014)	2014	Discovering newsworthy information From Twitter	A small Twitter Dataset

Eisenstein et al. They 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 (Eisenstein et al., 2010) (Tang et al., 2013) (Yin et al., 2011) (Sizov, 2010). **Cristani et al.** They 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 (Cristani et al., 2008).

Zhang et al. In this work, 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 (Zhang et al., 2015). **McInerney et al.** they 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 to identifying the significant tweet event, but for a large dataset of news articles, the result will be negative (McInerney and Blei, 2014).

E. 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. **Linstead et al.** For the first time, they 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 approach, can be effective for project organization, software refactoring (Linstead et al., 2007). **Gethers et al.** They 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 (Gethers and Poshyvanyk, 2010).

Asuncion et al. the authors 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 (Asuncion et al., 2010). **Thomas et al.** They 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 (Thomas, 2011). **Chen et al.** used a generative statistical model(LDA model) for analyzing source code evolution 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 (Chen et al., 2012). **Linstead et al.** 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(Linstead et al., 2008).

Tian et al. 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 (Tian et al., 2009). **Lukinet al.** 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 for evaluate and analyze of bugs in these source codes (Lukins et al., 2008, Lukins et al., 2010). **Yang et al.** They 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, to 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 (Yang et al., 2017).

Table10. Impressive works LDA-based in software engineering

Study- Author	Year	Purpose	Dataset
(Linstead et al., 2007)	2007	Mining software and extracted concepts from code	SourceForge and Apache(d 1,555 projects)
(Gethers and Poshyvanyk, 2010)	2010	Identifying latent topics and find their relationships in source code	Thirteen open source software systems
(Asuncion et al., 2010)	2010	Generating traceability links	ArchStudio software project
(Chen et al., 2012)	2012	Find relationship between the conceptual concerns in source code.	source code entities
(Linstead et al., 2008)	2008	Analyzing Software Evolution	, open source Java projects, Eclipse and ArgoUML

(Tian et al., 2009)	2009	Automatic Categorization of Software systems	43 open-source software systems
(Lukins et al., 2008)	2008	Source code retrieval for bug localization	Mozilla, Eclipse source code
(Lukins et al., 2010)	2010	Automatic bug localization and evaluate its effectiveness	Open source software such as (Rhino, and Eclipse)
(Yang et al., 2017)	2017	Detection of malicious Android apps	1612 malicious application

F. Topic modeling in Social Network / Microblogs (such as Twitter)

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. Recently, researchers have proposed many LDA approaches to analyzing user tweets on Twitter. **Weng et al.** In this paper, 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 (Weng et al., 2010). **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 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 (Hong et al., 2011).

Bhattacharya et al. In this paper, they 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 (Bhattacharya et al., 2014). **Kim et al.** They suggested, a recommendation system based on LDA for obtaining the behaviors of users on Twitter, called TWILITE. In more detail, TWILITE 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 (Kim and Shim, 2014). **Wang et al.** They 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 (Wang et al., 2012).

Table11. Impressive works LDA-based in social network

Study- Author	Year	Purpose	Dataset
Weng et all (Weng et al., 2010)	2010	Finding influential twitterers on social network(Twitter)	Top-1000 Singapore-based twitterers
Bhattacharya et all (Bhattacharya et al., 2014)	2014	Building a topical recommendation systems	A twitter dataset
Kim et all (Kim and Shim, 2014)	2014	A recommendation system for Twitter	A twitter dataset
Cordeiro et all(Cordeiro, 2012)	2012	Analysis and discovered events on Twitter	A twitter dataset
Tan et all (Tan et al., 2014)	2014	Analyze public sentiment variations regarding a certain tar on Twitter	A twitter dataset
Roberts et all (Roberts et al., 2012)	2012	Analysis of the emotional and stylistic distributions on Twitter	A twitter dataset
Ren et all (Ren et al., 2016)	2016	A topic-enhanced word embedding for Twitter sentiment classification	SemEval-2014
Li et al (Li et al., 2016b)	2016	Categorize emotion tendency on Sina Wibo	A Sina Wibo dataset

Godin et al. In this paper, they introduced a novel method based LDA model to hashtag recommendation on Twitter that can categories posts with them (hashtags)(Godin et al., 2013). **Lin et al.** They investigated the cold-start issue with using 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 (Lin et al., 2013).

Cordeiro et al. presented a technique to analysis and discovered events by an LDA model. Authors found that this method can detect events in inferred topics from tweets 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 (Cordeiro, 2012). **Pier et al.** In this paper, they investigated the issue of how to effectively discovery 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 (Prier et al., 2011). **Tan and et al.** 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 can rank 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 (Tan et al., 2014). **Roberts et al.** 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(Roberts et al., 2012). **Srijith et al.** This paper 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 (Srijith et al., 2017).

Ren et al. 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 (Ren et al., 2016). **Wang et al.** 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 the 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 (Wang et al., 2016).

G. 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. **Chen et al.** 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. And for the next step, used from achieved features of LDA to classify a new document by collaborative representation classifier (CRC). **Geber et al.** 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 (Chen et al., 2015, Gerber, 2014, Sharma et al., 2015).

Table12. Impressive works LDA-based in crime prediction

Study- Author	Year	Purpose	Dataset
(Wang et al., 2012)	2012	Automatic semantic analysis on Twitter posts	A corpus of tweets from Twitter(manual)
(Gerber, 2014)	2014	Crime prediction using tagged tweets	City of Chicago Data: https://data.cityofchicago.org
(Chen et al., 2015)	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

4.1 library/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 can consider as a standard source for testing and evaluation. **Table7**, Show some well-known tools for topic modeling and **Table8**, 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 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. (Steyvers and Griffiths, 2007)

Table13. Some well-known tools for topic modeling

Tools	Implementation/ Language	Inference/Parameter	source code availability
Mallet (McCallum, 2002)	Java	Gibbs sampling	http://mallet.cs.umass.edu/topics.php
TMT (Ramage and Rosen, 2011)	Java	Gibbs sampling	https://nlp.stanford.edu/software/tmt/tmt-0.4/
Mr.LDA (Zhai et al., 2012)	Java	Variational Bayesian inference	https://github.com/lintool/Mr.LDA
JGibbLDA (Phan and Nguyen, 2006)	Java	Gibbs sampling	http://jgibbllda.sourceforge.net/
Gensim (Řehůřek and Sojka, 2011)	Python	Gibbs sampling	https://radimrehurek.com/gensim
TopicXP (Řehůřek and Sojka, 2011)	Java(Eclipse plugin)		http://www.cs.wm.edu/semeru/TopicXP/
Matlab Topic Modeling (Steyvers and Griffiths, 2011)	Matlab	Gibbs sampling	http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
Yahoo_LDA (Chang, 2011)	C++	Gibbs sampling	https://github.com/shravanmn/Yahoo_LDA
Lda in R (Ahmed et al., 2012)	R	Gibbs sampling	https://cran.r-project.org/web/packages/lda/

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 8**, shows lists of some famous and popular datasets in various languages.

Table 14. Some well-known Dataset for topic modeling

Dataset	Language	Date of publishing	Short-detail	Availability address
Reuters (Reuters21578) (Lewis, 1997)	English	1997	Newsletters in various categories	http://kdd.ics.uci.edu/databases/reuters21578/reuters21578
ReutersV 1 (Reuters-Volume I) (Lewis et al., 2004)	English	2004	Newsletters in various categories	
UDI-TwitterCrawl-Aug2012 (Li et al., 2012)	English	2012	-a twitter dataset from millions of tweets	https://wiki.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012
SemEval-2013 Dataset (Manandhar and Yuret, 2013)	English	2013	-a twitter dataset from millions of tweets	
Wiki10[179]	English	2009	a Wikipedia Document in various category	http://nlp.uned.es/social-tagging/wiki10/
Weibo dataset (Zhang et al., 2013)	Chinese	2013	a popular Chinese microblogging network	
Bag of Words[180]	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 (Wang and Blei, 2011)	English	2011	a bibliography sharing service of	http://www.citeulike.org/faq/data.adp

			academic papers	
DBLP Dataset[183] (Lange and Naumann, 2011)	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(Asgari and Chappelier, 2013)	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 (Qin et al., 2016) (Cong et al., 2012)	Chinese		A Chinese corpus from Chinese Wikipedia	
Pascal VOC 2007 (Everingham et al., 2008) (Everingham et al., 2010)	English	2007	a dataset of natural images	http://host.robots.ox.ac.uk/pascal/VOC/voc2007/
AFP_ARB corpus(Larkey and Connell, 2001)	Arabic	2001	A collection of newspaper articles in Arabic from Agence France Presse	
20Newsgroups4 corpus[(Rennie, 2008)	English	Jan 2008	Newsletters in various categories	http://qwone.com/~jason/20Newsgroups/
New York Times (NYT)dataset (Sandhaus, 2008)	English	Oct 2008	Newsletters in various categories	

5. Seven important issues in Challenges and Open research

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 will discuss 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 classification. For example, an image of a class highway is more likely annotated with words “road” and “traffic”, “car ” than words “fish ” “scuba” and “boat”. **Chong et al.** develop a new probabilistic model for jointly modeling the image, its annotations, and its category 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(Chong et al., 2009).

Lienou and et al. 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. **Wick and et al.** 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. **Vaduva and et al.** 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. **Philbin and et al.** proposed a geometrically consistent latent topic model to detect significant objects, called Latent Dirichlet Allocation (gLDA). and then introduced methods for the 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. (Lienou et al., 2010, Philbin et al., 2011, Vaduva et al., 2013, Wick et al., 2007).

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; **Nakano et al.** The authors 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. **Hu et al.** They 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) (Hu et al., 2014, Nakano et al., 2014).

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. **Yu and et al.** 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. **Bisgin and et al.** 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 the 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 (Bisgin et al., 2014, Yang and Kiang, 2015, Yu et al., 2014).

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 the 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. **Alashri and et al.** employed sentiment analysis, topic modeling, and trends detection through wavelet transform to topics and trends discovery. They extracted 9,700 posts and 12,050,595 comments of five USA presidential candidates (Ted Cruz , Donald Trump, Hillary Clinton, John Kasich and Bernie Sanders) from their official Facebook pages (Alashri et al., 2016, Wang et al., 2016). **Shi and et al.** The authors proposed a novel probabilistic graphical model for the pattern discovery in comments that called MCTA. This model can cope with the language gap and obtain the common semantics with considering various languages from News Reader Comments (such as Chinese and English newreader comments) (Shi et al., 2016). **Hou and et al.** presented a context and co-mention method using knowledge linking method and a topic-level alignment method to build the links between external resources and news from social media. It can also be said that they applied a unified probabilistic model for predict news and relationship discovery within events and topics with considering the background knowledge of users' comments(Hou et al., 2015) .

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. **Henderson and et al.** 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. **Yu and et al.** 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 (Henderson and Eliassi-Rad, 2009, Yu et al., 2015a).

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 (Diao et al., 2012) (Yin et al., 2014), 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.** 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. **Wang et al.** presented a solution framework based on user behavior and synergetic modeling of multi-modal content using a topic analytic for cross media topic analysis and detection of the behavior of users activities on the web. **Yuan et al.** proposed a framework based on a probabilistic topic modeling method to detection of “user interests” and user behavior pattern discovery in the mobile Web usage log. They applied this model on a real dataset in Beijing that include 3 million users (Giri et al., 2014, Wang et al., 2014b, Yuan et al., 2014). **Siersdorfer and et al.** The authors focused on analysis comment rating behavior of users on social medias and gathered more than 10 million user comments from YouTube and Yahoo! News websites. For YouTube, they restricted their analysis on tag annotations for content and employed Latent Dirichlet Allocation (LDA) to obtain term probabilities and each tag of a video defined as a mixture of latent topics. Also, they used a linear support vector machines (SVMs) to detection of comments likely to attract replies (Siersdorfer et al., 2014).

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 (Chaney and Blei, 2012, Kim et al., 2017, Murdock and Allen, 2015, Gretarsson et al., 2012). **Chuang and et al.** The authors 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 (Chuang et al., 2012). **Millar and et al.** the authors proposed a combined approach based on Latent Dirichlet Allocation for dimensionality reduction and self-organizing maps to document Clustering and Visualization, that called LDA-SOM [160]. **Sievert and et al.** introduced LDAvis as an interactive visualization system using LDA that capable the providing a global view of the topics, and has a flexible feature for exploring relationship between topics and terms and obtain better understand from a fitted LDA model. It can also be said that the system can find significant topics and cluster them in various categories (Millar et al., 2009, Siersdorfer et al., 2014).

6. Conclusion

Topic modeling provides methods for latent knowledge discovery, finding relationships among data, understanding, and summarizing huge electronic archives. 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.

Reference

- AHMED, A., ALY, M., GONZALEZ, J., NARAYANAMURTHY, S. & SMOLA, A. J. Scalable inference in latent variable models. Proceedings of the fifth ACM international conference on Web search and data mining, 2012. ACM, 123-132.
- ALAM, M. H., RYU, W.-J. & LEE, S. 2016. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. *Information Sciences*, 339, 206-223.
- ALASHRI, S., KANDALA, S. S., BAJAJ, V., RAVI, R., SMITH, K. L. & DESOUZA, K. C. An analysis of sentiments on facebook during the 2016 US presidential election. Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, 2016. IEEE, 795-802.
- ALSUMAIT, L., BARBARÁ, D. & DOMENICONI, C. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, 2008. IEEE, 3-12.
- ASGARI, E. & CHAPPELIER, J.-C. Linguistic Resources and Topic Models for the Analysis of Persian Poems. CLfL@ NAACL-HLT, 2013. 23-31.
- ASUNCION, H. U., ASUNCION, A. U. & TAYLOR, R. N. Software traceability with topic modeling. Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1, 2010. ACM, 95-104.
- BAGHERI, A., SARAEE, M. & DE JONG, F. 2014. ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, 40, 621-636.
- BALASUBRAMANYAN, R., COHEN, W. W., PIERCE, D. & REDLAWSK, D. P. Modeling polarizing topics: When do different political communities respond differently to the same news? ICWSM, 2012.
- BAUER, S., NOULAS, A., SÉAGHDHA, D. O., CLARK, S. & MASCOLO, C. Talking places: Modelling and analysing linguistic content in foursquare. Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), 2012. IEEE, 348-357.
- BHATTACHARYA, P., ZAFAR, M. B., GANGULY, N., GHOSH, S. & GUMMADI, K. P. Inferring user interests in the twitter social network. Proceedings of the 8th ACM Conference on Recommender systems, 2014. ACM, 357-360.
- BISGIN, H., LIU, Z., FANG, H., KELLY, R., XU, X. & TONG, W. 2014. A phenome-guided drug repositioning through a latent variable model. *BMC bioinformatics*, 15, 267.

- BLEI, D. M. & JORDAN, M. I. Modeling annotated data. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 2003. ACM, 127-134.
- BLEI, D. M. & LAFFERTY, J. D. Dynamic topic models. Proceedings of the 23rd international conference on Machine learning, 2006. ACM, 113-120.
- BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3, 993-1022.
- CHANEY, A. J.-B. & BLEI, D. M. Visualizing Topic Models. ICWSM, 2012.
- CHANG, J. 2011. lda: Collapsed Gibbs sampling methods for topic models. R.
- CHANG, J. & BLEI, D. M. Relational topic models for document networks. International conference on artificial intelligence and statistics, 2009. 81-88.
- CHEN, B., ZHU, L., KIFER, D. & LEE, D. What Is an Opinion About? Exploring Political Standpoints Using Opinion Scoring Model. AAAI, 2010.
- CHEN, L., WANG, Y., YU, Q., ZHENG, Z. & WU, J. WT-LDA: user tagging augmented LDA for web service clustering. International Conference on Service-Oriented Computing, 2013. Springer, 162-176.
- CHEN, S.-H., SANTOSO, A., LEE, Y.-S. & WANG, J.-C. Latent dirichlet allocation based blog analysis for criminal intention detection system. Security Technology (ICCST), 2015 International Carnahan Conference on, 2015. IEEE, 73-76.
- CHEN, T.-H., THOMAS, S. W., NAGAPPAN, M. & HASSAN, A. E. Explaining software defects using topic models. Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on, 2012. IEEE, 189-198.
- CHENG, V. C., LEUNG, C. H., LIU, J. & MILANI, A. 2014. Probabilistic aspect mining model for drug reviews. *IEEE transactions on knowledge and data engineering*, 26, 2002-2013.
- CHENG, Z. & SHEN, J. 2016. On effective location-aware music recommendation. *ACM Transactions on Information Systems (TOIS)*, 34, 13.
- CHIEN, J.-T. & CHUEH, C.-H. 2011. Dirichlet class language models for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19, 482-495.
- CHONG, W., BLEI, D. & LI, F.-F. Simultaneous image classification and annotation. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009. IEEE, 1903-1910.
- CHOO, J., LEE, C., REDDY, C. K. & PARK, H. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19, 1992-2001.
- CHUANG, J., MANNING, C. D. & HEER, J. Termite: Visualization techniques for assessing textual topic models. Proceedings of the international working conference on advanced visual interfaces, 2012. ACM, 74-77.
- COHEN, R., AVIRAM, I., ELHADAD, M. & ELHADAD, N. 2014. Redundancy-aware topic modeling for patient record notes. *PloS one*, 9, e87555.
- COHEN, R. & RUTHS, D. Classifying political orientation on Twitter: It's not easy! ICWSM, 2013.
- CONG, Y., QIN, Z., YU, J. & WAN, T. Cross-Modal Information Retrieval—A Case Study on Chinese Wikipedia. International Conference on Advanced Data Mining and Applications, 2012. Springer, 15-26.
- CORDEIRO, M. Twitter event detection: combining wavelet analysis and topic inference summarization. Doctoral symposium on informatics engineering, 2012. 11-16.

- CRISTANI, M., PERINA, A., CASTELLANI, U. & MURINO, V. Geo-located image analysis using latent representations. *Computer Vision and Pattern Recognition*, 2008. CVPR 2008. IEEE Conference on, 2008. IEEE, 1-8.
- DIAO, Q., JIANG, J., ZHU, F. & LIM, E.-P. Finding bursty topics from microblogs. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 2012. Association for Computational Linguistics, 536-544.
- EIDELMAN, V., BOYD-GRABER, J. & RESNIK, P. Topic models for dynamic translation model adaptation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 2012. Association for Computational Linguistics, 115-119.
- EISENSTEIN, J., O'CONNOR, B., SMITH, N. A. & XING, E. P. A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010. Association for Computational Linguistics, 1277-1287.
- EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K., WINN, J. & ZISSERMAN, A. 2008. The pascal visual object classes challenge 2007 (voc 2007) results (2007).
- EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K., WINN, J. & ZISSERMAN, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 303-338.
- FANG, Y., SI, L., SOMASUNDARAM, N. & YU, Z. Mining contrastive opinions on political texts using cross-perspective topic model. *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012. ACM, 63-72.
- FU, X., LI, J., YANG, K., CUI, L. & YANG, L. 2016. Dynamic online HDP model for discovering evolutionary topics from Chinese social texts. *Neurocomputing*, 171, 412-424.
- FU, X., YANG, K., HUANG, J. Z. & CUI, L. 2015. Dynamic non-parametric joint sentiment topic mixture model. *Knowledge-Based Systems*, 82, 102-114.
- GERBER, M. S. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115-125.
- GETHERS, M. & POSHYVANYK, D. Using relational topic models to capture coupling among classes in object-oriented software systems. *Software Maintenance (ICSM)*, 2010 IEEE International Conference on, 2010. IEEE, 1-10.
- GIRI, R., CHOI, H., HOO, K. S. & RAO, B. D. User behavior modeling in a cellular network using latent dirichlet allocation. *International Conference on Intelligent Data Engineering and Automated Learning*, 2014. Springer, 36-44.
- GODIN, F., SLAVKOVIKJ, V., DE NEVE, W., SCHRAUWEN, B. & VAN DE WALLE, R. Using topic models for twitter hashtag recommendation. *Proceedings of the 22nd International Conference on World Wide Web*, 2013. ACM, 593-596.
- GREENE, D. & CROSS, J. P. Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis. *Proceedings of the ACM Web Science Conference*, 2015. ACM, 2.
- GRETARSSON, B., O'DONOVAN, J., BOSTANDJIEV, S., HÖLLERER, T., ASUNCION, A., NEWMAN, D. & SMYTH, P. 2012. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3, 23.
- GRIFFITHS, T. L. & STEYVERS, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101, 5228-5235.

- GUO, J., XU, G., CHENG, X. & LI, H. Named entity recognition in query. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009. ACM, 267-274.
- HEINTZ, I., GABBARD, R., SRINIVASAN, M., BARNER, D., BLACK, D. S., FREEDMAN, M. & WEISCHEDEL, R. Automatic extraction of linguistic metaphor with lda topic modeling. Proceedings of the First Workshop on Metaphor in NLP, 2013. 58-66.
- HENDERSON, K. & ELIASSI-RAD, T. Applying latent dirichlet allocation to group discovery in large graphs. Proceedings of the 2009 ACM symposium on Applied Computing, 2009. ACM, 1456-1461.
- HONG, L., DAN, O. & DAVISON, B. D. Predicting popular messages in twitter. Proceedings of the 20th international conference companion on World wide web, 2011. ACM, 57-58.
- HONG, L., FRIAS-MARTINEZ, E. & FRIAS-MARTINEZ, V. Topic Models to Infer Socio-Economic Maps. AAI, 2016. 3835-3841.
- HOU, L., LI, J., WANG, Z., TANG, J., ZHANG, P., YANG, R. & ZHENG, Q. 2015. Newsminer: Multifaceted news analysis for event search. *Knowledge-Based Systems*, 76, 17-29.
- HU, P., LIU, W., JIANG, W. & YANG, Z. 2014. Latent topic model for audio retrieval. *Pattern Recognition*, 47, 1138-1143.
- HU, Y., JOHN, A., WANG, F. & KAMBHAMPATI, S. ET-LDA: Joint Topic Modeling for Aligning Events and their Twitter Feedback. AAI, 2012. 59-65.
- HUANG, Z., LU, X. & DUAN, H. 2013. Latent treatment pattern discovery for clinical processes. *Journal of medical systems*, 37, 9915.
- JAGARLAMUDI, J. & DAUMÉ III, H. Extracting Multilingual Topics from Unaligned Comparable Corpora. ECIR, 2010. Springer, 444-456.
- JIANG, D., VOSECKY, J., LEUNG, K. W.-T., YANG, L. & NG, W. 2015. SG-WSTD: A framework for scalable geographic web search topic discovery. *Knowledge-Based Systems*, 84, 18-33.
- JIANG, Z., ZHOU, X., ZHANG, X. & CHEN, S. Using link topic model to analyze traditional chinese medicine clinical symptom-herb regularities. e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on, 2012. IEEE, 15-18.
- JO, Y. & OH, A. H. Aspect and sentiment unification model for online review analysis. Proceedings of the fourth ACM international conference on Web search and data mining, 2011. ACM, 815-824.
- KIM, M., KANG, K., PARK, D., CHOO, J. & ELMQVIST, N. 2017. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE transactions on visualization and computer graphics*, 23, 151-160.
- KIM, Y. & SHIM, K. 2014. TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems*, 42, 59-77.
- LACOSTE-JULIEN, S., SHA, F. & JORDAN, M. I. DiscLDA: Discriminative learning for dimensionality reduction and classification. Advances in neural information processing systems, 2009. 897-904.
- LANGE, D. & NAUMANN, F. Frequency-aware similarity measures: why Arnold Schwarzenegger is always a duplicate. Proceedings of the 20th ACM international conference on Information and knowledge management, 2011. ACM, 243-248.
- LARKEY, L. S. & CONNELL, M. E. Arabic Information Retrieval at UMass in TREC-10. TREC, 2001.

- LEE, S., KIM, S., LEE, S., YOON, H., LEE, D., CHOI, J. & LEE, J.-R. 2016. LARGen: automatic signature generation for Malwares using latent Dirichlet allocation. *IEEE Transactions on Dependable and Secure Computing*.
- LEVY, K. E. & FRANKLIN, M. 2014. Driving regulation: using topic models to examine political contention in the US trucking industry. *Social Science Computer Review*, 32, 182-194.
- LEWIS, D. D. 1997. Reuters-21578 text categorization collection.
- LEWIS, D. D., YANG, Y., ROSE, T. G. & LI, F. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5, 361-397.
- LI, C., CHEUNG, W. K., YE, Y., ZHANG, X., CHU, D. & LI, X. 2015a. The author-topic-community model for author interest profiling and community discovery. *Knowledge and Information Systems*, 44, 359-383.
- LI, C., RANA, S., PHUNG, D. & VENKATESH, S. 2016a. Hierarchical Bayesian nonparametric models for knowledge discovery from electronic medical records. *Knowledge-Based Systems*, 99, 168-182.
- LI, F., HUANG, M. & ZHU, X. Sentiment Analysis with Global Topics and Local Dependency. AAI, 2010. 1371-1376.
- LI, J., CARDIE, C. & LI, S. TopicSpam: a Topic-Model based approach for spam detection. *ACL (2)*, 2013. 217-221.
- LI, R., WANG, S., DENG, H., WANG, R. & CHANG, K. C.-C. Towards social user profiling: unified and discriminative influence model for inferring home locations. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012. ACM, 1023-1031.
- LI, W. & MCCALLUM, A. Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd international conference on Machine learning*, 2006. ACM, 577-584.
- LI, X., OUYANG, J. & ZHOU, X. 2015b. Supervised topic models for multi-label classification. *Neurocomputing*, 149, 811-819.
- LI, Y., ZHOU, X., SUN, Y. & ZHANG, H. 2016b. Design and implementation of Weibo sentiment analysis based on LDA and dependency parsing. *China Communications*, 13, 91-105.
- LIENOU, M., MAITRE, H. & DATCU, M. 2010. Semantic annotation of satellite images using latent Dirichlet allocation. *IEEE Geoscience and Remote Sensing Letters*, 7, 28-32.
- LIN, C. X., ZHAO, B., MEI, Q. & HAN, J. PET: a statistical model for popular events tracking in social communities. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010. ACM, 929-938.
- LIN, J., SUGIYAMA, K., KAN, M.-Y. & CHUA, T.-S. Addressing cold-start in app recommendation: latent user models constructed from twitter followers. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013. ACM, 283-292.
- LINSTEAD, E., LOPES, C. & BALDI, P. An application of latent Dirichlet allocation to analyzing software evolution. *Machine Learning and Applications*, 2008. ICMLA'08. Seventh International Conference on, 2008. IEEE, 813-818.
- LINSTEAD, E., RIGOR, P., BAJRACHARYA, S., LOPES, C. & BALDI, P. Mining concepts from code with probabilistic topic models. *Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering*, 2007. ACM, 461-464.

- LIU, B., LIU, L., TSYKIN, A., GOODALL, G. J., GREEN, J. E., ZHU, M., KIM, C. H. & LI, J. 2010. Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, 26, 3105-3111.
- 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.
- LIU, Z., ZHANG, Y., CHANG, E. Y. & SUN, M. 2011. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, 26.
- 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 Intelligent Systems*, 1-1.
- LU, H.-M., WEI, C.-P. & HSIAO, F.-Y. 2016. Modeling healthcare data using multiple-channel latent Dirichlet allocation. *Journal of biomedical informatics*, 60, 210-223.
- LUI, M., LAU, J. H. & BALDWIN, T. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2, 27-40.
- LUKINS, S. K., KRAFT, N. A. & ETZKORN, L. H. Source code retrieval for bug localization using latent dirichlet allocation. Reverse Engineering, 2008. WCRE'08. 15th Working Conference on, 2008. IEEE, 155-164.
- LUKINS, S. K., KRAFT, N. A. & ETZKORN, L. H. 2010. Bug localization using latent dirichlet allocation. *Information and Software Technology*, 52, 972-990.
- MADAN, A., FARRAHI, K., GATICA-PEREZ, D. & PENTLAND, A. S. Pervasive sensing to model political opinions in face-to-face networks. International Conference on Pervasive Computing, 2011. Springer, 214-231.
- MANANDHAR, S. & YURET, D. Second joint conference on lexical and computational semantics (* sem), volume 2: Proceedings of the seventh international workshop on semantic evaluation (semeval 2013). Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013.
- MAO, X.-L., MING, Z.-Y., CHUA, T.-S., LI, S., YAN, H. & LI, X. SSDLDA: a semi-supervised hierarchical topic model. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, 2012. Association for Computational Linguistics, 800-809.
- MCCALLUM, A., CORRADA-EMMANUEL, A. & WANG, X. 2005. Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series*, 3.
- MCCALLUM, A. K. 2002. Mallet: A machine learning for language toolkit.
- MCFARLAND, D. A., RAMAGE, D., CHUANG, J., HEER, J., MANNING, C. D. & JURAFSKY, D. 2013. Differentiating language usage through topic models. *Poetics*, 41, 607-625.
- MCINERNEY, J. & BLEI, D. M. Discovering newsworthy tweets with a geographical topic model. NewsKDD: Data Science for News Publishing workshop Workshop in conjunction with KDD2014 the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2014.
- MIAO, J., HUANG, J. X. & ZHAO, J. 2016. TopPRF: A probabilistic framework for integrating topic space into pseudo relevance feedback. *ACM Transactions on Information Systems (TOIS)*, 34, 22.
- MILLAR, J. R., PETERSON, G. L. & MENDENHALL, M. J. Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps. FLAIRS Conference, 2009. 69-74.

- MINKA, T. & LAFFERTY, J. Expectation-propagation for the generative aspect model. Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, 2002. Morgan Kaufmann Publishers Inc., 352-359.
- MURDOCK, J. & ALLEN, C. Visualization Techniques for Topic Model Checking. AAI, 2015. 4284-4285.
- NAKANO, T., YOSHII, K. & GOTO, M. Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014. IEEE, 5202-5206.
- NGUYEN, D. Q., BILLINGSLEY, R., DU, L. & JOHNSON, M. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3, 299-313.
- PANICHELLA, A., DIT, B., OLIVETO, R., DI PENTA, M., POSHYVANYK, D. & DE LUCIA, A. How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms. Proceedings of the 2013 International Conference on Software Engineering, 2013. IEEE Press, 522-531.
- PAUL, M. & DREDZE, M. Factorial LDA: Sparse multi-dimensional text models. Advances in Neural Information Processing Systems, 2012. 2582-2590.
- PAUL, M. & GIRJU, R. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. *Urbana*, 51, 36.
- PAUL, M. J. & DREDZE, M. 2011. You are what you Tweet: Analyzing Twitter for public health. *Icwsn*, 20, 265-272.
- PHAN, X.-H. & NGUYEN, C.-T. 2006. Jgibbllda: A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference.
- PHILBIN, J., SIVIC, J. & ZISSERMAN, A. 2011. Geometric latent dirichlet allocation on a matching graph for large-scale image datasets. *International journal of computer vision*, 95, 138-153.
- PRIER, K. W., SMITH, M. S., GIRAUD-CARRIER, C. & HANSON, C. L. Identifying health-related topics on twitter. International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, 2011. Springer, 18-25.
- QIAN, S., ZHANG, T., XU, C. & SHAO, J. 2016. Multi-modal event topic model for social event analysis. *IEEE Transactions on Multimedia*, 18, 233-246.
- QIN, Z., CONG, Y. & WAN, T. 2016. Topic modeling of Chinese language beyond a bag-of-words. *Computer Speech & Language*, 40, 60-78.
- RAMAGE, D., HALL, D., NALLAPATI, R. & MANNING, C. D. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, 2009. Association for Computational Linguistics, 248-256.
- RAMAGE, D. & ROSEN, E. 2011. Stanford topic modeling toolbox. Dec.
- RAO, Y. 2016. Contextual sentiment topic model for adaptive social emotion classification. *IEEE Intelligent Systems*, 31, 41-47.
- RAO, Y., LEI, J., WENYIN, L., LI, Q. & CHEN, M. 2014. Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, 17, 723-742.
- ŘEHŮŘEK, R. & SOJKA, P. 2011. Gensim—Statistical Semantics in Python.
- REN, Y., WANG, R. & JI, D. 2016. A topic-enhanced word embedding for Twitter sentiment classification. *Information Sciences*, 369, 188-198.
- RENNIE, J. 2008. The 20 Newsgroups data set. [http](http://www.cse.cmu.edu/~jrennie/).
- ROBERTS, K., ROACH, M. A., JOHNSON, J., GUTHRIE, J. & HARABAGIU, S. M. EmpaTweet: Annotating and Detecting Emotions on Twitter. LREC, 2012. 3806-3813.

- ROSEN-ZVI, M., GRIFFITHS, T., STEYVERS, M. & SMYTH, P. The author-topic model for authors and documents. Proceedings of the 20th conference on Uncertainty in artificial intelligence, 2004. AUAI Press, 487-494.
- SANDHAUS, E. 2008. The New York Times Annotated Corpus. Philadelphia, PA: Linguistic Data Consortium.
- SAVAGE, T., DIT, B., GETHERS, M. & POSHYVANYK, D. Topic XP: Exploring topics in source code using Latent Dirichlet Allocation. Software Maintenance (ICSM), 2010 IEEE International Conference on, 2010. IEEE, 1-6.
- SHARMA, V., KULSHRESHTHA, R., SINGH, P., AGRAWAL, N. & KUMAR, A. Analyzing Newspaper Crime Reports for Identification of Safe Transit Paths. HLT-NAACL, 2015. 17-24.
- SHI, B., LAM, W., BING, L. & XU, Y. Detecting Common Discussion Topics Across Culture From News Reader Comments. ACL (1), 2016.
- SIERSDORFER, S., CHELARU, S., PEDRO, J. S., ALTINGOVDE, I. S. & NEJDL, W. 2014. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web (TWEB)*, 8, 17.
- SIZOV, S. Geofolk: latent spatial semantics in web 2.0 social media. Proceedings of the third ACM international conference on Web search and data mining, 2010. ACM, 281-290.
- SONG, M., KIM, M. C. & JEONG, Y. K. 2014. Analyzing the political landscape of 2012 korean presidential election in twitter. *IEEE Intelligent Systems*, 29, 18-26.
- SRIJITH, P., HEPPLER, M., BONTCHEVA, K. & PREOTIUC-PIETRO, D. 2017. Sub-story detection in Twitter with hierarchical Dirichlet processes. *Information Processing & Management*, 53, 989-1003.
- STEYVERS, M. & GRIFFITHS, T. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427, 424-440.
- STEYVERS, M. & GRIFFITHS, T. 2011. Matlab topic modeling toolbox 1.4. URL http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.
- TAN, S., LI, Y., SUN, H., GUAN, Z., YAN, X., BU, J., CHEN, C. & HE, X. 2014. Interpreting the public sentiment variations on twitter. *IEEE transactions on knowledge and data engineering*, 26, 1158-1170.
- TANG, H., SHEN, L., QI, Y., CHEN, Y., SHU, Y., LI, J. & CLAUSI, D. A. 2013. A multiscale latent Dirichlet allocation model for object-oriented clustering of VHR panchromatic satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 51, 1680-1692.
- THOMAS, S. W. Mining software repositories using topic models. Proceedings of the 33rd International Conference on Software Engineering, 2011. ACM, 1138-1139.
- THOMAS, S. W., ADAMS, B., HASSAN, A. E. & BLOSTEIN, D. Modeling the evolution of topics in source code histories. Proceedings of the 8th working conference on mining software repositories, 2011. ACM, 173-182.
- TIAN, K., REVELLE, M. & POSHYVANYK, D. Using latent dirichlet allocation for automatic categorization of software. Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on, 2009. IEEE, 163-166.
- TITOV, I. & MCDONALD, R. Modeling online reviews with multi-grain topic models. Proceedings of the 17th international conference on World Wide Web, 2008. ACM, 111-120.
- VADUVA, C., GAVAT, I. & DATCU, M. 2013. Latent Dirichlet allocation for spatial analysis of satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 51, 2770-2786.

- VULIĆ, I., DE SMET, W. & MOENS, M.-F. Identifying word translations from comparable corpora using latent topic models. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 2011. Association for Computational Linguistics, 479-484.
- WALLACH, H. M., MIMNO, D. M. & MCCALLUM, A. Rethinking LDA: Why priors matter. Advances in neural information processing systems, 2009. 1973-1981.
- WANG, C. & BLEI, D. M. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. Advances in neural information processing systems, 2009. 1982-1989.
- WANG, C. & BLEI, D. M. Collaborative topic modeling for recommending scientific articles. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011. ACM, 448-456.
- WANG, H., DING, Y., TANG, J., DONG, X., HE, B., QIU, J. & WILD, D. J. 2011. Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PloS one*, 6, e17243.
- WANG, J., ZHOU, J., XU, H., MEI, T., HUA, X.-S. & LI, S. 2014a. Image tag refinement by regularized latent Dirichlet allocation. *Computer Vision and Image Understanding*, 124, 61-70.
- WANG, S., WANG, Z., JIANG, S. & HUANG, Q. Cross media topic analytics based on synergetic content and user behavior modeling. Multimedia and Expo (ICME), 2014 IEEE International Conference on, 2014b. IEEE, 1-6.
- WANG, T., CAI, Y., LEUNG, H.-F., LAU, R. Y., LI, Q. & MIN, H. 2014c. Product aspect extraction supervised with online domain knowledge. *Knowledge-Based Systems*, 71, 86-100.
- WANG, X., GERBER, M. S. & BROWN, D. E. 2012. Automatic Crime Prediction Using Events Extracted from Twitter Posts. *SBP*, 12, 231-238.
- WANG, X. & MCCALLUM, A. Topics over time: a non-Markov continuous-time model of topical trends. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006. ACM, 424-433.
- WANG, Y.-C., BURKE, M. & KRAUT, R. E. Gender, topic, and audience response: an analysis of user-generated content on facebook. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013. ACM, 31-34.
- WANG, Y., LUO, J., NIEMI, R., LI, Y. & HU, T. Catching Fire via "Likes": Inferring Topic Preferences of Trump Followers on Twitter. ICWSM, 2016. 719-722.
- WANG, Y. & MORI, G. Max-margin Latent Dirichlet Allocation for Image Classification and Annotation. BMVC, 2011. 7.
- WENG, J. & LEE, B.-S. 2011. Event detection in twitter. *ICWSM*, 11, 401-408.
- WENG, J., LIM, E.-P., JIANG, J. & HE, Q. TwitterRank: finding topic-sensitive influential twitterers. Proceedings of the third ACM international conference on Web search and data mining, 2010. ACM, 261-270.
- WICK, M., ROSS, M. & LEARNED-MILLER, E. Context-sensitive error correction: Using topic models to improve OCR. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, 2007. IEEE, 1168-1172.
- WILSON, A. T. & CHEW, P. A. Term weighting schemes for latent dirichlet allocation. human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics, 2010. Association for Computational Linguistics, 465-473.
- WU, H., BU, J., CHEN, C., ZHU, J., ZHANG, L., LIU, H., WANG, C. & CAI, D. 2012a. Locally discriminative topic modeling. *Pattern Recognition*, 45, 617-625.

- WU, Y., LIU, M., ZHENG, W. J., ZHAO, Z. & XU, H. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2012b. NIH Public Access, 422.
- XIANGHUA, F., GUO, L., YANYAN, G. & ZHIQIANG, W. 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems*, 37, 186-195.
- XIAO, C., ZHANG, P., CHAOVALITWONGSE, W. A., HU, J. & WANG, F. Adverse Drug Reaction Prediction with Symbolic Latent Dirichlet Allocation. AAAI, 2017. 1590-1596.
- XIE, P., YANG, D. & XING, E. P. Incorporating Word Correlation Knowledge into Topic Modeling. HLT-NAACL, 2015. 725-734.
- XIE, W., ZHU, F., JIANG, J., LIM, E.-P. & WANG, K. 2016. Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28, 2216-2229.
- YANG, M.-C. & RIM, H.-C. 2014. Identifying interesting Twitter contents using topical analysis. *Expert Systems with Applications*, 41, 4330-4336.
- YANG, M. & KIANG, M. Extracting Consumer Health Expressions of Drug Safety from Web Forum. System Sciences (HICSS), 2015 48th Hawaii International Conference on, 2015. IEEE, 2896-2905.
- YANG, X., LO, D., LI, L., XIA, X., BISSYANDÉ, T. F. & KLEIN, J. 2017. Characterizing malicious Android apps by mining topic-specific data flow signatures. *Information and Software Technology*.
- YANO, T., COHEN, W. W. & SMITH, N. A. Predicting response to political blog posts with topic models. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009. Association for Computational Linguistics, 477-485.
- YANO, T. & SMITH, N. A. What's Worthy of Comment? Content and Comment Volume in Political Blogs. ICWSM, 2010.
- 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.
- YIN, H., CUI, B., CHEN, L., HU, Z. & HUANG, Z. A temporal context-aware model for user behavior modeling in social media systems. Proceedings of the 2014 ACM SIGMOD international conference on Management of data, 2014. ACM, 1543-1554.
- YIN, Z., CAO, L., HAN, J., ZHAI, C. & HUANG, T. Geographical topic discovery and comparison. Proceedings of the 20th international conference on World wide web, 2011. ACM, 247-256.
- 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, 717-730.
- YU, K., ZHANG, J., CHEN, M., XU, X., SUZUKI, A., ILIC, K. & TONG, W. 2014. Mining hidden knowledge for drug safety assessment: topic modeling of LiverTox as a case study. *BMC bioinformatics*, 15, S6.
- YU, R., HE, X. & LIU, Y. 2015a. Glad: group anomaly detection in social media analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10, 18.
- YU, X., YANG, J. & XIE, Z.-Q. 2015b. A semantic overlapping community detection algorithm based on field sampling. *Expert Systems with Applications*, 42, 366-375.
- YUAN, B., XU, B., WU, C. & MA, Y. Mobile Web User Behavior Modeling. International Conference on Web Information Systems Engineering, 2014. Springer, 388-397.

- YUAN, J., GAO, F., HO, Q., DAI, W., WEI, J., ZHENG, X., XING, E. P., LIU, T.-Y. & MA, W.-Y. Lightlda: Big topic models on modest computer clusters. Proceedings of the 24th International Conference on World Wide Web, 2015. International World Wide Web Conferences Steering Committee, 1351-1361.
- ZENG, J., LIU, Z.-Q. & CAO, X.-Q. 2016. Fast online EM for big topic modeling. *IEEE Transactions on Knowledge and Data Engineering*, 28, 675-688.
- ZHAI, K., BOYD-GRABER, J., ASADI, N. & ALKHOUSA, M. L. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. Proceedings of the 21st international conference on World Wide Web, 2012. ACM, 879-888.
- ZHAI, Z., LIU, B., XU, H. & JIA, P. 2011. Constrained LDA for grouping product features in opinion mining. *Advances in knowledge discovery and data mining*, 448-459.
- ZHANG, H., GILES, C. L., FOLEY, H. C. & YEN, J. Probabilistic community discovery using hierarchical latent gaussian mixture model. *AAAI*, 2007. 663-668.
- ZHANG, J., LIU, B., TANG, J., CHEN, T. & LI, J. Social Influence Locality for Modeling Retweeting Behaviors. *IJCAI*, 2013. 2761-2767.
- ZHANG, L., SUN, X. & ZHUGE, H. 2015. Topic discovery of clusters from documents with geographical location. *Concurrency and Computation: Practice and Experience*, 27, 4015-4038.
- ZHANG, X.-P., ZHOU, X.-Z., HUANG, H.-K., FENG, Q., CHEN, S.-B. & LIU, B.-Y. 2011. Topic model for chinese medicine diagnosis and prescription regularities analysis: case on diabetes. *Chinese journal of integrative medicine*, 17, 307-313.
- ZHANG, Y., CHEN, M., HUANG, D., WU, D. & LI, Y. 2017. iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, 66, 30-35.
- ZHAO, F., ZHU, Y., JIN, H. & YANG, L. T. 2016. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. *Future Generation Computer Systems*, 65, 196-206.
- ZHAO, W. X., JIANG, J., WENG, J., HE, J., LIM, E.-P., YAN, H. & LI, X. Comparing twitter and traditional media using topic models. *European Conference on Information Retrieval*, 2011. Springer, 338-349.
- ZHENG, X., LIN, Z., WANG, X., LIN, K.-J. & SONG, M. 2014. Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. *Knowledge-Based Systems*, 61, 29-47.
- ZHU, J., AHMED, A. & XING, E. P. MedLDA: maximum margin supervised topic models for regression and classification. Proceedings of the 26th annual international conference on machine learning, 2009. ACM, 1257-1264.
- ZIRN, C. & STUCKENSCHMIDT, H. 2014. Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, 90, 38-53.
- ZOGHBI, S., VULIĆ, I. & MOENS, M.-F. 2016. Latent Dirichlet allocation for linking user-generated content and e-commerce data. *Information Sciences*, 367, 573-599.