A Survey of Topic Models in Text Classification

Linzhong Xia
School of Intelligent Manufacturing and Equipment
Shenzhen Institute of Information Technology
Shenzhen, China
e-mail: xialz@sziit.edu.cn

Dean Luo
School of Electronic Communication Technology
Shenzhen Institute of Information Technology
Shenzhen, China
e-mail: luoda@sziit.edu.cn

Chunxiao Zhang
School of Electronic Communication Technology
Shenzhen Institute of Information Technology
Shenzhen, China
e-mail: zhangcx@sziit.edu.cn

Zhou Wu
Sino-German School
Shenzhen Institute of Information Technology
Shenzhen, China
e-mail: wuz@sziit.edu.cn

***Abstract*—A massive of text that is generated every minute is increasing dramatically. Therefore, it is more and more important to find an effective model to automatic classify the amount of text. Topic models is the most powerful techniques in text classification. There are many research results in the field of topic model have been published in scholarly journals. The Latent Dirichlet Allocation (LDA) is one of the most popular topic models in text classification. Researchers have proposed many topic evolution models based on LDA to solve some specific problems in applications of text classification. And some joint models which based on topic models combined other algorithms have been studied to enhance the performance of text classification. In this paper, we investigated three categories topic models for text classification and briefly introduced their advantages and disadvantages in the applications of text mining. Also, we introduce the generated process of documents and illustrate the graphical model for each topic models.**

Keywords-topic model; latent dirichlet allocation; text classification; topic evolution model

I. INTRODUCTION

The explosion of electronic document archives gained a great deal of attention in recent years. A report sponsored by EMC predicts that the data volume will grow to 40 trillion gigabytes by 2020, leading to a 50-time growth from the beginning of 2010 [1]. Now the most important thing is to find an efficient tools or techniques to automatically organizing, searching, indexing, and browsing those electronic text data. The technique of topic models is one of the most important approach to solve the problem of text classification. This technique can discover the patterns which often reflect the underlying topics. Topic models are a well-know and significant modern machine learning technique. Given D is a set of documents composed of a set of terms W, T is a set of latent topics, that are created based on a statistical inference on term set W. So, documents are mixtures of topics, where a topic is a probability distribution over words. On the other hand, a document can be generated by a simple probabilistic procedure.

Topic models are a frequently used text classification tool for discovery of underlying semantic in a text body. For example, if that a document is about a special topic, one would expect special words to appear in the document: "student" and "school" will appear more often in documents about education, "GDP" and "finance" will appear in documents about economics. Thus, a document typically concerns multiple topics in different proportions.

Topic models rely on the bag of words (BOW) assumption which is ignoring the information of the order of words. So, generate a new document by choosing a distribution over topics, after that, each word can be chosen at random depends on a distribution over words of each topic.

Topic models are prominent for a lot of areas. Topic models are applied in various fields, such as: text classification [2]-[10], source code analysis [11], [12], emotion classification [13], [14], image classification [15], [16], opinion and aspect mining [17], [18], event detection [19], [20], system recommendation [21], [22], and so on.

The main goal of this paper is to provide a survey of topic models based on Latent Dirichlet Allocation (LDA). In a word, the main contribution of this paper is to investigate text classification scholarly articles which are related to topic models based on LDA and discover the research development.

This paper is organized as follows. Section II provides a literature review of the method of topic models. Section III overviews the detailed study of topic evolution models and some methods about joint of LDA and other algorithms in text classification. Section Ⅳ is the conclusion.

II. THE METHOD OF TOPIC MODELS

In this section, the main development history of topic models will be presented that deal with words, texts and topics.

978-1-7281-0831-5/19/$31.00 ©2019 IEEE

A. Latent Semantic Analysis (LSA)

The first major progress in text processing was due to the vector space model (VSM), in which a document is represented as a vector [23]. Although the VSM had been widely used in text processing in 1980s, it suffered from some inherent shortages to capture inter- and intra- document statistical structure, discriminate synonymy or polysemy and provide a small reduction only in the description of the corpus.

To overcome the shortages of the VSM, Deerwester et al. proposed a method of latent semantic analysis (LSA) [24]. The key idea of LSA is the analysis of the underlying semantic of document. LSA solves the problem of synonymy or polysemy by mapping the same documents or words into a different space and doing the comparison in the space with the method of Singular Value Decomposition (SVD) of term-document matrix. In the process of mapping, the high dimensional document vector will be transformed to low dimensional vector. Although LSA overcomes some of the drawbacks of the VSM, it suffers from some of limitations also. First, the computational cost of the SVD is expensive. Second, the new feature space which obtained by the SVD is difficult to interpret.

B. Probabilistic Latent Semantic Analysis (PLSA)

As a major advance in the application of Bayesian methods to document modeling, Jan Puzicha and Thomas Hofmann introduced probabilistic latent semantic analysis (PLSA) as an alternative to LSA [25]. The key idea of PLSA is a statistical model called aspect model. It is a latent variable model that associates an unobserved latent variable $z_k \in \{z_1, z_2, ..., z_K\}$ with each document d and represents each aspect by a distribution over words $p(\mathbf{w} \mid \mathbf{z})$. The joint distribution of a document d and a word w_{di} can be written as (1):

$$p(d, w_{di}) = p(d)\sum_{z=1}^{K} p(w_{di} \mid z)p(z \mid d) \qquad (1)$$

where w_{di} represents the ith word of the document d, $i \in \{1, 2, 3, ..., N_d\}$ and the N_d is the number of word in document d, $p(d)$ represents the probability of a document d in the data set, $p(w_{di} \mid z)$ represents the conditional probability of w_{di} under topic z, $p(z \mid d)$ represents conditional probability of topic z under document d. The graphical model of PLSA is shown in Fig. 1.

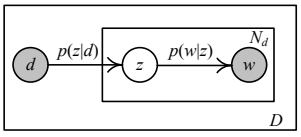

Figure 1. A graphical model of PLSA.

As in Fig. 1, the documents and words are conditionally independent. Each word for a document can be generated based on the underlying topics. Therefore, the generative process can be described as follows:

1) *Select a document with probability* $p(d)$.
2) *For each word i in document* d *:*
   a) *Select a latent class* z_i *with probability* $p(z_i \mid d)$.
   b) *Select a word* w_{di} *with probability* $p(w_{di} \mid z_i)$.

The main idea of PLSA is recognizing and distinguishing between different contexts of word usage without recourse to a dictionary or thesaurus. It includes two important implications: it allows us to disambiguate polysemy; it exposes topic similarities by grouping together words that shared a common context [26]. However, PLSA is not a real generative model as the variable d is a dummy random variable that is indexed by the documents in a training set [27]. Therefore, PLSA is inclined to overfit the training data.

C. Latent Dirichlet Allocation (LDA)

LDA is an algorithm for text classification that is based on statistical topic models and it is very widely used. The different with PLSA and LSA is that the LDA can capture the exchangeability of both words and documents [27]. LDA is a generative probabilistic model of a corpus. The documents of the corpus are represented as random mixtures over latent topics. Each latent topic of documents is characterized by a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well. The graphical model of LDA is shown in Fig. 2.

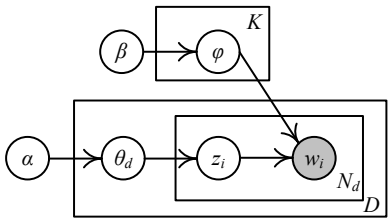

Figure 2. A graphical model of LDA.

As in Fig. 2, given a corpus D consisting of M documents, with document d having $N_d \in \{N_1, N_2, ..., N_M\}$ words, the generative process of LDA models D as follows:

1) Draw K multinomials φ_k *from a Dirichlet prior* β*, one for each topic* k.

2) Draw D multinomials θ_d *from a Dirichlet prior* α*, one for each document* d.

3) For each word w_{di} *in the document and each document* d *in the corpus:*

a) Draw a topic z_i *from multinomial* θ_d*;* $(p(z_i \mid \alpha))$.

b) Draw a word w_i *from multinomial* φ_z *;* $(p(w_i \mid z_i, \beta))$.

To compute the generative process corresponds to inferring the latent variables and learning the distributions of underlying topics. The process is generated by using (2):

$$p(\Theta, z, \Phi \mid \mathbf{w}, \alpha, \beta) = \frac{p(\mathbf{w}, \Theta, z, \Phi \mid \alpha, \beta)}{\int_{\varphi_{1:K}} \int_{\theta_{1:D}} p(\mathbf{w} \mid \alpha, \beta)} \qquad (2)$$

In LDA, exploring the data and extracting the topics correspond to computing the distribution of document-topic θ_d and the distribution of topic-word φ_k . Griffiths and Steyvers proposed a simple and effective strategy for estimating θ_d and φ_k [28]. It is an approximate iterative technique that we call the method of Gibbs sampling. The specific process of Gibbs sampling as follows:

Step 1. Initialize topic for each word of the corpus: $z_{d,n} = k : \ Mult(1/K)$. Where $n \in (1,2,...,N_d)$ and $k \in (1,2,...,K)$.

Step 2. When a new word is observed, given the word w_i equal to t , we can compute the current topic of the word by using (3):

$$P(z_i = k \mid z_{\neg i}, w) = \frac{c^{(t)}_{k,\neg i} + \beta_t}{\sum_{t=1}^{V} c^{(t)}_{k,\neg i} + \beta_t} \bullet \frac{c^{(k)}_{d,\neg i} + \alpha_k}{\sum_{k=1}^{K} c^{(k)}_{d,\neg i} + \alpha_k} \tag{3}$$

where $c^{(t)}_{k,\neg i}$ represents the number of times word $w_i = t$ is assigned to topic k , not including the current token instance i ; $\sum_{t=1}^{V} c^{(t)}_{k,\neg i}$ represents the number of times all words of the corpus is assigned to topic k , not including the current token instance i ; $c^{(k)}_{d,\neg i}$ is the number of times topic k is assigned to some words in document d , not including the current token instance i ; $\sum_{k=1}^{K} c^{(k)}_{d,\neg i}$ is the number of words of the document d which include word w_i , not including the current token instance i ; V is the total number of terms in corpus; K is the total number of topics in corpus.

Step 3. Repeat the process of step 2 until all the distribution of topics achieve convergence. After the word token of each word is finalized, the parameters of θ_d and φ_k can be computed by using (4) and (5):

$$\theta_{d,k} = \frac{c^{(k)}_d + \alpha_k}{\sum_{k=1}^{K} c^{(k)}_d + \alpha_k} \tag{4}$$

$$\varphi_{k,t} = \frac{c^{(t)}_k + \beta_t}{\sum_{t=1}^{V} c^{(t)}_k + \beta_t} \tag{5}$$

where $\theta_{d,k}$ represents the probabilistic of topic k in document d ; $c^{(k)}_d$ represents the number of words is assigned to topic k in document d ; $\sum_{k=1}^{K} c^{(k)}_d$ represents the total number of words in document d ; $\varphi_{k,t}$ represents the probabilistic of word t in topic k ; $c^{(t)}_k$ represents the number of word t is assigned to topic k ; $\sum_{t=1}^{V} c^{(t)}_k$ represents the total number of word is assigned to topic k in corpus.

III. Methods about Topic Evolution Models in Text Classification

There are many limitations for LDA in text classification. To overcome those limitations, there are many topic evolution models have been created. In this section, we will review several important papers related to topic evolution models.

A. Correlated Topic Models (CTM)

LDA is a useful tool for the statistical analysis of text datasets and other discrete data. But the LDA model unable to model correlation between topics. For example, a scholarly paper about genetics may be likely to also be about disease, but unlikely to also be about x-ray astronomy. The limitation of LDA model stems from the independence assumption in the Dirichlet distribution on the topic proportions.

To overcome the limitation of the LDA model, the correlated topic model (CTM) has been created by Blei et al. [29]. In CTM model, it is allowed that one latent topic can be correlated with another latent topic. The description of the interdependency via covariance matrix which come from the logistic normal distribution. The graphical model of CTM model is shown in Fig. 3.

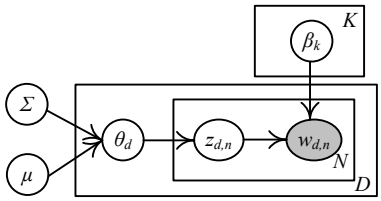

Figure 3. A graphical model of CTM.

As in Fig. 3, an N-word document generative process as follows:

1) *Draw* $\theta \mid \{\mu, \Sigma\} : \ ? \{\mu, \Sigma\}$.
2) *For* $n \in \{1,2,...N\}$ *:*
   a) *Draw topic assignment* $z_n \mid \theta$ *from* $Mult(f(\theta))$.
   b) *Draw word* $w_n \mid \{z_n, \beta_{1:K}\}$ *from* $Mult(\beta_{z_n})$.

where $\{\mu, \Sigma\}$ represents a K-dimensional mean and covariance matrix. This generative process of CTM model is like LDA model except that the topic proportions of CTM model are drawn from a logistic normal distribution.

B. Dynamic Topic Models (DTM)

In an exchangeable topic models, the documents of a corpus are assumed to be exchangeable in sequence. But in many document collections, the implicit assumption of exchangeable documents is inappropriate. Document collections such as scholarly papers, e-mails, news articles and so on reflect evolving content.

Blei and Lafferty have developed a dynamic topic model (DTM) which captures the evolution of topics in a sequentially organized document collection [30]. In DTM, we assume that the corpus is divided by time slice, such as year, month and so on. K topics in documents which

modeled by DTM for each time slice, where the topics associated with time slice t evolve from the topics associated with slice $t-1$. The graphical model of DTM is shown in Fig. 4.

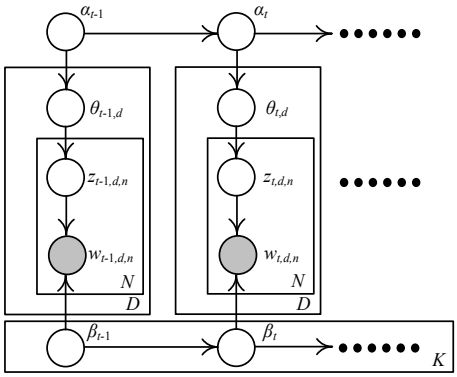

Figure 4. A graphical model of DTM.

As in Fig. 4, the generative process for time slice t of a corpus is as follows:

1) *Draw topics* $\beta_t \mid \beta_{t-1} : ?(\beta_{t-1}, \sigma^2 I)$.
2) *Draw* $\alpha_t \mid \alpha_{t-1} : ?(\alpha_{t-1}, \delta^2 I)$.
3) *For each document* d_t *:*

a) *Draw* $\theta_{t,d} : ?(\alpha_{t-1}, a^2 I)$.

b) *For each word* $w_{t,d}$ *:*

Draw $z_{t,d,n} : Mult(\pi(\theta_{t,d}))$.

Draw $w_{t,d,n} : Mult(\pi(\beta_{t,z_{t,d,n}}))$

C. Supervised Topic Models

In most topic models, only the words in the document are modelled. These kind of topic models are unsupervised. The unsupervised LDA has been used to extract features for classification. And the unsupervised LDA act to reduce data dimension. But when the goal is prediction, the unsupervised LDA may not be a good model.

Blei and McAuliffe have developed a supervised topic model which considers the grades of essays, the numerical ratings of movie reviews, and web pages with counts of how many online community members liked them [31]. In supervised LDA [sLDA], document is paired with a response and the goal is to infer latent topics predictive of the response. For example, if we predict a movie rating with words based on its reviews by using sLDA, the results of the prediction may be "terrible", "average", and "perfect", without regard to genre. But if you predict it by using unsupervised LDA, the result of the prediction may be genre. The graphical model of sLDA is shown in Fig. 5.

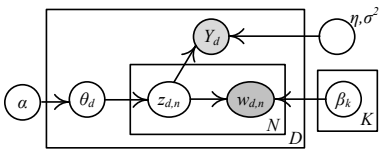

Figure 5. A graphical model of sLDA.

As in Fig. 5, the generative process of each document in corpus and response as follows:

1) *Draw topic proportion* $\theta_d \mid \alpha : Dir(\alpha)$.
2) *For each word* $w_{d,n}$ *:*

a) *Draw topic assignment* $z_{d,n} \mid \theta_d : Multi(\theta_d)$.

b) *Draw word* $w_{d,n} \mid z_{d,n}, \beta_{1:K} : Multi(\beta_{z_{d,n}})$.

3) *Draw response variable* $y \mid z_{d,1:N}, \eta, \sigma^2 : ?(\eta^T \bar{z}_d, \sigma^2)$. *Here we define* $\bar{z}_d := (1/N)\sum_{n=1}^{N} z_{d,n}$.

D. Labeled LDA (L-LDA)

LDA is a kind of unsupervised topic model, it is not appropriate way to model a multi-labeled corpus because it can't incorporate a supervised label set into the model process. In LDA, the number of topics which learn from a corpus is fixed and the LDA compulsory distribute proportion to each topic for each document. Therefore, the topics which learned by LDA are hard to interpret, and the model provides no tools for tuning the generated topics to suit an end-use application. Meanwhile, the sLDA has been limited to associate with only a single label for each document. So, the sLDA is inappropriate for multiply labeled corpus.

Daniel Ramage et al. have developed a new topic model which called Label LDA (L-LDA), it associates each label with one topic in direct correspondence [32]. The L-LDA can solve the problem of the compulsory distribution in LDA and the problem of interpretability of topics over LDA. The L-LDA is a supervised topic model which use only those topics that correspond to a document's label set. The graphical model of L-LDA is shown in Fig. 6.

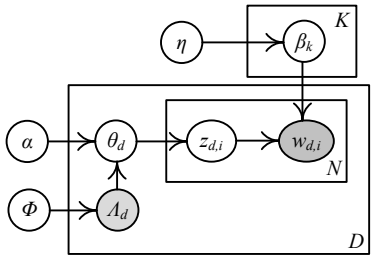

Figure 6. A graphical model of L-LDA.

As in Fig. 6, the generative process for L-LDA as follows:

1) *Generate* $\beta_k \mid \eta : Dir(\eta)$ *for each topic* $k \in \{1, 2, ..., K\}$.
2) *For each document* d *in corpus:*

a) *For each topic* k *:*

Generate $\Lambda_{d,k} \in \{0,1\} \mid \Phi_k : Bernoulli(\Phi_k)$.

b) *Generate* $\boldsymbol{\alpha}_d = L_d \times \boldsymbol{\alpha}$.

c) *Generate* $\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}_d : Dir(\boldsymbol{\alpha}_d)$.

d) *For each* $i \in \{1, ..., N_d\}$ *:*

Generate $z_{d,i} \in \{\lambda_{d,1}, ..., \lambda_{d,M_d}\} \mid \boldsymbol{\theta}_d : Multi(\boldsymbol{\theta}_d)$.

Generate $w_{d,i} \mid \beta_{z_{d,i}} : Multi(\beta_{z_{d,i}})$.

where labels $\mathbf{\Lambda}_d = (l_1, ..., l_K)$ and $l_k \in \{0,1\}$, L_d is a matrix of size $M_d \times K$ for each document d, $L_{d,ij} = \{1, if\ \lambda_{d,i} = j; 0, therwise\}$, $i \in \{1, ..., M_d\}$, $j \in \{1, ..., K\}$, $\lambda_d = \{k \mid \Lambda_{d,k} = 1\}$, $M_d = |\lambda_d|$.

E. gLDA

In LDA model, we assume that the words occur independently, and the document is comprised by the "bag of words". Therefore, the words didn't relevant or less relevant to the document topics at some time. To overcome the limitation of topic-words relevance in LDA, it is worth to make study how to solve the problem.

Zhao et al. developed a new topic model which called gLDA [33]. In gLDA, documents have been divided into several categories, and there are special set of "topics" for different category. Therefore, we can make the document generated from the most relevant category by adding topic-category distribution parameter on the foundation of LDA. The graphical model of gLDA is shown in Fig. 7.

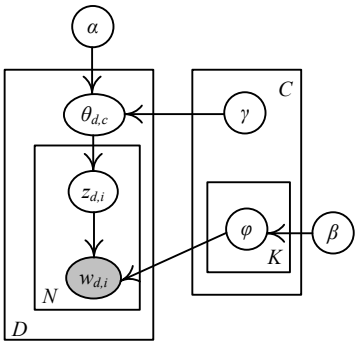

Figure 7. A graphical model of gLDA.

As in Fig. 7, the generative process for gLDA as follows:

1) For each topic k, select a word distribution $\varphi_k \mid \beta : Dir(\beta)$.

2) For each document d:

a) Select a category $c \mid \gamma : Dir(\gamma)$.

b) Select a distribution over topics $\theta_{d,c} \mid \alpha : Dir(\alpha)$.

c) For each word in document d:

Select a topic $z_{d,i} \mid \theta_{d,c} : Dir(\theta_{d,c})$.

Generate a word $w_{d,i} \mid \varphi_{z_{d,i}} : Dir(\varphi_{z_{d,i}})$.

where c represents different categories.

F. The Joint of Topic Model and Other Algorithms

The purpose of text classification is to assign the most suitable label to a specified document. But there are many challenges to text classification, such as the documents are lack label information, the format of document is sparse, the length of document is different and so on [34]. The problem of lack label information can be solved by adopting a semi-supervised learning (SSL) approach [9]. The problem of structure sparse for document can be solved by adopting multiple document representation schemes approach [35]. In a word, we can't solve all problems by adopting single model. Therefore, the joint of topic model and other algorithms is the best choice to enhance the performance of text classification for complicated corpus.

The model of K-nearest Neighbor (KNN) is a traditional method in the area of text classification. It can calculate the category of document by considering the similarity of feature words. The semantic similarity has been ignored in KNN. Chen et al developed a joint model which take advantage of KNN's and LDA's advantages to solve the both problems [36]. The joint model of LDA-KNN which compared with traditional models has superior classification performance in automatic text categorization.

The model of Vector Space Model (VSM) is a traditional method in the area of text classification, too. It can capture the category of document by considering the weight statistics and similarity calculation. But there are some issues for VSM model, such as the data latitude is too high, lack of understanding and so on. Liu et al developed a joint model of VSM model and LDA model [37]. They computed the similarity by adopting VSM model and LDA model respectively. The result of the mixed similarity has been obtained by combining with linear addition method. Then through the K-means algorithm for text clustering based on the result of the mixed similarity. The results show that this joint model is effective.

The number of today's web text is exploding. But the length of text is shorter and shorter. The problem that follows is the problem of the sparse text. The traditional topic models can't achieve good results in short text classification. Sun et al designed a short text classification method based on word vector (Word2vec) and LDA model [38]. They train the LDA model by the way of Gibbs sampling firstly. Then train the word vectors by adopting the Word2vec model and vectorized with the Topic High Frequency Word. Finally, through the support vector machine (SVM) algorithm to classify the short texts. The results show that this joint model can significantly improve the performance of text classification.

IV. CONCLUSION

Topic models have been widely applicated in text mining, latent knowledge discovery, summarized huge electronic archives and so on. In this survey, three categories topic models for text classification have been introduced. The first category, it has discussed the traditional topic models including LSA, PLSA, and LDA. The second category, it has introduced the topic evolution models including CTM, DTM, sLDA, L-LDA, and gLDA. The third category, it has explained some joint models of topic model and other algorithms including LDA-KNN, LDA-VSM+K-means, LDA-Word2vec+SVM.

This paper overview some of the most important topic model approaches in the field of text classification. It does not go into specific details. It only describes the high-level view of these models that relates to topic models in text classification. Furthermore, it has mentioned the advantages and the limitations for each topic model. Also, it has been mentioned that each of these topic models has improved and

modified over the previous one. In a word, this paper investigated the most important scholarly articles highly related to topic model based on LDA in the field of the text classification. Given the importance of this survey, we believe this paper can be an important source and good opportunities for text classification with topic models based on LDA for researchers and future works.

ACKNOWLEDGMENT

This work is supported by Engineering Applications of Artificial Intelligence Technology Laboratory of Shenzhen Institute of Information Technology (Number: PT201701).

REFERENCES

[1] J. Ganz and D. Reinsel, "THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the far East," Technical Report 1. IDC, Framingham, Dec. 2012, pp. 1-16.

[2] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," Proc. of the twenty-third international conference on machine learning (ICML 2006), ACM Press, Jun. 2006, pp. 577-584, doi: 10.1145/1143844.1143917.

[3] L. AlSumait, D. Barbara, and C. Domeniconi, "On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," 2008 Eighth IEEE International Conference on Data Mining, IEEE Press, Dec. 2008, pp. 3-12, doi: 10.1109/ICDM.2008.140.

[4] Y. Wang, E. Agichtein, and M. Benzi, "TM-LDA: efficient online modeling of latent topic transitions in social media," Proc. of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, Aug. 2012, pp. 123-131, doi: 10.1145/2339530.2339552.

[5] G. Liu, X. Xu, Y. Zhu, and G. Li, "An improved Latent Dirichlet Allocation Model for Hot Topic Extraction," 2014 IEEE International Conference on Big Data and Cloud Computing (BdCloud), IEEE Press, Dec. 2014, pp. 800-809, doi: 10.1109/BDCloud.2014.55.

[6] D. Ramage, C. D. Manning, and S. Dumais, "Partially labeled topic models for interpretable text mining," Proc. of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, Aug. 2011, pp. 457-465, doi: 10.1145/2020408.2020481.

[7] X. Mao, Z. Ming, T. Chua, S. Li, H. Yan, and X. Li, "SSHLDA: a semi-supervised hierarchical topic model," Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics Press, Jul. 2012, pp. 800-809.

[8] Z. Liu, Y. Zhang, E. Chang, and M. Sun, "PLDA+: Parallel latent Dirichlet allocation with data placement and pipeline processing," ACM Transactions on Intelligent Systems and Technology, vol. 2, Apr. 2011, pp. 26:1-18, doi: 10.1145/1961189.1961198.

[9] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec," Information Sciences, vol. 477, Mar. 2019, pp. 15-29, doi: 10.1016/j.ins.2018.10.006.

[10] X. Chen, Y. Xia, P. Jin, and J. Carroll, "Dataless Text Classification with Descriptive LDA," Proc. of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI Press, Jan. 2015, pp. 2224-2231.

[11] E. Linstead, P. Rigor, S. Bajracharya, C. Lopes, and P. Baldi, "Mining concepts from code with probabilistic topic models," Proc. of the twenty-second IEEE/ACM international conference on automated software engineering, ACM Press, Nov. 2007, pp. 461-464, doi: 10.1145/1321631.1321709.

[12] S. K. Lukins, N. A. Kraft, and L. H. Etzkorn, "Bug localization using latent Dirichlet allocation," Information and Software Technology, vol. 52, Sep. 2010, pp. 972-990, doi: 10.1016/j.infsof.2010.04.002.

[13] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu, "EmpaTweet: Annotating and Detecting Emotions on Twitter," Proc. of LREC2012, 2012, pp. 3806-3813.

[14] Y. Rao, "Contextual sentiment topic model for adaptive social emotion classification," IEEE Intelligent Systems, vol. 31, Jan. 2016, pp. 41-47, doi: 10.1109/MIS.2015.91.

[15] Y. Wang, and G. Mori, "Max-margin Latent Dirichlet Allocation for Image Classification and Annotation," The 22nd Proc. of British Machine Vision Conference (BMVC), BMVA Press, Sep. 2011, pp. 112:1- 11, doi: 10.5244/C.25.112.

[16] M. Kandemir, T. Kekec, and R. Yeniterzi, "Supervising topic models with Gaussian processes," Pattern Recognition, vol. 77, May 2018, pp. 226-236, doi: 10.1016/j.patcog.2017.12.019.

[17] X. Zheng, Z. Lin, X. Wang, K.-J Lin, and M. Song, "Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification," Knowledge-Based Systems, vol. 61, May 2014, pp. 29-47, doi: 10.1016/j.knosys.2014.02.003.

[18] H. Nabli, R. B. Djemaa, I. A. B. Amor, "Efficient cloud service discovery approach based on LDA topic modeling," The Journal of Systems and Software, vol. 146, Dec. 2018, pp. 233-248, doi: 10.1016/j.jss.2018.09.069.

[19] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," IEEE Transactions on Multimedia, vol. 18, Feb. 2016, pp. 233-246, doi: 10.1109/TMM.2015.2510329.

[20] S. Jia, and B. Wu, "Incorporating LDA based text mining method to explore new energy vehicles in China," IEEE Access, vol. 6, Nov. 2018, pp. 64596-64602, doi: 10.1109/ACCESS.2018.2877716.

[21] S. Zoghbi, I. Vulic, and M.-F. Moens, "Latent Dirichlet allocation for linking user-generated content and e-commerce data," Information Sciences, vol. 367, Nov. 2016, pp. 573-599, doi: 10.1016/j.ins.2016.05.047.

[22] Z. Cheng, and J. Shen, "On effective location-aware music recommendation," ACM Transactions on Information Systems, vol. 34, Apr. 2016, pp. 1-32, doi: 10.1145/2846092.

[23] G. Salton, and M. J. McGill, Introduction to modern information retrieval, New York: McGraw-Hill. 1983.

[24] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," Journal of The American Society for Information Science, vol. 41, Sep. 1990, pp. 391-407, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[25] T. Hofmann, "Probabilistic latent semantic indexing," Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, Aug. 1999, pp. 50-57, doi: 10.1145/312624.312649.

[26] T.Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Machine learning, vol. 42, 2001, pp. 177-196, doi: 10.1023/a:1007617005950.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, May 2003, pp. 993-1022, doi: 10.1162/jmlr.2003.3.4-5.993.

[28] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. of National Academy of Sciences, Apr. 2004, pp. 5228-5235, doi: 10.1073/pnas.0307752101.

[29] D.M. Blei and J. D. Lafferty, "Correlated Topic Models," Advances in Neural Information Processing Systems 18 (NIPS 2005), NIPS Press, Dec. 2005, pp. 147-154, doi: 10.1145/1143844.1143859.

[30] D. M. Blei and J. D. Laffery, "Dynamic Topic Models," Proc. of the 23rd international conference on Machine learning, ACM Press, Jun. 2006, pp. 113-120, doi: 10.1145/1143844.1143859.

[31] D. M. Blei and J. D. McAuliffe, "Supervised topic models," Advances in Neural Information Processing, Mar. 2010, pp. 1-22, doi: 10.1109/ICPR.2014.65.

[32] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labled corpora," Proc. of the 2009 Conference on Empirical Methods in

Natural Language Processing (EMNLP), ACL Press, Aug. 2009, pp. 248-256.

[33] D. Zhao, J. He, and J. Liu, "An Improved LDA Algorithm for Text Classification," 2014 International Conference on Information Science, Electronics and Electrical Engineering, IEEE Press, Nov. 2014, pp. 217-221, doi: 10.1109/InfoSEEE.2014.6948100.

[34] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and LDA topic models," Expert Systems With Applications, vol. 80, Sep. 2017, pp. 83-93, doi: 10.1016/j.eswa.2017.03.020.

[35] B. S. Harish, D. S. Guru, and S. Manjunath, "Represention and classification of text document: a brief review," Internatonal Journal of Computer Applications, Jan. 2010, pp. 110-119.

[36] W. Chen and X. Zhang, "Research on text categorization model based on LDA-KNN," 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), IEEE Press, Mar. 2017, pp. 2719-2726, doi: 10.1109/IAEAC.2017.8054520.

[37] X. Liu, H. Xiong, and N. Shen, "A Hybrid Model of VSM and LDA for Text Clusteing," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Sep. 2017, pp. 230-233.

[38] F. Sun and H. Chen, "Feature Extension for Chinese Short Text Classification Based on LDA and Word2vec," 2018 13th IEEE International Conference on Industrial Electronics and Applications (ICIEA), Jun. 2018, pp. 1189-1194.