Front. Comput. Sci., 2022, 16(6): 166350
https://doi.org/10.1007/s11704-021-0606-3

RESEARCH ARTICLE

PSLDA: a novel supervised pseudo document-based topic model for short texts

Mingtao SUN[1], **Xiaowei ZHAO**[2], **Jingjing LIN**[3], **Jian JING**[2], **Deqing WANG**[2], **Guozhu JIA** (✉)[1]

1 School of Economics and Management, Beihang Univeristy, Beijing 100191, China
2 School of Computer Science, Beihang University, Beijing 100191, China
3 School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China

Abstract Various kinds of online social media applications such as Twitter and Weibo, have brought a huge volume of short texts. However, mining semantic topics from short texts efficiently is still a challenging problem because of the sparseness of word-occurrence and the diversity of topics. To address the above problems, we propose a novel supervised pseudo-document-based maximum entropy discrimination latent Dirichlet allocation model (PSLDA for short). Specifically, we first assume that short texts are generated from the normal size latent pseudo documents, and the topic distributions are sampled from the pseudo documents. In this way, the model will reduce the sparseness of word-occurrence and the diversity of topics because it implicitly aggregates short texts to longer and higher-level pseudo documents. To make full use of labeled information in training data, we introduce labels into the model, and further propose a supervised topic model to learn the reasonable distribution of topics. Extensive experiments demonstrate that our proposed method achieves better performance compared with some state-of-the-art methods.

Keywords supervised topic model, short text, pseudo-document

1 Introduction

The unprecedented growth of various kinds of Web applications, especially online social media such as Twitter and Weibo, have witnessed a huge volume of short texts. It may be difficult to find the knowledge contained in these short texts because of their sparseness [1], so how to mine the knowledge efficiently from short texts is still a challenging problem for researchers.

Topic models aim to extract discriminative and coherent latent semantic topics from a large number of documents. Many topic modeling algorithms such as probabilistic latent semantic analysis (PLSA) [2], latent Dirichlet allocation (LDA) and their variants [3−6] have been widely adopted for discovering latent semantic knowledge from text corpus. In the above algorithms for probabilistic graphical topic models (such as LDA), each document can be viewed as generated from various topics and each topic is viewed as a distribution over all the words in the document. These models utilize the word co-occurrence in the text collection to mine the potential semantic information in the document, and achieve better performance for long texts, such as news, papers.

However, traditional topic models often fail to mine topics of short texts such as tweets, due to the sparsity of features and the limited word co-occurrence information available in short texts. Thus many researchers have made efforts to address the issue: how to extract semantic and meaningful topics from short texts. One common way is to aggregate short texts into long ones with the help of external knowledge or augmented data into sparse texts. For example, both [7] and [8] proposed to aggregate the tweets of the same Twitter user before applying topic model algorithms, and their experiments showed that the proposed aggregation way indeed improved the performance of the topic model. And [9] aggregated tweets in a data pre-processing step for LDA through various pooling schemes. Besides, some studies took the use of auxiliary information into the generative process of their proposed models. For instance, Phan et al. [10] collected a very large general dataset and trained the topic model on it, then they applied the trained model to short texts. The inclusion of the mined topics from general dataset will reduce the sparsity of short texts and improve the topic coherence. However, the above methods all suffer from high time complexity because of the pre-trained processing on the external large-scale dataset.

Self-aggregation-based methods are proposed to perform topic modeling and text self-aggregating simultaneously during the topic inference stage. In this way, short texts are merged into long pseudo-documents before topic inference that can help improve word co-occurrence information. For example, SATM [11] and PTM [12] posited that each short text was sampled from a long pseudo-document unobserved in current text collection, and inferred latent topics from long pseudo-documents, without depending on auxiliary

Received December 21, 2020; accepted July 20, 2021

E-mail: jiaguozhu@buaa.edu.cn

information or metadata. Unfortunately, it is expensive and even prohibitive to collect enough training samples for an effective model, especially when these samples need fine-grained annotations.

To address the above problems for modeling short texts, we propose a new topic model named supervised pseudo-document-based maximum entropy discrimination latent Dirichlet allocation (PSLDA for short). In PSLDA, we assume that short texts are generated from the normal-sized latent documents. Because of this assumption, PSLDA is superior to other traditional text topic models when the length of the training text is small. In this way, the relationship between short texts and pseudo documents can be modeled. PSLDA effectively avoids over-fitting on short-length corpora and has a fixed number of parameters. In addition to the document content itself, there are related text categories, regression responses and other information. Through making use of these supervision information effectively and expand the original unsupervised model into supervised model, we can learn the theme distribution implied by the document and a classification or regression model at the same time, which can be better used to predict the task. To include labels into model, we propose an expected classifier to calculate the reasonable distribution of topics. So PSLDA can describe the observed short texts and make accurate predictions for similar texts as well.

The contributions of the proposed method are threefold: (1) We propose a novel PSLDA for modeling the topic distribution of short texts. It assumes that short texts are generated from pseudo documents, and the topic distribution is sampled from the pseudo documents. (2) Labeled information can also be included through an expected classifier to predict similar texts. (3) Experiments on three real-world datasets show that our model outperforms several state-of-the-art topic models on short text corpus, and gain better coherence on topic semantic.

2 Related work

2.1 Topic models

For modeling text, the topic model often depicts the process of generating words in a document from the perspective of the topic. That is, it treats a document as a process of generating topics from a document and then generating words from the topics. Actually, the topic model utilizes the word co-occurrence in the document set to mine the potential semantic information in the document. For example, classic Probabilistic Latent Semantic Indexing (pLSI) [2] considered a document as a mixture of topics while a topic as a mixture of words, which succeeded in uncovering topics of the documents. However, when the numbers of total documents and vocabulary increase, the complexity of the pLSI model also increases linearly. Then latent Dirichlet allocation (LDA) [3] improved pLSI by imposing Dirichlet priors α and β, which assumed the topics were obeyed with Dirichlet distribution. Finally, LDA has been widely employed in many topic modeling tasks and many different domains, such as text domain, image domain, and so on. Various topic-based model extension models, such as Dynamic Topic Model (DTM) [13], author-topic model [14], latent Feature LDA [15], and Hashtag-LDA [16], have been proposed to solve specific domain problems. [17] proposes an unified latent variable model (contraLDA) to mine contrastive opinion to identify, extract and organise opinions from user generated texts. [18] proposes two alternative distribution models to learn word representations by employing the neighbor and syntactic contexts.

2.2 Topic models for short text

However, none of them has considered their applicability to short texts. The major problem of topic modeling on short text is that the short text has less co-occurrence information, and their feature space is relatively sparse. Therefore, some researchers tried to address the problem of modeling short texts by aggregating short texts into normal ones, employing auxiliary corpus, or generating pseudo documents. For instance, Hong and Davison [7] aggregated the tweets of the same Twitter user before learning the topic model. Similarly, Weng et al. [19] proposed to aggregate tweets from the same user into pseudo-documents, and then used them to train classic topic models. Since they focused on user’s topical interests rather than a single tweet, the above aggregation was meaningful because one user’s interests were related to his whole tweets.

Besides, some scholars imposed various assumptions to improve the topic model. For example, Jin et al. [20] improved the effectiveness of topic mining on short texts by constructing a topic model from auxiliary long texts, and then transferred the semantic knowledge from long texts to short texts through transfer learning methods. Lin et al. [21] introduced sparse constraints on both topic distribution and topic word distribution of the document in order to model topics of short texts. The BTM (Biterm Topic Model) [22] model was based on the co-occurrence relationship of words in the corpus. In BTM, the text is pre-processed to generate word pairs, and then the tokens (words in the classic topic model) are modeled with word pairs Modeling. Recently, Zuo et al. [23] proposed Word Net Topic Model (WNTM) to synthesize pseudo-documents from short texts, then performed topic modeling. Zuo et al. proposed the PTM model [12], which introduced the concept of the pseudo document to implicitly aggregate short texts against data sparsity.

Yin et al. [24] proposed the GSDMM model to cope with the sparse and high-dimensional problem of short texts, and can obtain the representative words of each cluster. GPU-DMM [25] exploited auxiliary word embeddings to enrich topic modeling for short texts is the main focus of this paper. There are also some methods for topic modeling of short texts that creates pseudo-documents representations from the original documents [26–28]. [27] creating larger pseudo-document representations from the original documents. [29] propose a semantics-assisted non-negative matrix factorization (SeaNMF) model to discover topics for the short texts. It effectively incorporates the word-context semantic correlations into the model, where the semantic relationships between the words and their contexts are learned from the skip-gram view of the corpus.

In recent years, deep learning has attracted a lot of attention, and many deep learning based methods have been proposed for topic modeling. These deep learning based methods can be divided into two categories: 1) AutoEncoder-based models. Miao et al. [30] first use VAE Document model to construct neural variational document model NVDM, and consider the topic vocabulary distribution. Ding et al. [31] propose a neural topic model focusing on the consistency of topic semantic expression. The model uses the pre-trained word vector to measure the semantic similarity between quantifier pairs and takes it as part of the NVDM optimization function. 2) Neural topic models based on sparse constraints. The neural sparse topic coding model NSTC proposed by Peng et al. [32] uses word vector to improve topic aggregation, and uses neural network based on backward propagation algorithm to simplify the parameter inference process. [33] propose a neural topic model framework based on metadata as label assisted modeling. The framework uses various metadata as label information to solve the problem of multi-label classification. At the same time, the sparsity of topic distribution is controlled by exponential priori. Although the DL-based methods can achieve high accuracy, the models are not interpretable. Besides, DL-based models are expensive to train. Our model is based on a probability graph model, which is an interpretable model widely used in the topic modeling field.

3 Model and inference

To cope with the rapid growth of extremely short texts such as Twitter or Weibo, we propose a pseudo-document-based supervised topic model for short text information mining. The solution is called PSLDA, a supervised pseudo document based (maximum entropy discrimination) latent Dirichlet allocation model.

In PSLDA, the most significant assumption is that short texts are gengaged from the normal size latent documents. Because of this assumption, PSLDA is superior to other traditional text topic models when length of the training text is small. The traditional topic models fail to mine high-quality topics from short texts mainly due to the lack of word co-occurrence patterns. It has been proven that when the number of tokens **N** in short texts **D** is too small, topic models such as LDA can not learn topics accurately even though **D** is extremely large. So if we assume that short texts are generated from pseudo documents P rather than D, with $\mathbf{P} << \mathbf{D}$, we can approximate that each pseudo document has $\mathbf{N}_0$ tokens on average, $\mathbf{N}_0 = \mathbf{DN}/\mathbf{P}$, which implies the potential improvement of word co-occurrence. In this way, we can obtain implicit aggregation of short texts to against data sparsity.

PSLDA effectively avoids overfitting on short-length corpora and has a fixed number of parameters. We will first introduce the basics of the model, and then discuss the inferences of PSLDA in the following parts of this section.

3.1 Basic model

Considering the limitations of space, we will introduce our model in a general text binary classification problem. Firstly we have a labeled training set $D_s = \{(d_s, y_s)\}_{s=1}^{D}$ and an unlabeled test set $D_t = \{d_t\}_{t=1}^{D_t}$ where d_s denotes short text in a training set, y_s is the label of the corresponding short text (also called response variable), and the range of response variable here Y is $y = \{-1, +1\}$, d_t denotes short text in a test set. Our task is to train a classifier $f : D_t \rightarrow Y$.

PSLDA is a hierarchical Bayesian model just as Fig. 1 shows. We assume there exists a set of unobserved latent documents (i.e., pseudo documents), and with the help of that pseudo documents we can handle the problem of short texts classification in a good manner just as the final experiment results shows.

This model consists of two parts — a Pseudo-document-based Topic Model (PTM) for observed short texts $\{\boldsymbol{d}_s\}_{s=1}^{D}$ proposed by Zuo et al. [12] which is used to model the relationship between short texts and pseudo documents, and an expected classifier with respect to the supervising signal $\boldsymbol{y} = \{y_s\}_{s=1}^{D}$. Combining the two parts, PSLDA has the ability to learn labeled short texts and efficiently avoid overfitting at the same time. The details of the model are elaborated as follows.

3.1.1 Pseudo-document-based topic model

Pseudo-document-based Topic Model (PTM) proposed by Zuo et al. [12] is an unsupervised topic model for short texts, which in a way motivates us. PTM is good at modeling the relationship between short texts and pseudo documents. In PTM we have D observed short texts $\{\boldsymbol{d}_s\}_{s=1}^{D}$, we also assume that there are P pseudo documents $\{\boldsymbol{d}_l\}_{l=1}^{P}$ and limit each short text belongs to only one pseudo document. Then the distribution of short texts over pseudo documents is modeled by a multinomial distribution ψ. Next, dealing with the relationship between the pseudo-document and the topics. Assuming that there are K multinomial distributions $\{\phi_z\}_{z=1}^{K}$ about topics over a V-sized vocabulary and the distribution of each pseudo document over topics is a multinomial distribution θ which also indicates the document's class. The plate notation of PTM corresponds to the part in Fig. 1, where the η, λ and response variable y are removed. Taking that into consideration, we have the specifical process of short text's generation as follows:

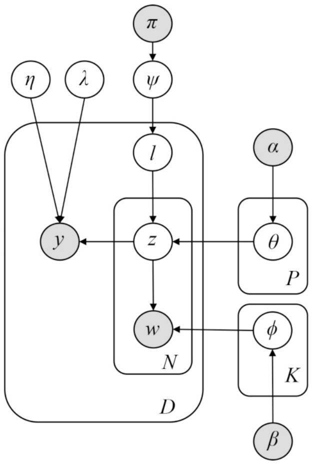

Fig. 1 Graphical representation of PSLDA

1. Sample $\psi - Dir(\pi)$
2. For each topic z:
   (a) Sample $\phi_z - Dir(\beta)$:
3. For each pseudo document d_l':
   (a) Sample $\theta_l \ - \ Dir(\alpha)$:
4. For each short text d_s:
   (a) Sample a pseudo document $l - Multi(\psi)$
   (b) For each word w_i in d_s:
      i. Sample a topic $z - Multi(\theta_l)$
      ii. Sample the ith word $w_i - Multi(\phi_z)$,

where $Dir(\cdot)$ is a Dirichlet distribution the conjuction distribution of multinomial distribution $Multi(\cdot)$, and α, β, and π are hyper-parameters of Dirichlet distribution. The meaning of the symbols are described in Table 1.

Let $W = \{d_s\}_{s=1}^{D}$, $Z = \{z_s\}_{s=1}^{D}$, $\Theta = \{\theta_l\}_{l=1}^{P}$, $L = \{l_s\}_{s=1}^{D}$, $\Phi = \{\phi_z\}_{z=1}^{K}$, and PTM infers the posterior distribution $p(\Theta, Z, L, \Phi \mid W)$.

3.1.2 Expected classifier

In addition to the document content itself, there are related text categories, regression responses and other information. Through making use of these supervision information effectively and expand the original unsupervised model into supervised model, we can learn the theme distribution implied by the document and a classification or regression model at the same time, which can be better used to predict the task. To include the information of labels to the model, we then add an expected classifier to the model, which is an alternative formulation of max-margin supervised topic models, for which we can develop simple and efficient inference algorithms. Because it is difficult to get the exact posterior according to PTM, the expected classifier can calculate the reasonable distribution of topics when the labels are included.

The principle we used to choose the Expected Classifier over a hypothesis space H of classifers is empirical risk minimization (ERM). According to the solution provided by Zhu et al. [34], given training set $D = \{(d_s, y_s)\}_{s=1}^{D}$, we have posterior $q(\eta, \Theta, Z, \Phi, L \mid D)$. It is intractable to have the exact posterior, we approximate by minimizing the training error $R_D(q) = \sum_d \mathbb{I}(\hat{y}_d \neq y_d)$. The discriminant function is defined as

$$F(W) = \mathbb{E}_{q(\eta, z|D)}\left[F(\eta, z; w)\right], F(\eta, z; w) = \eta^{\mathrm{T}}\bar{z}, \tag{1}$$

where $\bar{z} = \dfrac{1}{N}\sum_{n=1}^{N}\mathbb{I}(z_n^k = 1)$, and $\mathbb{I}(\cdot) = 1$ if prediction holds, otherwise $\mathbb{I}(\cdot) = 0$.

As described above, both parts express their respective meanings. The first part of PTM models the relationship between short texts, pseudo documents and topic words, and the second part Expected Classifier provides us a training method for calculating the reasonable distribution of topics when the labels are included. It is clear that the above two parts are strongly coupled via the latent topic assignments Z. And the final model PSLDA we get has both capability of describing the observed short texts and making accurate predictions for the similar texts.

Table 1 Math notations for PSLDA

Notation	Description
D	Number of documents
K	Number of topics
N	Number of words in document d
w	The word in document d
P	The number of pseudo documents
l	The pseudo document
y	The label of the short text
η	Prediction mode
λ	The augmented variables for a higher-dimensional distribution
z	The latent topic assignment
ψ	The multinomial distribution for pseudo document
ϕ_z	The topic-word multinomial for topic z
θ	The document-topic multinomial for document d
α	Hyperparameter of the Dirichlet prior on θ_l
β	Hyperparameter of the Dirichlet prior on ϕ_z
π	Hyperparameter of the Dirichlet prior on ψ

3.2 Inference

For our model, exact posterior inference is hard to achieve. We take a method called “augument-and-collapse” Gibbs sampling [34], which is a fast sampling algorithm and simple to derive. By introducing the data augumentation representation (λ) and integrating out η, ϕ, ψ and π, the collapsed posterior distribution for PSLDA is

$$\begin{aligned} & q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{Z}, \boldsymbol{L}) \\ & \propto p_0(\boldsymbol{\eta}) p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{L} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) \phi(\boldsymbol{y}, \boldsymbol{\lambda} \mid \boldsymbol{Z}, \boldsymbol{\eta}) \\ & = p_0(\boldsymbol{\eta}) \times \prod_{k=1}^{K} \frac{\delta(\boldsymbol{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \times \prod_{p=1}^{P} \frac{\delta(\boldsymbol{C}_p + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \\ & \quad \times \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right), \end{aligned} \tag{2}$$

where $\delta(x) = \dfrac{\prod_{i=1}^{dim(x)} \Gamma(x_i)}{\Gamma(\Sigma_{i=1}^{dim(x)} x_i)}$, C_k^t means a counter about term t being assigned to topic k over the whole data set and $\boldsymbol{C}_k = \left\{C_k^t\right\}_{t=1}^{V}$; and similarly, $\boldsymbol{C}_p = \left\{C_p^k\right\}_{k=1}^{K}$ means topic assignments to pseudo documents. $\zeta_d = \ell - y_d \eta^{\mathrm{T}} \bar{z}_d$, $\ell (\geqslant 1)$ is the cost of making mistakes.

As stated above, in our model we need pseudo document assignment $\boldsymbol{L}$,topic assignment $\boldsymbol{Z}$, prediction mode $\boldsymbol{\eta}$ and the augmented variables $\boldsymbol{\lambda}$. In the next part of this section, we give PSLDA’s inference details.

Sampling prediction model $\boldsymbol{\eta}$:

$$\begin{aligned} q(\boldsymbol{\eta} \mid \boldsymbol{\lambda}, \boldsymbol{Z}, \boldsymbol{L}) & \propto p_0(\eta) \prod_d \exp(-\frac{(\lambda_{d_s} + c\zeta_{d_s})^2}{2\lambda_{d_s}}) \\ & = \exp\left(-\sum_k \frac{\eta_k}{2\nu_2} - \sum_d \frac{(\lambda_{d_s} + c\zeta_{d_s})^2}{2\lambda_{d_s}}\right) \\ & = \mathcal{N}(\eta; \mu, \Sigma), \end{aligned} \tag{3}$$

where we assume its prior is an isotropic Gaussian distribution $p_0(\eta) = \prod_k N(\eta_k; 0, \nu^2)$, the mean is $\mu = \Sigma\left(c \sum_d y_d \dfrac{\lambda_d + c\ell}{\lambda_d} \bar{z}_d\right)$ and the coveriance matrix is $\Sigma = \left(\dfrac{1}{\nu^2} I + c^2 \sum_d \dfrac{\bar{z}\bar{z}_d^{\mathrm{T}}}{\lambda_d}\right)^{-1}$ So The

K-dimensional multivariate Gaussian distribution can be easily done, and the inverse can be finished by using Cholesky decomposition within $O(K^3)$ time.

Sampling the augmented variables $\boldsymbol{\lambda}$:

$$\begin{aligned} q(\boldsymbol{\lambda} \mid \boldsymbol{\eta}, \boldsymbol{Z}, \boldsymbol{L}) &\propto \frac{1}{\sqrt{2\pi\lambda_{d_s}}} \exp\left(-\frac{(\lambda_{d_s} + c\zeta_{d_s})^2}{2\lambda_{d_s}}\right) \\ &\propto \frac{1}{\sqrt{2\pi\lambda_{d_s}}} \exp\left(-\frac{c\zeta_{d_s}^2}{2\lambda_d} - \frac{\lambda_{d_s}}{2}\right) \\ &= \mathcal{GIG}\left(\lambda_{d_s}; \frac{1}{2}, 1, c^2\zeta_{d_s}^2\right), \end{aligned} \quad (4)$$

where $\mathcal{GIG}(x; p, a, b) = C(p, a, b)x^{p-1} \exp\left(-\frac{1}{2}\left(\frac{b}{x} + ax\right)\right)$ is a generalized inverse Gaussian distribution and $C(p, a, b)$ is a normalization constant. Therefore, we can derive that λ_d^{-1} follows an inverse Gaussian distribution

$$p(\lambda_d^{-1} \mid \boldsymbol{Z}, \boldsymbol{\eta}) = \mathcal{IG}\left(\lambda_d^{-1}; \frac{1}{c \mid \zeta_d \mid}, 1\right), \quad (5)$$

where $\mathcal{IG}(x; a, b) = \sqrt{\frac{b}{2\pi x^3}} \exp\left(-\frac{b(x-a)^2}{2a^2 x}\right)$ for $a > 0$ and $b > 0$.

Sampling topic assignment $\boldsymbol{Z}$:

$$\begin{aligned} &q(\boldsymbol{Z} \mid \boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{L}) \\ &\propto \prod_{p=1}^{P} \frac{\delta(\boldsymbol{C}_p + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \times \prod_{k=1}^{K} \frac{\delta(\boldsymbol{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \\ &\times \prod_{d=1}^{D} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right). \end{aligned} \quad (6)$$

By canceling common factors, we can derive the conditional distribution of one variable z_{d_n} given others $\boldsymbol{Z}_\neg$ as:

$$\begin{aligned} &q(z_{dn}^k = 1 \mid \boldsymbol{Z}_\neg, \boldsymbol{\eta}, \boldsymbol{\lambda}, w_{dn} = t, l_d = p) \\ &\propto \frac{C_{k,\neg n}^t + \beta_t}{\sum_t C_{k,\neg n}^t + \sum_{t=1}^V \beta_t} \times \frac{C_{p,\neg n}^k + \alpha_k}{\sum_k C_{p,\neg n}^k + \sum_{k=1}^K \alpha_k} \\ &\times \exp\left(\frac{\gamma y_d (c\ell + \lambda_d)\eta_k}{\lambda_d} - c^2 \frac{\gamma_2 \eta_k^2 + 2\gamma(1-\gamma)\eta_k \Lambda_{d_n}^k}{2\lambda_d}\right), \end{aligned} \quad (7)$$

where $C_{\cdot,\neg n}^{\cdot}$ indicates that term n is excluded from the corresponding document or topic; $\ell = \frac{1}{N_d}$; and $\Lambda_{d_n}^k = \frac{1}{N_d - 1}$ $\sum_{k'} \eta_{k'} C_{k'}^{d,\neg n}$ is the discriminant function value without word n. We can see that the first two terms are from the PTM model for observed word counts and the third term is from the supervised signal y.

Sampling pseudo document assignment $\boldsymbol{L}$:

$$\begin{aligned} &q(\boldsymbol{l}_d = l \mid \boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{Z}, \boldsymbol{L}_\neg) \\ &\propto \frac{C_{l,\neg d} + \pi}{D - 1 + P\pi} \frac{\dfrac{\prod_{z \in d}(C_l^z + \alpha)}{\prod_{z \in \neg d}(C_l^z + \alpha)}}{\prod_{i=1}^{C_d}(C_{l,\neg d} + K\alpha + i - 1)} \\ &= \frac{C_{l,\neg d} + \pi}{D - 1 + P\pi} \frac{\prod_{z \in d} \prod_{j=1}^{C_d^z} (C_{l,\neg d}^z + \alpha + j - 1)}{\prod_{i=1}^{C_d}(C_{l,\neg d} + K\alpha + i - 1)}, \end{aligned} \quad (8)$$

where $C_{l,\neg d}$ is the number of short texts assigned to the lth pseudo document $\boldsymbol{l}_d$. C_d is the length of the dth short text $\boldsymbol{w}_d$, C_d^z counts the number of topic z being assigned to the short text $\boldsymbol{w}_d$, Similarly C_l^z counts the number of topic z being assigned to the pseudo document $\boldsymbol{l}_d$.

After above finished, we get the equations of sampling η,Z,L, and λ and use them to construct a Markov chain with an initial condition. As for sampling from an inverse Gaussian distribution we apply the transformation method with multiple roots [35]. Finally, we initially set $\lambda = 0.1$ and randomly draw Z and L from the uniform distributions. On this basis, we can draw a sample $\hat{\eta}$ as the Gibbs classifier which we will use to make predictions on testing data.

4 Experiment

In order to show the effectiveness of our method on short-text, three real-world short-text corpora are used in our experiments. Statistics of the three datasets can be found in Table 2. A summary of the corpora used is listed below.

DBLP This dataset is mainly composed of the titles of the papers. With a focus on the popular research field, we collect nearly 60,000 conference papers around the following six research directions: *data mining*, *computer vision*, *database*, *information retrieval*, *natural language processing*, and *machine learning*. And we build this corpus by combining the title of the paper and its corresponding category.

NEWS This dataset mainly consists of news. This corpus is provided by Dua, D. and Graff, C. [36]. They collect four categories (i.e., Entertainment, Science and Technology, Business, and Health) of news data from March 10, 2014 to August 10, 2014 among 71425 news webs. They provide over 400000 data, each entry in it contains the following attributes: ID, TITLE, URL, PUBLISHER, CATEGORY, STORY, HOSTNAME, and TIMESTAMP. In our experiment which mainly focuses on short-text, we extract the data from CATEGORY and TITLE excluding stop words and data punctuations in the dataset.

TWEET This corpus is a tweets collection published by Zubiaga et al. [37]. They crawl tweets which contain URL. With the effort of the Open Directory Project (ODP) and the website URLs point to, they can label tweets with the website categories. This dataset contains almost 200,000 entries over ten categories and around 360k labeled tweets. In our experiment, we select nine topic-related categories.

4.1 Classification performance

External tasks such as text classification are often used to evaluate topic models. Our method mainly focuses on short-text classification. In order to evaluate the classification performance we conduct experiments to compare the final latent semantic representations (i.e., topic models) learned by our method and the following baselines.

Table 2 Data statistics

Data set	#Doc.	Voc.size	Avg.doc length
NEWS	422932	36344	7.3
TWEET	182671	21480	8.5
DBLP	59568	7723	6.3

Latent Dirichlet Allocation With SVM (LDA+S-VM). We treat LDA model [3] as a feature extractor. We combine the latent representation of the documents in LDA model and its label as the training dataset for an SVM classifier. For the new documents, we can use the LDA model to get its latent representations, and get the predictions by the above SVM classifer. The LDA algorithm we selected in this paper is GibbsLDA++ [38], and we use a linearSVM for classification. LDA+SVM is a basic method for supervised text classification.

Supervised Latent Dirichlet Allocation (sLDA). In order to use side information (i.e., ratings or labels) attached to the documents in the task of text classification, Blei et al. [39] extended LDA (Latent Dirichlet Allocation) to supervised LDA (sLDA), which is one of the most classical topic models. In this experiment, we use a C++ implementation of variational EM for supervised latent Dirichlet allocation (LDA) written by Wang [40] for comparison.

Maximum Entropy Discrimination Latent Dirichlet Allocation (MedLDA). Zhu et al. [41] proposed the MedLDA model. MedLDA is claimed to perform better in seeking predictive representations of data and more discriminative topic bases for the corpus than other supervised topic models which employ likelihood-driven objective functions for learning and inference. In this paper, we compare our method with MedLDA which is implemented by a variational algorithm.

Pseudo-document-based Topic Model (PTM) and Sparsity-enhanced PTM(SPTM) [12]. PTM is our benchmark method, which introduces the concept of pseudo-documents to implicitly aggregate short texts to prevent data sparseness. By modeling the topic distribution of potential pseudo-objects documents rather than short essays, PTM is expected to benefit excellent performance. A Sparsity-enhanced PTM (SPTM for short) is also proposed by applying Spike and Slab prior, the purpose is to eliminate the bad association between pseudo-documents and potential topics.

Self-aggregate Topic Model (SATM). SATM [11] aggregates short texts into pseudo documents without auxiliary information. However,its parameters grow in number with the size of a short text collection, which makes it prone to overfitting.

In this paper, we use Macro averaged precision, Recall and F-measure to evaluate classification results provided by [42].

For all algorithms, we fix the number of topics to 30. Other parameters in the baseline methods are the same as parameters provided by the author in the code package. In our method, the number of pseudo-document is fixed to 300. In order to ensure fairness and prevent over-fitting, we make the following settings based on five-fold cross-validation.

For each dataset, we first randomly divide it into five equal parts, select one of them in turn, use this part as the test set, leaving the remaining parts as the training set. In this way, we can have five datasets to verify the algorithms, but in order to prevent overfitting caused by randomness, we repeat the above process 5 times. Finally, one corpus can be used to verify each algorithm 25 times, and we believe the results we achieved in this way are reliable.

Table 3 represents the results of classification. It is clear that our method outperforms MedLDA, SLDA, and LDA+SVM on TWEET and NEWS. Also, it is easy to find that LDA+SVM gets almost the worst results comparing to the other algorithm, which means that LDA+SVM is not suitable for tasks such as short-text classification, and we will no longer do some other quantitative analysis on LDA+SVM. The result of our algorithm is very close to the best one, albeit it is not the best with the DBLP dataset.

But such a gap in DBLP caught our attention, considering the meaning of pseudo-document topic model, which means short-text should first be aggregated into pseudo documents, combined with the statistics of corpora, we make an assumption that dataset size has an influence on the topic models.

And we conduct the following experiments to verify the above assumption. First, we select the NEWS dataset, which has a large size. In the case of ensuring the data characteristics, we extract 5%, 10%, 15%, 25%, 50%, and 75% of the data from this dataset to form new datasets. Although SLDA beats our algorithm and MedLDA on DBLP, DBLP is a small dataset and SLDA performs not well on NEWS. Also, SLDA has a high time complexity, since a single experiment can take a week, let alone all the experiments. Based on the above considerations, in this part, we choose MedLDA and our algorithm for comparison. Figure 2 summarizes the results. It can be seen that our assumption is verified. In this table we can find that at the very beginning MedLDA performs better than our algorithm when the size is only 10%, and then the larger the size gets, the higher correctness and stability our algorithm has. We also record the execution time of the algorithms, and from the results given in Fig. 2 it can be seen that our algorithm achieves better performance on big datasets at the cost of time.

In this part, it is demonstrated that our algorithm is superior

Table 3 The precision, recall and F-measure on three datasets

Method	DBLP			NEWS			TWEET		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
PSLDA	$0.6356_{\pm 2.2e\text{-}6}$	$0.6376_{\pm 2.2e\text{-}6}$	$0.6344_{\pm 6.4e\text{-}7}$	$\mathbf{0.8884}^{**}_{\pm 2.1e\text{-}5}$	$\mathbf{0.8824}^{**}_{\pm 1.8e\text{-}5}$	$\mathbf{0.8836}^{**}_{\pm 2.3e\text{-}5}$	$\mathbf{0.7252}^{**}_{\pm 7.6e\text{-}7}$	$\mathbf{0.7092}^{**}_{\pm 7.4e\text{-}6}$	$\mathbf{0.7136}^{**}_{\pm 2.2e\text{-}6}$
MedLDA	$0.6664_{\pm 1.1e\text{-}2}$	$0.658_{\pm 9.9e\text{-}3}$	$0.6572_{\pm 1.1e\text{-}2}$	$0.7344_{\pm 5.3e\text{-}4}$	$0.8008_{\pm 1.6e\text{-}4}$	$0.7616_{\pm 3.0e\text{-}4}$	$0.544_{\pm 6.9e\text{-}3}$	$0.5548_{\pm 5.7e\text{-}3}$	$0.5456_{\pm 6.4e\text{-}3}$
SLDA	$0.652_{\pm 6.9e\text{-}5}$	$0.6567_{\pm 4.9e\text{-}5}$	$0.6567_{\pm 7.3e\text{-}5}$	$0.853_{\pm 8.3e\text{-}5}$	$0.854_{\pm 8.6e\text{-}5}$	$0.854_{\pm 8.6e\text{-}5}$	$0.63_{\pm 1.2e\text{-}4}$	$0.643_{\pm 8.9e\text{-}5}$	$0.6313_{\pm 1.3e\text{-}4}$
LDA+SVM	$0.604_{\pm 1.2e\text{-}4}$	$0.59_{\pm 2.2e\text{-}4}$	$0.594_{\pm 3.5e\text{-}4}$	$0.7384_{\pm 7.9e\text{-}5}$	$0.7348_{\pm 3.9e\text{-}4}$	$0.7356_{\pm 5.7e\text{-}5}$	$0.5536_{\pm 6.8e\text{-}3}$	$0.5168_{\pm 7.5e\text{-}4}$	$0.5292_{\pm 7.3e\text{-}4}$
SPTM	$0.661_{\pm 2.4e\text{-}4}$	$0.667_{\pm 7.6e\text{-}4}$	$0.663_{\pm 7.2e\text{-}3}$	$0.760_{\pm 1.8e\text{-}3}$	$0.761_{\pm 7.9e\text{-}5}$	$0.759_{\pm 3.1e\text{-}3}$	$0.551_{\pm 6.2e\text{-}4}$	$0.558_{\pm 4.1e\text{-}3}$	$0.550_{\pm 1.1e\text{-}2}$
SATM	$0.657_{\pm 6.3e\text{-}5}$	$0.662_{\pm 5.1e\text{-}6}$	$0.654_{\pm 4.3e\text{-}3}$	$0.697_{\pm 7.6e\text{-}4}$	$0.702_{\pm 8.1e\text{-}3}$	$0.686_{\pm 8.4e\text{-}3}$	$0.599_{\pm 6.0e\text{-}4}$	$0.605_{\pm 7.1e\text{-}5}$	$0.594_{\pm 2.5e\text{-}4}$
PTM	$\mathbf{0.667}^{**}_{\pm 4.1e\text{-}3}$	$\mathbf{0.672}^{**}_{\pm 2.9e\text{-}2}$	$\mathbf{0.668}^{**}_{\pm 6.1e\text{-}4}$	$0.755_{\pm 5.7e\text{-}3}$	$0.757_{\pm 5.4e\text{-}4}$	$0.754_{\pm 1.5e\text{-}2}$	$0.561_{\pm 8.1e\text{-}4}$	$0.568_{\pm 7.6e\text{-}3}$	$0.559_{\pm 4.7e\text{-}3}$

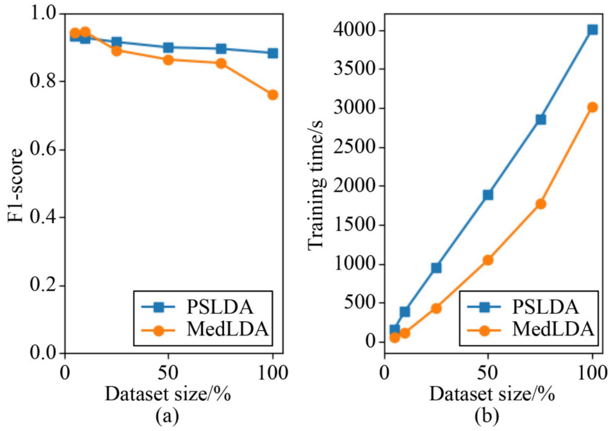

Fig. 2 Effect of dataset size on (a) F1-score and (b) training time

to the classic (i.e., SLDA) or mainstream (i.e., MedLDA) supervised topic model, and it is more obvious on large datasets.

4.2 Topic coherence

In the previous section, we have proved that our algorithm outperforms when working on tasks such as text classification, especially when the dataset size is very large. In this part, we will dive deep into the topic models generated above, and compare the interpretability of those topics.

Many researchers have proved that better perplexity does not indicate interpretability of topics, which means perplexity is not a good metric for the evaluation of topic models although it was wildly used previously. In order to address this problem, many topic coherence measures have been proposed. According to Roder et al. [43], topic coherence measures performs better than *perplexity* in topic evaluation. They also compare 237,912 coherence measures and show that CV measure is the best coherence measure. Based on the above considerations, we employ CV coherence based on Wikipedia corpus, focus on the top 10 topic words, to evaluate topics.

Table 4 Topic coherence of different models on three datasets

	DBLP	TWEET	NEWS
PSLDA	$\mathbf{0.4274}^{**}_{\pm 7.0e-3}$	$\mathbf{0.4322}^{**}_{\pm 7.4e-3}$	$\mathbf{0.4667}^{**}_{\pm 1.6e-3}$
MedLDA	$0.4185_{\pm 5.7e-3}$	$0.4087_{\pm 5.6e-3}$	$0.4153_{\pm 1.7e-3}$
SLDA	$0.3767_{\pm 6.9e-5}$	$0.3710_{\pm 4.9e-5}$	$0.4070_{\pm 6.7e-3}$
LDA	$0.3930_{\pm 1.4e-4}$	$0.4003_{\pm 5.1e-2}$	$0.4161_{\pm 5.6e-3}$

Table 4 shows the results of the evaluation. We see that for all datasets our algorithm scores higher than the baseline methods. And due to the small amount of data in DBLP, MedLDA is not able to generate 30 complete topics, many of which have fewer than 10 topic words, so we have abandoned them when calculating CV coherence. The result means that the topic model derived by our method has better interpretability and therefore proves that using pseudo-document technology to reorganize short-text datasets is an effective method, either dealing with text classification issues or generating interpretable topics.

4.3 Validation of model parameter sensitivity

Parameters of our model include K, P, D, and N (i.e., topic numbers, pseudo documents number, $\Delta\ell=1/D$ and number of iterations). In this part, we fix K to 30 and take P = 300, N = 200, $\Delta\ell$ = 0.1 as the standard, adjust the range of P, N, D in turn, to explore the influence of the change of parameters on the results of our model. We can see in Fig. 3, with the number of pseudo-documents increasing, the stability of the overall document of the algorithm is improved. However, it can be seen that the accuracy of the algorithm decreases when the number of pseudo-documents increases to a certain number. The parameter $\Delta\ell$ has a great influence on the algorithm, and the stability of the algorithm will decline as the value of this parameter increase. At last, our algorithm converges very fast and only needs a few iterations to complete the learning.

4.4 Qualitative analysis of topic detection

In order to show and compare the effectiveness of topics detected by our model and other models, we conduct a

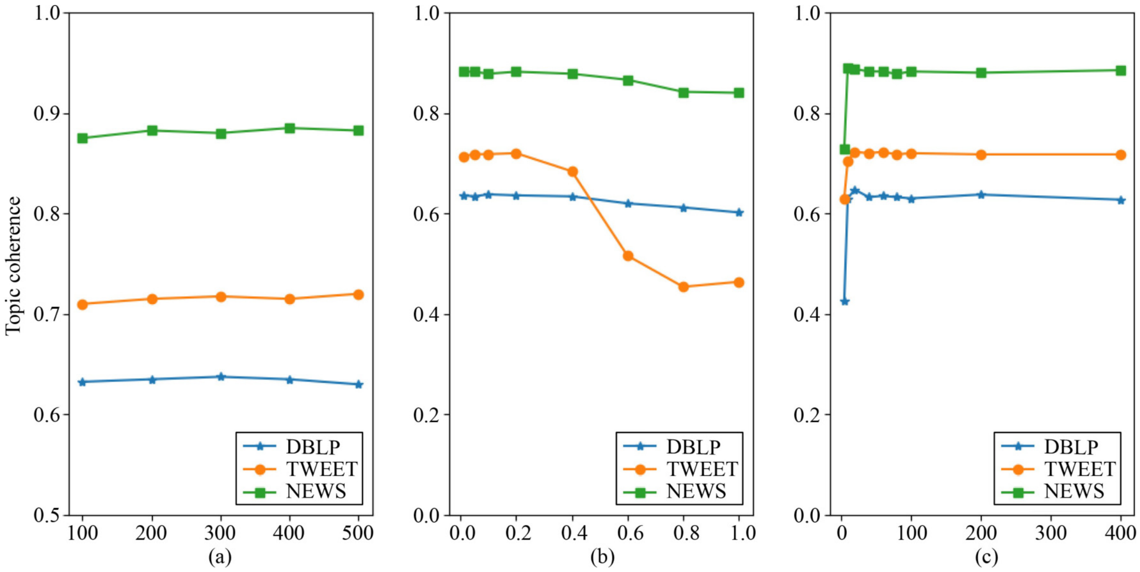

Fig. 3 (a) Impact of number of pseudo documents; (b) impact of $\Delta\ell$; (c) convergence with number of iterations

Table 5 Top 10 words of 2 topics detected by SLDA, MedLDA and our model from three datasets

Dataset	Topic	Model	Topic words	#irrelevant
DBLP	Topic 1	SLDA	Image, segmentation, motion, images, visual, face, scene, **invariant**, **flow**, **depth**	3
		MedLDA	Image, based, recognition, learning, object, detection, 3d, tracking, segmentation, motion	0
		PSLDA	Image, based, recognition, learning, object, detection, 3d, tracking, motion, segmentation	0
	Topic 2	SLDA	Models, language, machine, translation, **statistical**, **natural**, phrase, grammars, sentence, japanese	2
		MedLDA	Based, translation, language, learning, machine, **parsing**, word, semantic, model, models	1
		PSLDA	Based, word, translation, model, machine, dependency, analysis, learning, classification, supervised	0
TWEET	Topic 1	SLDA	Apple, ipad, android, app, **way**, phone, **design**, windows, apps, store	2
		MedLDA	**Yelp**, **checked**, video, engadget, apple, android, ipad, google, **just**, **review**	4
		PSLDA	Android, engadget, app, windows, video, nokia, phone, google, apps, **beta**	1
	Topic 2	SLDA	Amazon, usa, **set**, **star**, **baby**, **complete**, gift, picture, drive, easy	4
		MedLDA	Amazon, kindle, usa, black, newly, **tagged**, edition, book, battery, bonanza	1
		PSLDA	Amazon, kindle, usa, black, newly, **tagged**, edition, battery, book, laptop	1
NEWS	Topic 1	SLDA	Samsung, microsoft, **one**, galaxi, iphone, S5, smartphone, **unveil**, ipad, xbox	2
		MedLDA	Samsung, galaxi, **one**, **new**, S5, buy, price, apple, facebook, **T**	3
		PSLDA	Google, apple, new, **change**, facebook, samsung, microsoft, user, galaxi, **climat**	2
	Topic 2	SLDA	**On**, **die**, record, Jay-Z, new, age, **beyond**, **second**, **break**, **attack**	6
		MedLDA	New, star, video, kardashian, kim, **To**, season, **A**, **In**, show	3
		PSLDA	Kardashian, kim, kany, west, **wed**, bieber, justin, Jay-Z, star, selena	1

qualitative analysis of detected topic words. We choose two topics and show the top 10 words of each topic detected by SLDA, MedLDA, and our model from three datasets in Table 5.

We ask volunteers to manually mark irrelevant words in each model and topic (bold words in the table). The result shows that:

1) SLDA performs worst and detects most irrelevant topic words, while the meaning of a topic can be easily inferred by reading the topic words detected by our model. For example, Topic 2 of DBLP means “machine translation”, Topic 1 of TWEET means “mobile phone” and Topic 2 of NEWS means “singer”.

2) From the topic words of Topic 2 of NEWS we can see that our model detected many singer names, such as “Jay-Z”, “Selena”, and “Justin Bieber”, which are highly relevant to the topic; while the topic words detected by MedLDA mainly consist of some general words, such as “star”, “video”, and “show”. As a result, the topics detected by our model are more interpretable and effective.

5 Conclusion

Short texts are difficult to model because of their sparseness. In this paper, we proposed the PSLDA model for short texts, by leveraging much fewer pseudo documents to self aggregate tremendous short texts, PSLDA learned topic distributions without using auxiliary contextual information. Then we presented the inference of our model and conducted various experiments to validate the effectiveness. In the future, we will adopt our model for more domain-specific tasks.

References

1. Rosso P, Errecalde M, Pinto D. Analysis of short texts on the web: introduction to special issue. Language Resources and Evaluation, 2013, 47(1): 123–126
2. Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999, 50–57
3. Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993–1022
4. Li Z, Zhang H, Wang S, Huang F, Li Z, Zhou J. Exploit latent Dirichlet allocation for collaborative filtering. Frontiers of Computer Science, 2018, 12(3): 571–581
5. Chen W, Cai F, Chen H, De Rijke M. Personalized query suggestion diversification in information retrieval. Frontiers of Computer Science, 2020, 14(3): 143602
6. Miyazawa S, Song X, Xia T, Shibasaki R, Kaneda H. Integrating GPS trajectory and topics from twitter stream for human mobility estimation. Frontiers of Computer Science, 2019, 13(3): 460–470
7. Hong L, Davison B D. Empirical study of topic modeling in twitter. In: Proceedings of the 1st Workshop on Social Media Analytics. 2010, 80–88
8. Davison B D, Suel T, Craswell N, Liu B. WSDM’10: Third ACM International Conference on Web Search and Data Mining. New York: ACM, 2010
9. Mehrotra R, Sanner S, Buntine W, Xie L. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2013, 889–892
10. Phan X H, Nguyen C T, Le D T, Nguyen L M, Horiguchi S, Ha Q T. A hidden topic-based framework toward building applications with short Web documents. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 961–976
11. Quan X, Kit C, Ge Y, Pan S J. Short and sparse text topic modeling via self-aggregation. In: Proceedings of the 24th International Conference on Artificial Intelligence. 2015, 2270–2276
12. Zuo Y, Wu J, Zhang H, Lin H, Wang F, Xu K, Xiong H. Topic modeling of short texts: a pseudo-document view. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 2105–2114
13. Blei D M, Lafferty J D. Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. 2006, 113–120
14. Meek C, Chickering M, Halpern J. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff: AUAI Press, 2004
15. Nguyen D Q, Billingsley R, Du L, Johnson M. Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 2015, 3: 299–313
16. Zhao F, Zhu Y, Jin H, Yang L T. A personalized hashtag recommendation approach using lda-based topic model in microblog environment. Future Generation Computer Systems, 2016, 65: 196–206
17. Ibeke E, Lin C, Wyner A, Barawi M H. Extracting and understanding contrastive opinion through topic relevant sentences. In: Proceedings of

the 8th International Joint Conference on Natural Language Processing. 2017, 395–400

18. Tian C, Rong W, Zhou S, Zhang J, Ouyang Y, Xiong Z. Learning word representation by jointly using neighbor and syntactic contexts. Neurocomputing, 2021, 456: 136–146
19. Weng J, Lim E P, Jiang J, He Q. TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. 2010, 261–270
20. Jin O, Liu N N, Zhao K, Yu Y, Yang Q. Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011, 775–784
21. Lin T, Tian W, Mei Q, Cheng H. The dual-sparse topic model: mining focused topics and focused terms in short text. In: Proceedings of the 23rd International Conference on World Wide Web. 2014, 539–550
22. Cheng X, Yan X, Lan Y, Guo J. BTM: topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2928–2941
23. Zuo Y, Zhao J, Xu K. Word network topic model: a simple but general solution for short and imbalanced texts. Knowledge and Information Systems, 2016, 48(2): 379–398
24. Yin J, Wang J. A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014, 233–242
25. Li C, Wang H, Zhang Z, Sun A, Ma Z. Topic modeling for short texts with auxiliary word embeddings. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016, 165–174
26. Li X, Li C, Chi J, Ouyang J. Short text topic modeling by exploring original documents. Knowledge and Information Systems, 2018, 56(2): 443–462
27. Bicalho P, Pita M, Pedrosa G, Lacerda A, Pappa G L. A general framework to expand short text for topic modeling. Information Sciences, 2017, 393: 66–81
28. Pedrosa G, Pita M, Bicalho P, Lacerda A, Pappa G L. Topic modeling for short texts with co-occurrence frequency-based expansion. In: Proceedings of the 5th Brazilian Conference on Intelligent Systems (BRACIS). 2016, 277–282
29. Shi T, Kang K, Choo J, Reddy C K. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: Proceedings of 2018 World Wide Web Conference. 2018, 1105–1114
30. Miao Y, Yu L, Blunsom P. Neural variational inference for text processing. In: Proceedings of the 33rd International Conference on Machine Learning. 2016, 1727–1736
31. Ding R, Nallapati R, Xiang B. Coherence-aware neural topic modeling. In: Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. 2018, 830–836
32. Zhu J, Xing E P. Sparse topical coding. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence. 2011, 831–838
33. Card D, Tan C, Smith N A. Neural models for documents with metadata. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018, 2031–2040
34. Zhu J, Chen N, Perkins H, Zhang B. Gibbs max-margin topic models with data augmentation. The Journal of Machine Learning Research, 2014, 15(1): 1073–1110
35. Michael J R, Schucany W R, Haas R W. Generating random variates using transformations with multiple roots. The American Statistician, 1976, 30(2): 88–90
36. Dua D, Graff C. UCI machine learning repository. See archiveics.uci.edu/ml/index website, 2017
37. Zubiaga A, Ji H. Harnessing web page directories for large-scale classification of tweets. In: Proceedings of the 22nd International Conference on World Wide Web. 2013, 225–226
38. Phan X H, Nguyen C T. GibbsLDA++: A C/C++ implementation of latent dirichlet allocation (LDA). Boston: Free Software Foundation, 2007
39. Blei D M, McAuliffe J D. Supervised topic models. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. 2007, 121–128
40. Chong W, Blei D, Li F F. Simultaneous image classification and annotation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2009, 1903–1910
41. Zhu J, Ahmed A, Xing E P. MedLDA: maximum margin supervised topic models. The Journal of Machine Learning Research, 2012, 13(1): 2237–2278
42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: machine learning in Python. The Journal of Machine Learning Research, 2011, 12: 2825–2830
43. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 2015, 399–408

Mingtao Sun is a PhD candidate in School of Economics and Management, Beihang University, China. His research interests include Big Data processing and Education Administration.

Xiaowei Zhao is currently pursuing the PhD degree in Computer Science with Beihang University, China. Her main research interests include transfer learning and sentiment analysis.

Jingjing Lin is currently a senior student at the School of Instrumentation and Optoelectronic Engineering, Beihang University, China. Her research interests include text classification, natural language inference, and sentiment analysis.

Jian Jing received the MS degree in the Engineering of Computer Techonlogy from the Beihang University, China in 2021. His research interests include knowledge reasoning, algorithms and big data processing.

Deqing Wang received the PhD degree in computer science from Beihang University, China in 2013. He is currently an Associate Professor with the School of Computer Science and the Deputy Chief Engineer with the National Engineering Research Center for Science Technology Resources Sharing and Service, Beihang University, China. His research focuses on text categorization and data mining for software engineering and machine learning.

Guozhu Jia received the PhD degree from Aalborg University, Denmark. He is currently a Professor of School of Economics and Management, Beihang University, China and a member of Expert Committee of China Manufacturing Servitization Alliance. He is also a director of China Innovation Method Society.