Information Sciences 607 (2022) 79–91

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Context reinforced neural topic modeling over short texts

Jiachun Feng [a], Zusheng Zhang [a], Cheng Ding [a], Yanghui Rao [a,*], Haoran Xie [b], Fu Lee Wang [c]

[a] *School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China*
[b] *Department of Computing and Decision Sciences, Lingnan University, Hong Kong*
[c] *School of Science and Technology, Hong Kong Metropolitan University, Hong Kong*

ARTICLE INFO

Article history:
Received 5 November 2020
Received in revised form 20 October 2021
Accepted 27 May 2022
Available online 01 June 2022

Keywords:
Neural topic model
Short texts
Context reinforcement

ABSTRACT

As one of the prevalent topic mining methods, neural topic modeling has attracted a lot of interests due to the advantages of low training costs and strong generalisation abilities. However, the existing neural topic models may suffer from the feature sparsity problem when applied to short texts, due to the lack of context in each message. To alleviate this issue, we propose a Context Reinforced Neural Topic Model (CRNTM), whose characteristics can be summarized as follows. First, by assuming that each short text covers only a few salient topics, the proposed CRNTM infers the topic for each word in a narrow range. Second, our model exploits pre-trained word embeddings by treating topics as multivariate Gaussian distributions or Gaussian mixture distributions in the embedding space. Extensive experiments on two benchmark short corpora validate the effectiveness of the proposed model on both topic discovery and text classification.

1. Introduction

Topic modeling is an unsupervised learning method for discovering the latent semantic structure (i.e., topics) of texts, which can be applied to a variety of tasks in natural language processing, e.g., text classification, sentiment analysis, and recommender systems. Generally, the parameters of a topic model can be estimated by variational inference [3] or Gibbs sampling [13], both of which, however, require model-specific derivation when there is any change to the model structure. Moreover, with the growth of data scale, the generative process of a document is getting tricky and expensive, which may lead to high training costs. These limitations make it difficult for researchers to extend the topic models to new variations flexibly.

With the development of deep learning, variational auto-encoder (VAE) [17] has been developed as a promising solution for topic modeling. Benefiting from the flexibility of neural networks, VAE is competent to learn complicated non-linear distributions and is convenient to be applied to various tasks. Furthermore, by using backpropagation for tuning parameters, VAE is highly efficient in training when compared with the models based on variational inference or Gibbs sampling. Considering the above advantages, several topic models built on VAE have been proposed, such as neural variational document model (NVDM) [28], neural variation latent Dirichlet allocation (NVLDA) [34], Gaussian softmax model (GSM) [27], and neural variational correlated topic modeling (NVCTM) [23].

Although the existing VAE-based models can reduce the training cost impressively, they still suffer from the feature sparsity problem when applied to short texts. In this case, the number of words in each message is relatively small, while the

* Corresponding author.

https://doi.org/10.1016/j.ins.2022.05.098
0020-0255/© 2022 Elsevier Inc. All rights reserved.

vocabulary is large and the range of topics is broad. To alleviate the above issue, several approaches have been proposed, such as Graph-based inference network for the biterm topic model (GraphBTM) [40] and neural sparsemax topic model (NSMTM) [21]. However, due to significant word non-overlap in short texts, the relatedness information between word pairs may not be fully captured by these models. Therefore, learning context information from a short message is still challenging for them.

In this paper, we propose a VAE-based topic model for short texts, where the context information for each text is learned effectively. First, as can be observed, a short text generally covers only a small set of topics due to the limited text length. Therefore, we propose to filter the irrelevant topics by setting a *topic controller* for each topic, so that a short text can be summarized with only a few salient topics. Through this way, the range of topic inference is narrowed down and the context of each short message is indirectly enhanced, which can help alleviate the feature sparsity issue to some extent. Second, considering that word embeddings are effective in capturing the semantic relatedness between words [5], we propose to incorporate them into our model to explicitly enrich the context information for each short message. Specifically, we model each topic by a multivariate Gaussian distribution or a Gaussian mixture distribution in the embedding space, through which the relatedness of synonymous word pairs can be effectively inferred regardless of the lack of word-overlap. In this way, our model can discover topics that are more meaningful and interpretable. We name the proposed model as Context Reinforced Neural Topic Model (CRNTM) and conclude the main contributions of our work as follows:

- With the assumption that a short message only focuses on a few salient topics, we propose to filter the irrelevant ones by setting a *topic controller* for each topic, through which the range of topic inference can be narrowed down and the feature sparsity problem can be indirectly alleviated.
- Pre-trained word embeddings are incorporated into our model to enrich the limited context information for each short message. By treating topic distributions over words as multivariate Gaussian distributions or Gaussian mixture distributions in the embedding space, the feature sparsity issue can be explicitly dealt with and our model can produce more interpretable topics.

The rest of this paper is organized as follows. We discuss related research work in Section 2, and describe our proposed model in Section 3. Experimental settings and results are presented in Section 4, and the conclusions are drawn in Section 5.

2. Related work

2.1. Neural topic modeling

Many real-world tasks rely critically on solving the optimal value of a complex objective function [1,25,15], in which, neural networks have been considered as promising methods [19,29,2] owing to their strong abilities to fit non-linear distributions. For topic modeling, effective models built on neural networks have been proposed to discover latent topics from texts, and most of them are based on VAE. In this vein, NVDM [28] is a neural variational inference framework for generative and conditional models on text. It consists of an encoder (inference network) and a softmax decoder (generative model). The inference network, implemented by multilayer perceptrons (MLP), is used to compress the bag-of-words document representation into a continuous latent variable, from which the generative model is utilized to reconstruct the document by generating words independently. Based on NVDM, GSM [27] is proposed to construct the topic distribution explicitly with a softmax function applied to the projection of the Gaussian random vector. However, NVDM and GSM share a same drawback that the topic distribution is assumed to be an isotropic Gaussian, which makes them incapable of capturing the topic correlations. Therefore, NVCTM [23] is developed to model the relationships among topics by reshaping topic distributions. While NVDM, GSM, and NVCTM apply the Gaussian distribution as their prior, NVLDA [34], ProdLDA [34], and DVAE [4] employ the Dirichlet prior. Specifically, NVLDA constructs a Laplace approximation to the Dirichlet distribution, and ProdLDA, as the improved version of NVLDA, replaces the mixture assumption in latent Dirichlet allocation (LDA) [3] with a weighted product of experts. For DVAE, it decouples the properties of sparsity and smoothness by rewriting the Dirichlet parameter vector into a product of a sparse binary vector and a smoothness vector.

2.2. Topic discovery over short texts

Topic models provide a valuable solution to latent semantic mining and understanding over texts. Nonetheless, they usually suffer from the feature sparsity problem when applied to short texts [39], because the word co-occurrences are generally lacking in the text with limited length.

To enrich the contextual information of short texts, several studies introduce external documents as auxiliary resources [33,16,32]. However, since the introduced documents are required to be semantically close to the original corpus, this way might be stiff and inflexible. Some approaches try to alleviate the feature sparsity issue by firstly aggregating short texts into lengthy pseudo-documents and then applying a well established topic model. For this category of methods, short texts can be aggregated by utilizing the side information, e.g., user characteristics tags [12], user ID [39], and timestamp [7]. Other kinds of methods try to enrich the word co-occurrences by modifying the prior of Bayesian models. For instance, the Biterm

Topic Model (BTM) [6], which models the word co-occurrences at the corpus level, is able to lengthen short texts indirectly by converting documents into biterm sets.

While all the above models are built on Bayesian analysis, models based on neural networks have also been introduced for modeling short texts. Zhu et al. [40] proposed a graph-based inference network named GraphBTM to accelerate BTM. This model samples a fixed number of texts as training instances to deal with the feature sparsity problem. Lin et al. [21] developed a neural model named NSMTM by providing sparse posterior distributions over topics based on the Gaussian sparsemax construction. In NSMTM, the variational distribution is inferred with the relax Wasserstein divergence (RW), rather than the Kullback–Leibler (KL) divergence as in the common VAE-based models. Gupta et al. [14] developed a neural autoregressive topic model named iDocNADE in a language modeling fashion. They also incorporated word embeddings as fixed priors into the model to introduce complementary information. However, their approach is unable to model topic distributions explicitly.

In summary, the previous models built on VAE, including NVDM, NVCTM, ProdLDA, and DVAE, have been proven as effective to mine topics from texts. However, these models may suffer from the feature sparsity problem when applied to short texts. To alleviate this issue, Bayesian models that focus on context enhancement have been introduced and proven to be effective for modeling texts with limited lengths. Nonetheless, they are difficult to be improved and extended since most of them are built on variational inference or Gibbs sampling. Although neural network-based models, such as GraphBTM, NSMTM, and iDocNADE, were developed to enhance the flexibility, learning context information from short texts is still challenging for them due to the lack of word co-occurrences. Therefore, in this paper, we propose our VAE-based CRNTM to deal with the feature sparsity problem and obtain a good model scalability.

3. Model description

In this section, we describe our context reinforced neural topic model (CRNTM) in details. The overall architecture of CRNTM is illustrated in Fig. 1, which consists of two modules: an inference network for learning latent topics and a Gaussian decoder for reconstructing documents.

3.1. Problem definition

Given a corpus with D short texts, we denote the corresponding vocabulary as $W = \{w_1, w_2, \dots, w_V\}$, with V being the vocabulary size. Following [28], each document is processed into a bag-of-words (BOW) vector, i.e., $x_d = [x_{d,1}, x_{d,2}, \dots, x_{d,V}]$, where $x_{d,i}$ represents the number of times for word w_i appearing in document d.

In the inference network, we use $\theta_d \in \mathbb{R}^K$ to denote the topic distribution of document d and use $z_k \in \{z_1, z_2, \dots, z_K\}$ to represent the topic assignment for an observed word, where K is the number of topics inherent in the given corpus. We set a *topic controller* $\lambda_{d,k} \in [0,1]$ for each topic z_k in document d: the topic will be kept when $\lambda_{d,k} = 1$, or it will be filtered out when $\lambda_{d,k} = 0$. The *topic controller* $\lambda_d = \{\lambda_{d,1}, \lambda_{d,2}, \dots, \lambda_{d,K}\}$ can be drawn from the Beta distribution $\mathscr{B}(\alpha_d, \beta_d)$, where α_d and β_d are the corresponding parameters.

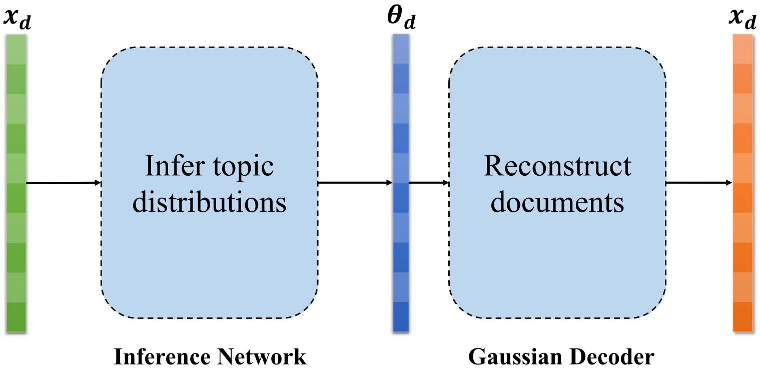


Fig. 1. Structure of our CRNTM.

For the Gaussian decoder, we denote the word embedding matrix corresponding to the vocabulary as $WE \in \mathbb{R}^{V \times r}$, where r denotes the dimension of word embeddings. In specific, the embedding of word w_i is represented as WE_i. Furthermore, we use $TW \in \mathbb{R}^{K \times V}$ to denote the topic-word matrix, in which $TW_{(k,i)}$ represents the conditional probability of word w_i over topic z_k. In this study, $TW_{(k,i)}$ is drawn from a multivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$ or a Gaussian mixture distribution, where $\mu_k \in \mathbb{R}^r$ and $\Sigma_k \in \mathbb{R}^{r \times r}$ are learnable parameters.

3.2. Inference network

The inference network is applied to learn the topic distribution of a given document. The structure of the inference network is illustrated in Fig. 2, from which we can see that it consists of two parts: one for learning a temporal distribution θ'_d, and the other for learning a topic controller λ_d. The final document-topic distribution θ_d can be obtained by merging θ'_d and λ_d. In the following, we will detail these two parts.

Based on the structure of VAE, our CRNTM infers the parameters μ_d and Σ_d via layers of networks that are elaborately designed for the observed data. Being fed with the input document x_d, the inference network firstly outputs an encoded vector π_d. Then, π_d is linearly transformed to obtain μ_d and Σ_d, both of which are used to parameterize the Gaussian distribution $\mathcal{N}(\mu_d, \Sigma_d)$. The above process can be described by:

$$\pi_d = \text{MLP}_1(x_d), \tag{1}$$
$$\mu_d = l_1(\pi_d), \ \log\sigma_d = l_2(\pi_d), \tag{2}$$
$$\Sigma_d = \text{diag}(\sigma_d^2), \tag{3}$$

where MLP_1 denotes a multilayer perceptron. $l_1(\cdot)$ and $l_2(\cdot)$ are linear transformations. Note that the diagonal elements σ_d^2 of the covariance matrix Σ_d are non-negative, which can be ensured by $\log\sigma_d$. Now we can obtain the temporal topic distribution θ'_d by the following processes:

$$\epsilon_d \sim \mathcal{N}(0, I^2), \tag{4}$$
$$h_d = \mu_d + \epsilon_d \ * \sigma_d, \tag{5}$$
$$\theta'_d = \text{softmax}(W_\theta \cdot h_d + b_\theta), \tag{6}$$

where h_d is drawn from the Gaussian distribution $\mathcal{N}(\mu_d, \Sigma_d)$ with reparameterization, allowing the parameters to be optimized by backpropagation.

A short message generally contains a few words, which may result in the feature sparsity problem during the inference process. However, it can be observed that a short message usually focuses on a small set of topics, and this inspires us to alleviate the aforementioned issue by narrowing down the scope of topic inference. Instead of allowing the topic mixtures navigate freely in the simplex, our CRNTM forces each short message to cover a narrow range of topics, which is achieved by setting a *topic controller* $\lambda_{d,k} \in [0,1]$ for each topic. Topic z_k is kept when $\lambda_{d,k} = 1$, and it will be filtered out when $\lambda_{d,k} = 0$. The topic controllers are drawn from the Beta distribution, i.e., $\lambda_d \sim \mathscr{B}(\alpha_d, \beta_d)$, which ensures that each component $\lambda_{d,k} \in \lambda_d$ is in the range of $[0,1]$. The parameters α_d and β_d are inferred as follows:

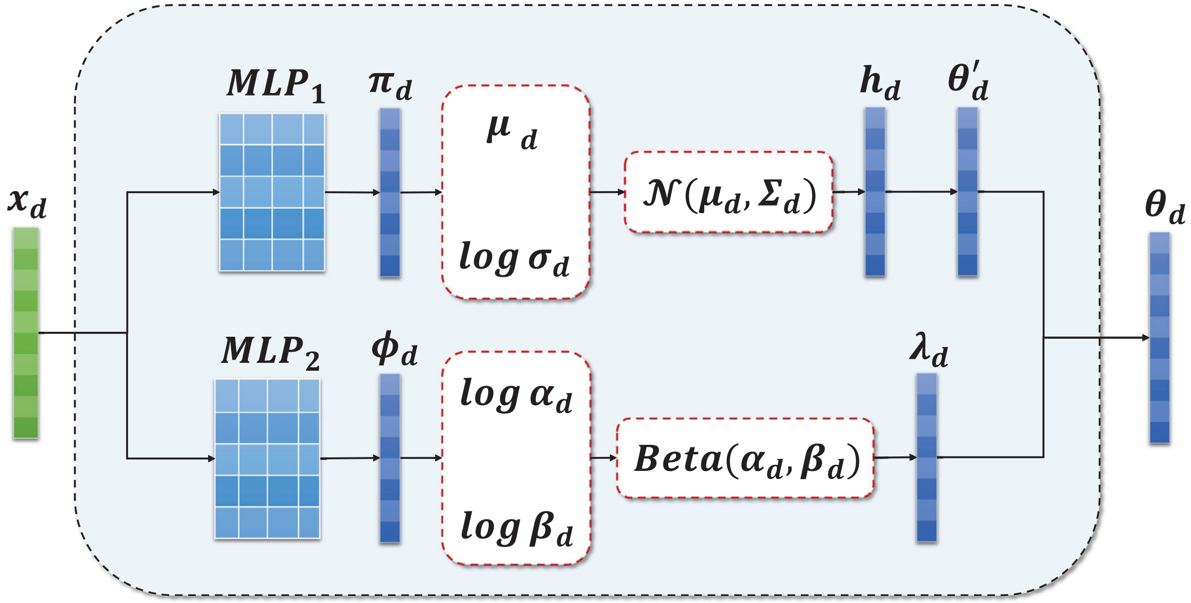


Fig. 2. Inference network.

$$\phi_d = \mathrm{MLP}_2(x_d), \tag{7}$$

$$\log \alpha_d = l_3(\phi_d),\ \log \beta_d = l_4(\phi_d), \tag{8}$$

where MLP_2 denotes a multilayer perceptron. $l_3(\cdot)$ and $l_4(\cdot)$ are linear transformations.

Since the Beta sampling can not be differentiated directly, it is intractable to update model parameters through backpropagation. Therefore, we use the reparameterization technique to obtain λ_d by following [30]. Note that the Beta samples can be generated using two Gamma variables, and the corresponding process can be formulated as follows:

$$\lambda_{\alpha_d} \sim \mathscr{G}(\alpha_d, 1),\ \lambda_{\beta_d} \sim \mathscr{G}(\beta_d, 1), \tag{9}$$

$$\lambda_d = \frac{\lambda_{\alpha_d}}{\lambda_{\alpha_d} + \lambda_{\beta_d}}. \tag{10}$$

For the Gamma distribution $\mathscr{G}(\alpha_d, 1)$ with $\alpha_d > 1$, the reparameterization can be accomplished by the reject sampling method, as follows:

$$\lambda_{\alpha_d} = (\alpha_d - \frac{1}{3})\left(1 + \frac{\epsilon_d}{\sqrt{9\alpha_d - 3}}\right)^3, \tag{11}$$

where $\epsilon_d \sim \mathscr{N}(0, I^2)$. In addition, we apply the shape augmentation method to convert $\alpha_d \leqslant 1$ to $\alpha_d > 1$. This can be formulated by $\lambda_{\alpha_d} = \rho^{\frac{1}{\alpha_d}} \tilde{\lambda}_{\alpha_d}$, where ρ is drawn from a uniform distribution, i.e., $\rho \sim U[0, 1]$, and $\tilde{\lambda}_{\alpha_d} \sim \mathscr{G}(\alpha_d + 1, 1)$ is obtained according to Eq. (11).

During the inference process, our CRNTM determines whether topic z_k is kept according to $\lambda_{d,k}$. By filtering out some irrelevant topics, the short texts are allowed to focus on a few salient topics, and thus the feature sparsity problem can be indirectly alleviated. Finally, the topic distribution of document x_d can be obtained by:

$$\theta_d = \theta'_d * \lambda_d. \tag{12}$$

3.3. Gaussian decoder

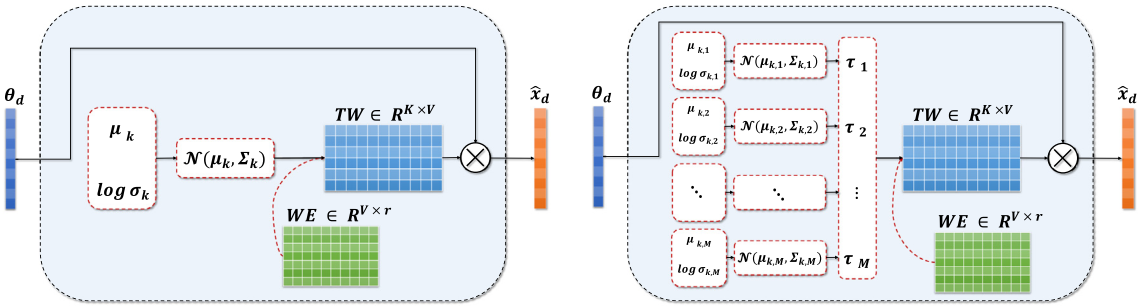
By setting a controller for each topic in the inference network, we can indirectly enhance the context of each short message and hence alleviate the feature sparsity issue to a certain extent. To further enrich the context information, we propose to introduce additional semantic knowledge for each word into the decoder via the corresponding embedding. Concretely, the decoder employs the multivariate Gaussian distribution $\mathscr{N}(\mu_k, \Sigma_k)$ to model the k-th topic in the embedding space, as shown in Fig. 3a. By incorporating the pre-trained word embeddings, the probability of word w_i conditioned to topic z_k can be calculated by:

$$TW_{(k,i)} = \frac{\exp\left(g(WE_i)\right)}{(2\pi)^{r/2}|\Sigma_k|^{1/2}}, \tag{13}$$

where

$$g(WE_i) = -\frac{1}{2}(WE_i - \mu_k)^T \Sigma_k^{-1}(WE_i - \mu_k), \tag{14}$$

and r is the dimension of word embeddings. It is worth noting that the parameters μ_k and Σ_k can be respectively regarded as the topic centroid and the topic concentration in the embedding space. According to the properties of Gaussian distribution,



(a) Gaussian distribution　(b) Gaussian mixture distribution

Fig. 3. Gaussian decoder.

the probabilities are higher for the words that are closer to the topic centroid, which indicates that they are more likely to reflect the same topic. From another point of view, the distance between these words is relatively small and thereby their semantic meanings are fairly close, which makes them appropriate to be used to summarize a certain topic. Therefore, our CRNTM is able to enrich the context information via the pre-trained word embeddings to further alleviate the feature sparsity problem. Finally, we can estimate the conditional probability of the i-th word in document d, i.e., $w_{d,i}$, by the following equation:

$$p(w_{d,i}|\theta_d,\lambda_d)=\sum_k \theta_{d,k}\cdot TW_{(k,i)}. \tag{15}$$

The above method can be easily adjusted by extending the Gaussian distribution to the Gaussian mixture distribution, as shown in Fig. 3b. In this case, we have

$$TW_{(k,i)}=\sum_{m=1}^{M}\tau_m \frac{\exp\left(g_m(WE_i)\right)}{(2\pi)^{r/2}|\Sigma_{k,m}|^{1/2}}, \tag{16}$$

where

$$g_m(WE_i)=-\frac{1}{2}(WE_i-\mu_{k,m})^T\Sigma_{k,m}^{-1}(WE_i-\mu_{k,m}), \tag{17}$$

with M being the number of the Gaussian components and τ_m being the corresponding coefficient.

3.4. Optimization objective

The optimization objective of CRNTM is to maximize the evidence lower bound $\mathscr{L}(d)$, which can be derived according to the variational inference method as follows:

$$\begin{aligned}\mathscr{L}(d)=\;&\iint q(\theta_d,\lambda_d|x_d)[-\log q(\theta_d,\lambda_d|x_d)+\log p(x_d,\theta_d,\lambda_d)]\mathrm{d}\theta_d\mathrm{d}\lambda_d\\ =\;&E_{q(\theta_d|x_d)q(\lambda_d|x_d)}[\log p(x_d|\theta_d,\lambda_d)]-D_{KL}[q(\theta_d|x_d)\|p(\theta_d)]\\ &-D_{KL}[q(\lambda_d|x_d)\|p(\lambda_d)].\end{aligned} \tag{18}$$

In the above equation, the first term is often regarded as the reconstruction loss, where

$$p(x_d|\theta_d,\lambda_d)=\prod_{i=1}^{n_d}p(w_{d,i}|\theta_d,\lambda_d). \tag{19}$$

For the second term $D_{KL}[q(\theta_d|x_d)\|p(\theta_d)], q(\theta_d|x_d)$ denotes the Gaussian distribution $\mathscr{N}(\mu_d,\Sigma_d)$ which is the variational approximation of the true posterior, and $p(\theta_d)$ is assumed to be a normal Gaussian prior $\mathscr{N}(0,I)$ by following [17,28,23]. Therefore, this term can be written as:

$$D_{KL}[q(\theta_d|x_d)\|p(\theta_d)]=\frac{1}{2}(-n+\mu_d^2-\log|\Sigma_d|+|\Sigma_d|). \tag{20}$$

Similarly, for the third term $D_{KL}[q(\lambda_d|x_d)\|p(\lambda_d)], q(\lambda_d|x_d)$ denotes the Beta distribution $\mathscr{B}(\alpha_d,\beta_d)$, and $p(\lambda_d)$ denotes a Beta prior which can be formulated as $\mathscr{B}(\alpha',\beta')$. Accordingly, we can write this term as follows:

$$\begin{aligned}D_{KL}[q(\lambda_d|x_d)\|p(\lambda_d)]=\;&\ln\tfrac{\Delta(\alpha',\beta')}{\Delta(\alpha_d,\beta_d)}-(\alpha'-\alpha_d)\psi(\alpha_d)-(\beta'-\beta_d)\psi(\beta_d)\\ &+(\alpha'-\alpha_d+\beta'-\beta_d)\psi(\alpha_d+\beta_d),\end{aligned} \tag{21}$$

where

$$\Delta(\alpha',\beta')=\frac{\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha'+\beta')},$$

with $\Gamma(\cdot)$ being the Gamma function and $\psi(\cdot)$ being the Digamma function. In our model, the *topic controller* λ_d acts as a switch to filter out irrelevant topics and keep salient topics with higher probabilities. Therefore, we preset the Beta prior as $\mathscr{B}(\alpha'=0.5,\beta'=0.5)$, from which we can draw samples that are approximate to 0 or 1.

4. Experiments

In this section, we firstly introduce the experimental setting, and then evaluate the effectiveness of our model by a series of experiments.

4.1. Datasets

To evaluate the model performance on both topic mining and text classification, we employ 20NewsGroups[1] and Snippets[2] with document labels as our datasets. 20NewsGroups is a collection of short news messages, which officially falls into a training set and a testing set with $11,314$ and $7,531$ samples. These short texts are grouped into 20 different categories. Snippets is collected from the results of web search transaction over 8 domain labels. The officially divided 10,060 and 2,280 search transaction documents are used for training and testing, respectively. For data preprocessing, we remove stop words and take the most frequent $2,000$ words and $5,000$ words as vocabularies. The statistics of the processed corpora are shown in Table 1, where *AvgD* and *L* denote the averaged number of words for each document and the number of categories, respectively.

4.2. Baseline methods

We use the following mainstream VAE-based methods as baselines for evaluation: NVDM [28], NVLDA & ProdLDA [34], GSM [27], TMN [37], NVCTM [23], and DVAE [4]. Among these methods, NVDM is one of the first neural document models. Besides, NVLDA, ProdLDA, and GSM are classical neural topic models. TMN consists of a neural topic model and a topic memory mechanism, which are trained in an end-to-end learning manner. NVCTM exploits the Centralized Transformation Flow (CTF) to capture the topic correlations by reshaping topic distributions. DVAE achieves a competitive topic coherence and a high log-likelihood by decoupling the properties of sparsity and smoothness in VAE-based topic models for short texts. Note that iDocNADE [14] and NSMTM [21] are not adopted for comparison, because the former does not model topic distributions explicitly while the training process of the latter is too sensitive to continue based on our implementation. Besides, since the ELBO is typically used and necessary to evaluate the performance of VAE based methods [28,27], we do not use Bayesian models such as [35,22,20] as baselines for fair comparison. Finally, GraphBTM [40] which only models a mini-corpus is unsuitable to be evaluated in this study.

4.3. Experimental settings

For the baselines of NVDM,[3] NVLDA & ProdLDA,[4] TMN,[5] and DVAE,[6] we directly use the publicly available codes. For GSM and NVCTM, we implement them based on the code of NVDM, where the length of CTF in NVCTM is set to 10 according to the preliminary experiments. We implement our models by TensorFlow[7] and denote the one applying Gaussian distribution as CRNTM_GD and the one applying Gaussian mixture distribution as CRNTM_GMD. The proposed CRNTM_GD and CRNTM_GMD both employ the 300-dimensional GloVe embeddings [31]. Moreover, the size of each hidden layer in our models is 500, and unless explicitly specified, the number of Gaussian components in CRNTM_GMD is set to 25. The source code, detailed parameter settings, and complementary results of our models can be found at GitHub.[8] We run the above models on one i7-7700 CPU, in which, it costs approximately 8.8 min and 57 min per 10 epochs to train CRNTM_GD and CRNTM_GMD on 20NewsGroups. All the baselines and our models are trained by the Adam optimizer, whose learning rate is $5e^{-5}$, with the batch size setting to 64. Table 2 shows the parameter setting details of our two models.

In the task of topic discovery, we use perplexity to evaluate the generalisation performance of models on the testing set. Perplexity is computed by $\exp\{-\frac{1}{D}\sum_{d=1}^{D}\frac{1}{N_d}\sum_{i=1}^{N_d}\log p(w_{d,i})\}$, where D is the number of documents, N_d is the number of words in document d, and $\log p(w_{d,i})$ is the log-likelihood of word w_i in document d. To evaluate the quality of discovered topics, we also use the normalized pointwise mutual information (NPMI) [18] as the metric. For a given model, its final NPMI is obtained by averaging the ones of all topics, while each topic is represented by 5, 10, and 15 words respectively. In the task of text classification, we use the document-topic distribution generated by each convergent model as the input of an MLP classifier,[9] in which, accuracy is used as the metric. For each task, the topic numbers are set to $25, 50$, and 100, respectively.

4.4. Comparison with baselines

Table 3 presents the perplexity results obtained by different models from the testing samples. We can observe that CRNTM_GD achieves competitive results when compared with all the baselines, which validates its effectiveness in fitting unknown data. We can further observe that CRNTM_GMD outperforms its basic version and obtains the best performance in most cases, owing to its stronger decoder than the one of CRNTM_GD. In addition, it can be found that TMN performs the best on the Snippets dataset, because it's basically a supervised model for text classification and the supervision from

[1] http://www.qwone.com/jason/20Newsgroups/20news-18828.tar.gz.
[2] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz.
[3] https://github.com/ysmiao/nvdm.
[4] https://github.com/akashgit/autoencoding_vi_for_topic_models.
[5] https://github.com/zengjichuan/TMN.
[6] https://github.com/sophieburkhardt/dirichlet-vae-topic-models.
[7] https://github.com/tensorflow/tensorflow.
[8] https://github.com/Deloris-NLP/CRNTM.
[9] https://scikit-learn.org/stable/modules/classes.html.

Table 1
The statistics of datasets.

Dataset	Train	Test	V	$AvgD$	L
20NewsGroups	11,314	7,531	2,000	12.3	20
Snippets	10,060	2,280	5,000	14.3	8

Table 2
Parameter setting details of CRNTM_GD and CRNTM_GMD.

Parameter	CRNTM_GD	CRNTM_GMD
Embedding Size	300	300
Hidden Size	500	500
Number of Gaussian Components	-	25
Learning Rate	5e-5	5e-5
Batch Size	64	64

the labels can help mining topics from extremely short texts. Table 4 reports the results of topic coherence which are calculated according to the training sets. Similarly, we can see that both of our models achieve excellent performance and CRNTM_GMD performs the best on the two datasets. It is mainly due to the topic controller employed by the inference network and the Gaussian decoder embedded with word vectors. These two mechanisms collectively help discover meaningful and interpretable topics from short texts. Table 5 shows the accuracy results for text classification. We can find that our models achieve the best results on 20NewsGroups. For the Snippets dataset, they may perform slightly inferior to NVCTM, yet the results obtained by them are still competitive when compared with the rest of baselines. To sum up, our proposed models have strong generalisation abilities. They are able to effectively mine topics from short texts and to learn appropriate text representations for the classification task.

4.5. Evaluation on Gaussian decoder via topic visualization

To investigate the quality of topics discovered by our models, we report top 15 words of 4 representative topics and visualize these topics by their embedding vectors using 20NewsGroups. Particularly, we extract μ_k of Gaussian distributions as

Table 3
Perplexity results of different models on both datasets, where the best scores are boldfaced.

Model	20NewsGroups			Snippets		
	25	50	100	25	50	100
NVDM	802	855	871	5144	5180	5328
NVLDA	1046	1252	1153	5336	5496	5374
ProdLDA	1106	1073	1035	5312	5379	5348
GSM	949	922	943	5237	5295	5434
TMN	1159	1136	1128	**3177**	**3197**	**3236**
NVCTM	758	738	744	5090	5121	5136
DVAE	1095	1066	1075	5090	5121	5136
CRNTM_GD	698	706	680	4822	4872	4861
CRNTM_GMD	**574**	**586**	**590**	4608	4695	4602

Table 4
Topic coherence results of different models on both datasets, where the best scores are boldfaced.

Model	20NewsGroups			Snippets		
	25	50	100	25	50	100
NVDM	0.041	0.061	0.053	0.068	0.067	0.069
NVLDA	0.065	0.062	0.061	0.042	0.045	0.041
ProdLDA	0.064	0.062	0.065	0.046	0.051	0.045
GSM	0.080	0.076	0.065	0.068	0.061	0.065
TMN	0.031	0.051	0.042	0.043	0.025	0.029
NVCTM	0.022	0.017	0.014	0.052	0.051	0.055
DVAE	0.065	0.075	0.069	0.039	0.052	0.040
CRNTM_GD	0.065	0.077	0.069	0.075	0.076	0.074
CRNTM_GMD	**0.088**	**0.081**	**0.079**	**0.082**	**0.084**	**0.085**

Table 5
Classification accuracies of different models on both datasets, where the best scores are boldfaced.

Model	20NewsGroups			Snippets		
	25	50	100	25	50	100
NVDM	0.64	0.64	0.67	0.15	0.17	0.16
NVLDA	0.40	0.45	0.42	0.12	0.13	0.13
ProdLDA	0.43	0.44	0.40	0.14	0.14	0.15
GSM	0.45	0.46	0.45	0.11	0.12	0.11
TMN	0.40	0.48	0.51	0.15	0.16	0.13
NVCTM	0.64	0.64	0.65	**0.16**	**0.18**	**0.18**
DVAE	0.32	0.37	0.34	0.08	0.09	0.06
CRNTM_GD	0.64	**0.65**	**0.68**	0.15	0.16	0.14
CRNTM_GMD	**0.69**	**0.65**	0.66	**0.16**	0.16	0.17

the topic centroid and utilize t-SNE [26] for visualization. Topic visualization of the results in CRNTM_GD is depicted in Fig. 4a. The points with different colors indicate different topics, and the centroid of topic k is denoted as Tk. For the convenience of comparison, we manually annotate each topic by referring to the ground truth category. Accordingly, $T1, T2, T3$, and $T4$ in CRNTM_GD are annotated as "soc.religion.christian", "talk.politics", "comp.sys.ibm.pc.hardware", and "comp.-graphics", respectively. We can observe that all top words of the same topics are close to each other and to the corresponding topic centroids in the continuous vector space. This validates that our Gaussian decoder can effectively capture the context information via word embeddings in mining topics. We also present 4 topics generated by CRNTM_GMD whose semantics are similar to those in CRNTM_GD to verify the effectiveness of Gaussian mixture distributions. Topic visualization of the results in CRNTM_GMD is shown in Fig. 4b. For clarity, the coefficients of Gaussian components are indicated by different point sizes and shades of colour. The bigger the points are and the stronger the color is, the higher coefficient of the corresponding Gaussian components is. The topic centroids and their probabilities are detailed in Fig. 4c. We can observe that for topic $T3$ named as "comp.graphics", the main components such as $T3:5, T3:3$, and $T3:0$ are close to the cluster of red points, while $T2:8$ and $T2:5$ are close to sub-clusters of top words in $T2$.

To make a comprehensive comparison, we present the results of all models on generating topic "soc.religion.christian" in Table 6. It can be observed that our models can discover quite meaningful topics, i.e., the topic words are coherent and they are all relevant to the given topic. Nonetheless, the topics learned by our models are not all of good qualities. For example, in a certain learned topic, the representative words are "keys", "door", "copy", "file", and "windows", respectively, where "door" is obviously not a topic word since all the rest words are related to computer. The reason of making such a mistake by our model may be concluded as follows: "door" is close to "keys" and "windows" in the embedding space and hence they will be grouped into the same topic by our Gaussian decoder, but they are not related to each other due to the polysemy existed in the latter two words.

4.6. Ablation study and hyperparameter analysis

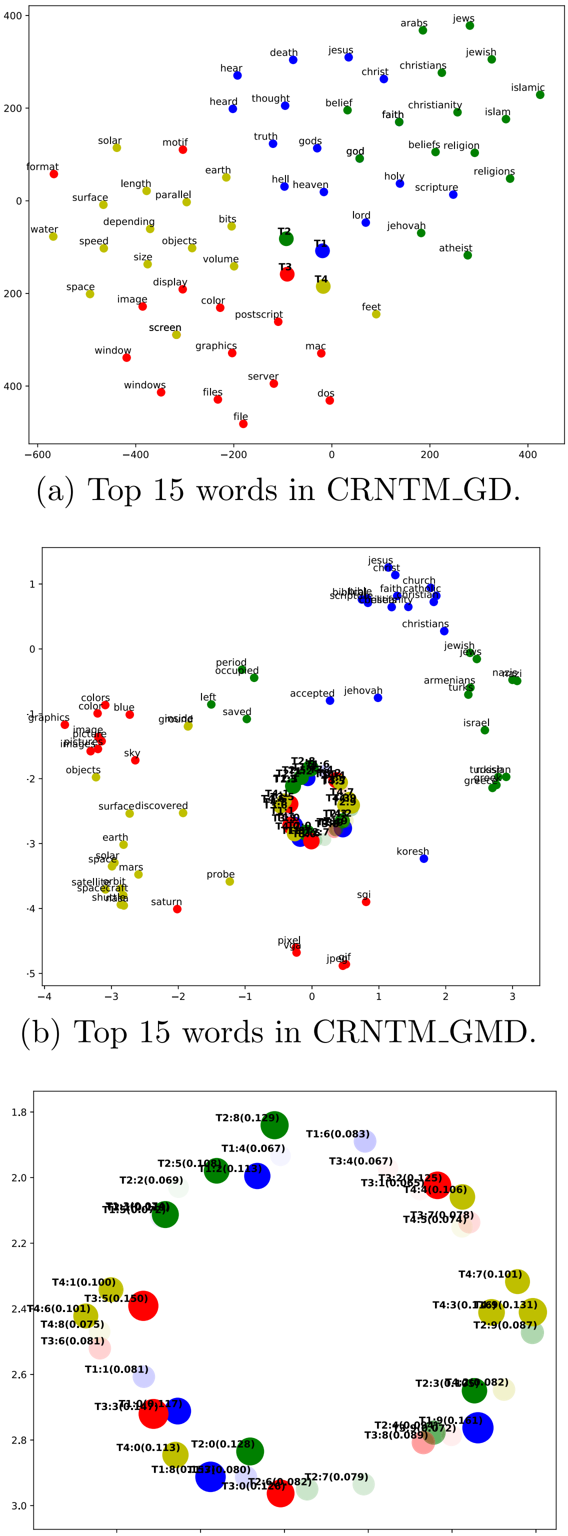
We conduct an ablation study to fully explore the impact of two key components in CRNTM_GMD, i.e., the *topic controller* and the Gaussian decoder. There are two main variants: (i) CRNTM-w/o-Ctrl which removes the topic controller, and (ii) CRNTM-w/o-GMD which removes the Gaussian decoder. Table 7 reports the results of CRNTM_GMD[10] and its variants. We can observe that both CRNTM-w/o-Ctrl and CRNTM-w/o-GMD perform worse than CRNTM_GMD, which demonstrates the effectiveness of our two key components. We also observe that CRNTM-w/o-GMD performs the worst on both datasets, owing to the lack of relatedness information provided by the pre-trained word vectors embedded in the Gaussian decoder.

We further study the impact of the number of Gaussian components on CRNTM_GMD. Table 8 presents the results of CRNTM_GMD on 20NewsGroups when varying the component number under 25 topics. We can observe that CRNTM_GMD with more Gaussian components generally performs better than that with less ones, which demonstrates that a more sophisticated mixture possesses a stronger capacity of learning high quality topics. The best topic coherence and classification accuracy are obtained when the component number is set to 25, and a larger value may not further boost the model performance.

5. Conclusion

In this paper, we propose a VAE-based topic model, i.e., CRNTM, to discover latent topics from short texts. The proposed CRNTM mainly alleviates the feature sparsity problem of short text mining by two ways. First, we introduce a *topic controller* into the inference network, which can help narrow down the range of topic inference and indirectly enhance the context of each message. Second, we incorporate pre-trained word embeddings into the decoder to explicitly enrich the context information, by modeling each topic through a multivariate Gaussian distribution or a Gaussian mixture distribution in the

[10] We do not consider CRNTM_GD for a straightforward description.



(a) Top 15 words in CRNTM_GD.

(b) Top 15 words in CRNTM_GMD.

(c) Topic centroids and their probabilities in CRNTM_GMD.

Fig. 4. Characteristics of 4 representative topics generated by our models on 20NewsGroups. $Tk : m$ represents the mth centroid of topic k.

Table 6
Top 10 words of manually labeled topic “soc.religion.christian” from all models on 20NewsGroups, where irrelevant words are underlined.

Model	Top words
NVDM	god sin scsi bible jesus rutgers homosexuality christian ide christians
NVLDA	god scsi sin drive jesus bible christian christians homosexuality love
ProdLDA	god christians jesus bible doctrine interpretation belief homosexuality christianity eternal
GSM	god jesus bible christ church people christian believe christians sin
TMN	sin myers eternal president mary god heaven christ doctor jobs
NVCTM	church catholic christians magnus scripture duke andrew turkey sex christianity
DVAE	jesus scripture christ bible doctrine sin christians god canon homosexuality
CRNTM_GD	jesus god christh eaven death holy truth gods faith lord
CRNTM_GMD	god christians bible christ jesus sin religion church lord doctrine

Table 7
Performance comparison in topic coherence between CRNTM_GMD and its variants.

Model	20NewsGroups	Snippets
CRNTM-w/o-Ctrl	0.035	0.033
CRNTM-w/o-GMD	0.030	0.021
CRNTM_GMD	0.088	0.082

Table 8
Performance of CRNTM_GMD with different Gaussian mixture numbers on 20NewsGroups, where M denotes the number of Gaussian components. Note that the best results are boldfaced.

M	Perplexity	Coherence	Accuracy
5	634	0.060	0.65
10	616	0.071	0.66
15	597	0.081	0.68
20	588	0.084	0.68
25	574	**0.088**	**0.69**
30	574	0.080	0.66
35	**571**	0.081	0.66

embedding space. We conduct extensive experiments on two benchmark datasets, which demonstrate that CRNTM is able to effectively deal with the feature sparsity issue and to mine interpretable topics from short texts. Specifically, our model is 24.27%, 25.37%, and 7.81% superior than the best baseline in perplexity, topic coherence, and classification accuracy, respectively. Apart from topic discovery over short texts, the proposed method could also shed light on addressing other challenging tasks with limited historical data [38] or sparse implicit information [36]. In our future work, we plan to introduce a feature-enhanced network [9,10,8,11] into our model, so as to fully explore the implicit features inherent in short texts. For enhancing the context of each text better, we also plan to take the information of the nearest and the shared neighbors [24] into account.

CRediT authorship contribution statement

Jiachun Feng: Methodology, Software, Writing – original draft. **Zusheng Zhang:** Software, Validation, Writing – original draft, Writing – review & editing. **Cheng Ding:** Software, Validation, Writing – review & editing. **Yanghui Rao:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Haoran Xie:** Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research described in this paper was supported by the National Natural Science Foundation of China (61972426), Guangdong Basic and Applied Basic Research Foundation (2020A1515010536), the Direct Grant (DR22A2) and the Faculty Research Grants (DB22B4 and DB22B7) of Lingnan University, Hong Kong, and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E01/19).

References

[1] G. Aghajani, N. Ghadimi, Multi-objective energy management in a micro-grid, Energy Rep. 4 (2018) 218–225.

[2] P. Akbary, M. Ghiasi, M.R.R. Pourkheranjani, H. Alipour, N. Ghadimi, Extracting appropriate nodal marginal prices for all types of committed reserve, Comput. Econ. 53 (2019) 1–26.

[3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[4] S. Burkhardt, S. Kramer, Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model, J. Mach. Learn. Res. 20 (2019) 1–27.

[5] X. Chang, Y.L. Yu, Y. Yang, E.P. Xing, Semantic pooling for complex event analysis in untrimmed videos, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1617–1632.

[6] X. Cheng, X. Yan, Y. Lan, J. Guo, BTM: topic modeling over short texts, IEEE Trans. Knowl. Data Eng. 26 (2014) 2928–2941.

[7] Q. Diao, J. Jiang, F. Zhu, E. Lim, Finding bursty topics from microblogs, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012, pp. 536–544.

[8] M. Duan, K. Li, K. Li, Q. Tian, A novel multi-task tensor correlation neural network for facial attribute prediction, ACM Trans. Intell. Syst. Technol. 12 (2020) 3:1–3:22.

[9] M. Duan, K. Li, X. Liao, K. Li, Q. Tian, Features-enhanced multi-attribute estimation with convolutional tensor correlation fusion network, ACM Trans. Multimedia Comput. Commun. Appl. 15 (2019) 91:1–91:23.

[10] M. Duan, K. Li, A. Ouyang, K.N. Win, K. Li, Q. Tian, Egroupnet: A feature-enhanced network for age estimation with novel age group schemes, ACM Trans. Multimedia Comput. Commun. Appl. 16 (2020) 42:1–42:23.

[11] M. Duan, A. Ouyang, G. Tan, Q. Tian, Age estimation using aging/rejuvenation features with device-edge synergy, IEEE Trans. Circuits Syst. Video Technol. 31 (2021) 608–620.

[12] J. Feng, Y. Rao, H. Xie, F.L. Wang, Q. Li, User group based emotion detection and topic discovery over short text, World Wide Web 23 (2020) 1553–1587.

[13] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 5228–5235.

[14] P. Gupta, Y. Chaudhary, F. Buettner, H. Schütze, Document informed neural autoregressive topic models with distributional prior, in: Proceedings of the 33rd Conference on Artificial Intelligence, 2019, pp. 6505–6512.

[15] M. Hamian, A. Darvishan, M. Hosseinzadeh, M.J. Lariche, N. Ghadimi, A. Nouri, A framework to expedite joint energy-reserve payment cost minimization using a custom-designed method based on mixed integer genetic algorithm, Eng. Appl. Artif. Intell. 72 (2018) 203–212.

[16] O. Jin, N.N. Liu, K. Zhao, Y. Yu, Q. Yang, Transferring topical knowledge from auxiliary long texts for short text clustering, in: Proceedings of the 20th Conference on Information and Knowledge Management, 2011, pp. 775–784.

[17] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: Proceedings of the 2nd International Conference on Learning Representations, 2014.

[18] J.H. Lau, D. Newman, T. Baldwin, Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 530–539.

[19] H. Leng, X. Li, J. Zhu, H. Tang, Z. Zhang, N. Ghadimi, A new wind power prediction method based on ridgelet transforms, hybrid feature selection and closed-loop forecasting, Adv. Eng. Inform. 36 (2018) 20–30.

[20] C. Li, H. Wang, Z. Zhang, A. Sun, Z. Ma, 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, pp. 165–174.

[21] T. Lin, Z. Hu, X. Guo, Sparsemax and relaxed wasserstein for topic sparsity, in: Proceedings of the 20th ACM International Conference on Web Search and Data Mining, 2019, pp. 141–149.

[22] T. Lin, W. Tian, Q. Mei, H. Cheng, The dual-sparse topic model: mining focused topics and focused terms in short text, in: Proceedings of the 23rd International World Wide Web Conference, 2014, pp. 539–550.

[23] L. Liu, H. Huang, Y. Gao, Y. Zhang, X. Wei, Neural variational correlated topic modeling, in: Proceedings of the 28th International World Wide Web Conference, 2019, pp. 1142–1152.

[24] R. Liu, H. Wang, X. Yu, Shared-nearest-neighbor-based clustering by fast search and find of density peaks, Inf. Sci. 450 (2018) 200–226.

[25] Y. Liu, W. Wang, N. Ghadimi, Electricity load forecasting by an improved forecast engine for building level consumers, Energy 139 (2017) 18–30.

[26] L. van der Maaten, Learning a parametric embedding by preserving local structure, in: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, 2009, pp. 384–391.

[27] Y. Miao, E. Grefenstette, P. Blunsom, Discovering discrete latent topics with neural variational inference, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 2410–2419.

[28] Y. Miao, L. Yu, P. Blunsom, Neural variational inference for text processing, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 1727–1736.

[29] F. Mirzapour, M. Lakzaei, G. Varamini, M. Teimourian, N. Ghadimi, A new prediction model of battery and wind-solar output in hybrid power system, J. Ambient Intell. Humanized Comput. 10 (2019) 77–87.

[30] C.A. Naesseth, W. Scott, F.J.R.R. Linderman, D.M. Blei, Reparameterization gradients through acceptance-rejection sampling algorithms, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017, pp. 489–498.

[31] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.

[32] X.H. Phan, M.L. Nguyen, S. Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 91–100.

[33] M. Sahami, T.D. Heilman, A web-based kernel function for measuring the similarity of short text snippets, in: Proceedings of the 15th International Conference on World Wide Web, 2006, pp. 377–386.

[34] A. Srivastava, C.A. Sutton, Autoencoding variational inference for topic models, in: Proceedings of the 5th International Conference on Learning Representations, 2017.

[35] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts, in: Proceedings of the 22nd International World Wide Web Conference, 2013, pp. 1445–1456.

[36] B. Yi, X. Shen, H. Liu, Z. Zhang, W. Zhang, S. Liu, N. Xiong, Deep matrix factorization with implicit feedback embedding for recommendation system, IEEE Trans. Industr. Inf. 15 (2019) 4591–4601.

[37] J. Zeng, J. Li, Y. Song, C. Gao, M.R. Lyu, I. King, Topic memory networks for short text classification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3120–3131.

[38] Q. Zhang, C. Zhou, Y.C. Tian, N. Xiong, Y. Qin, B. Hu, A fuzzy probability bayesian network approach for dynamic cybersecurity risk assessment in industrial control systems, IEEE Trans. Industr. Inf. 14 (2018) 2497–2506.

[39] W.X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: Proceedings of the 33rd European Conference on IR Research, 2011, pp. 338–349.

[40] Q. Zhu, Z. Feng, X. Li, Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4663–4672.