Research Article

A Method for Constructing Supervised Time Topic Model Based on Variational Autoencoder

Zhinan Gou (ID),[1] **Yan Li** (ID),[2] **and Zheng Huo**[1]

[1]*School of Information Technology, Hebei University of Economics and Business, Shijiazhuang, Hebei 050061, China*
[2]*School of Management Science and Engineering, Hebei University of Economics and Business, Shijiazhuang,
  Hebei 050061, China*

Correspondence should be addressed to Zhinan Gou; gouzhinan@heuet.edu.cn

Topic modeling is a probabilistic generation model to find the representative topic of a document and has been successfully applied to various document-related tasks in recent years. Especially in the supervised topic model and time topic model, many methods have achieved some success. The supervised topic model can learn topics from documents annotated with multiple labels and the time topic model can learn topics that evolve over time in a sequentially organized corpus. However, there are some documents with multiple labels and time-stamped in reality, which need to construct a supervised time topic model to achieve document-related tasks. There are few research papers on the supervised time topic model. To solve this problem, we propose a method for constructing a supervised time topic model. By analysing the generative process of the supervised topic model and time topic model, respectively, we introduce the construction process of the supervised time topic model based on variational autoencoder in detail and conduct preliminary experiments. Experimental results demonstrate that the supervised time topic model outperforms several state-of-the-art topic models.

1. Introduction

Nowadays many kinds of text information like news, blogs, books, and social network accompany the daily lives of people. Some traditional methods are difficult to get effective information from an increasingly large amount of data. The probabilistic topic model is a new technology that can help people to organize, index, search, and browse these kinds of large data automatically.

Latent Dirichlet allocation (LDA) [1] is a classical topic model to find the representative topic of a document. During the past decade, LDA has been successfully applied to various document-related tasks, such as text classification [2], clustering [3], and summarization [4]. However, as a static topic model, LDA has two limitations.

Firstly, the number of topics is hard to determine in the static topic model. To select the best number of topics, most of the methods use the lowest perplexity or the largest likelihood estimate value of the topic model by comparing

different numbers of topics [1, 5, 6]. In fact, these methods lead to the fact that the latent topic content is difficult to interpret. Secondly, the text information can be exchanged in the static topic model [1]. Moreover, the static topic model assumes that the text information is unordered with each other [7]. This simplified assumption is inappropriate and unrealistic [8].

In order to overcome these limitations, the supervised topic model and time topic model are proposed, respectively. In recent years, methods of the supervised topic model and time topic model have a variety of applications.

The objective of the supervised topic model is to learn topics from documents annotated with multiple labels. To the best of our knowledge, labeled LDA [9] is the classical supervised topic model that matches the multiple topics to the labels in the document. The number of topics is determined by the metadata of the document (such as labels), and topic terms have a better way to interpret topics [10]. In addition, the supervised topic model supports a variety of

applications, such as social event analysis [11], abnormal event detection [12], document classification [13], and tag recommendation [14].

The objective of the time topic model is to construct a topic model that evolves over time in a sequentially organized corpus. To the best of our knowledge, the dynamic topic model is the first time topic model that captures the evolution of topics in a sequentially organized corpus of documents [7]. Based on the time topic model, a series of applications are researched, for example, dynamic features extraction [15], automated behaviour analysis [16], travel recommendation [17], and tracking urban geotopics [18].

However, there are some documents with multiple labels and time-stamped in reality, for example, a scientific paper has keywords and is time-stamped. Park et al. proposed a similar supervised time topic model, the main idea of which is on generating the numerical time-series variables as supervised metadata that is a single topic of each time slice [19]. This model is not considered as a supervised time topic model which can learn a topic from documents with multiple labels and which are time-stamped and apply the topic to various document-related tasks. As a result, it is necessary to propose a method to construct a supervised time topic model. To the best of our knowledge, this is the first work of constructing a supervised time topic model. The contributions of this paper are as follows.

We reveal and discuss the limitations of the main current relevant works on the topic model, and thus to construct a supervised time topic model is very necessary. We propose a method for constructing a supervised time topic model based on variational autoencoders, denoted by ST-TM, which is designed to deal with documents that have multiple labels and are time-stamped. The reasoning and constructing process of ST-TM is presented in detail. For a preliminary evaluation, we compare ST-TM with the state-of-the-art methods in the experiments. The results show that our proposed method is more effective.

The rest of the paper is organized as follows: we firstly review related research works. Secondly, the method for constructing a supervised time topic model based on variational autoencoders is described in detail. Then, the experiments of our proposed method are conducted. Finally, this paper is concluded.

2. Related Works

In this section, we review some representative works on the supervised topic models and time topic models in detail, respectively. Based on analysing the limitations of related works, we present a supervised time topic model to address these limitations.

2.1. Supervised Topic Model. To the best of our knowledge, supervised latent Dirichlet allocation is the first supervised topic model that adds to a topic associated with each document and uses variational methods to handle intractable posterior expectations [20]. Since then, several supervised topic models have been proposed, for example,

discriminative LDA [21] and maximum entropy discrimination LDA [22]. The above methods train only a single topic for each document.

It is well known that a document with only a single topic is inappropriate. For example, a document on social education includes both social and educational topics. Based on this fact, some supervised topic models with multiple topics are proposed. Labeled LDA matches the multiple topics to the labels in the document [9]. The number of topics is determined by the metadata of the document (such as labels), and topic terms have a better way to interpret topics [10]. Partially labeled LDA learns latent topic structure within the scope of observed, human interpretable labels [23]. Nonparametric labeled LDA uses the Dirichlet process with mixed random measures as a base distribution of the hierarchical Dirichlet process framework [24]. Dependency-LDA further considers the label frequency and label dependency observed in the training data for constructing the supervised topic model [2].

We review Labeled LDA (L-LDA) that is representative work on the supervised topic model. In order to contain the supervision, L-LDA applies a $1:1$ correspondence between topics and labels. In addition to labels, keywords of scientific papers and categories of news are also considered as topics [10]. L-LDA is a probabilistic graphical model that describes a process for generating a labeled document collection. The graphical model of L-LDA is shown in Figure 1.

In Figure 1, two observable parts of the grey fill c and w are explicit variables that indicate labels and terms, respectively. Nodes with no fill are implicit variables and unobservable. Unlike LDA, both the label set c and the topic prior α influence the topic mixture θ. L-LDA assumes that each document d is restricted to a multinomial distribution θ_d over labels which are part of the corpus. Each label is represented as a topic c that is a multinomial distribution Φ_c over the terms. The generative process of L-LDA is shown in Table 1. Different from LDA, the whole generative process of L-LDA contains topic constraint. The more detailed descriptions of L-LDA are presented in literature [9].

2.2. Time Topic Model. The time topic model is developed to analyse the time evolution of topics in large document collections. To the best of our knowledge, the dynamic topic model is the first time topic model that captures the evolution of topics in a sequentially organized corpus of documents [7]. In order to simplify the reasoning procedure of the time topic model, the method uses a variational autoencoder to construct the time topic model [8]. To build cross-lingual tools, a multilingual dynamic topic model is proposed that can capture cross-lingual topics that evolve across time [25]. We review the dynamic topic model (DTM) that is representative work on time topic model. DTM supposes that the data is divided by time slice, for example, by decades. The graphical model of DTM is shown in Figure 2.

In Figure 2, β is the vector of natural parameters for topic evolving over time and α is the vector of mean parameters of the logistic normal distribution for the topic proportions.
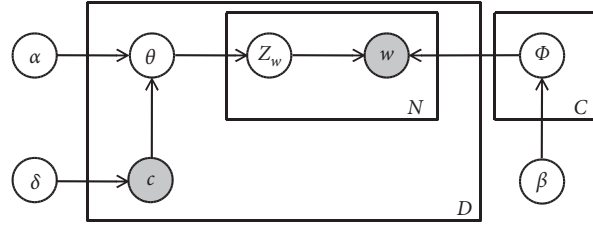
FIGURE 1: The graphical model of L-LDA.

TABLE 1: The generative process of L-LDA.

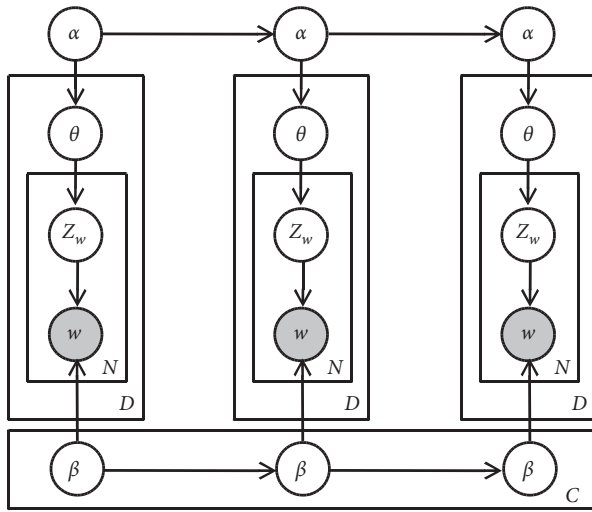| Steps	Description
1	For each topic $c \in C$:
2	Generate the multinomial topic distributions over vocabulary $\Phi_c \sim \text{Dirichlet}(\beta_c)$
3	End for
4	For each document $d \in D$:
5	Generate the multinomial topic distributions over document $\theta_d \sim \text{Dirichlet}(\alpha_d)$ according to the label set of each document d
6	For each term $W_{d,n} \in d$:
7	Generate topic $Z_{d,n} \sim \text{Multinomial}(\theta_d)$
8	Generate term $W_{d,n} \sim \text{Multinomial} \Phi_{Z_{d,n}}$
9	End for
10	End for

FIGURE 2: The graphical model of DTM.

TABLE 2: The generative process of DTM.

Steps	Description
1	Draw topics $\beta_t \vert \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$
2	Draw $\alpha_t \vert \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$
3	For each document $d \in D_t$:
4	Generate the multinomial topic distributions over document $\theta_{t,d} \sim \mathcal{N}(\alpha_t, a^2 I)$
5	For each term $W_{t,d,n} \in d$:
6	Generate topic $Z_{d,n} \sim \text{Multinomial}(\theta_{t,d})$
7	Generate term $W_{t,d,n} \sim \text{Multinomial}(\beta_{t,Z_{d,n}})$
8	End for
9	End for

these documents. As a result, it is necessary to propose a method to construct a supervised time topic model for documents with multiple labels and time-stamped.

However, the Dirichlet is not amenable to the sequential model. Literature [7] chains the vectors of parameters α and β in a state-space model that evolves with Gaussian noise, respectively. By chaining together topics and topic proportion distributions, DTM has sequentially tied a collection of topic models. The generative process for slice t of a sequential corpus is shown in Table 2. The whole process of DTM is restrained by time slice t; the more detailed descriptions of DTM are presented in the literature [7].

Using the above methods, the supervised topic model and time topic model are implemented for mining topics. However, some documents have multiple labels and are time-stamped in reality; for example, the scientific paper has keywords and is time-stamped. And there are few research papers on the supervised time topic model to learn topics for

3. Methods

In order to make the proposed method for constructing a supervised time topic model easier to describe and understand, we summarize some major notations needed in our formulation in Table 3.

The method of constructing a supervised time topic model is based on the generative process of the supervised topic model and time topic model. The difficulty is the reasoning of the loss function. In this section, we describe the method for constructing a supervised time topic model based on variational autoencoders in detail and denoted by ST-TM; the graphical model of ST-TM is shown in Figure 3.

The generative process of ST-TM is different from L-LDA and DTM by dividing time slices and having the supervision for each time slice, respectively. Moreover, unlike DTM, ST-TM removes transitive dependencies for α

TABLE 3: Notations description.

Notation	Description
t	Time slice t
c	The supervised topic
C	The set of topics in a text corpus
d	Document
W	Term
n	The number of terms in document d
β_t	The word-topic distribution parameters in time slice t
θ_t	The document-specific topic parameters in time slice t
z	The vector of words-topic
w_t	Input documents in time slice t
μ_t	The mean parameters for inputting documents in time slice t
σ_t	The covariance parameters for inputting documents in time slice t
w_t'	Generate documents in time slice t
μ_t'	The mean parameters for generating documents in time slice t
σ_t'	The covariance parameters for generating documents in time slice t

FIGURE 3: The graphical model of ST-TM.

and β so that the reasoning process of the model can further simplify. The generative process of ST-TM is shown in Table 4.

The implementation of ST-TM depends on a variational autoencoder that is a deep learning technique for learning latent representations. The variational autoencoder is composed of an encoder and decoder. In time slice t, the encoder is to generate a variational approximate posterior distribution $q_\phi(\theta_t|w_t)$ of the document-topic. Decoder estimates the optimal generation probability $p(w_t'|\theta_t, \beta_t)$ with $q_\phi(\theta_t|w_t)$ known. The network structure of the variational autoencoder of the topic model is shown in Figure 4.

In Figure 4, the model assumes $q(\theta|w)$ obeys diagonalization of the covariance matrices of the Gaussians distribution. The encoder trains the potential mean and variance for the vector of term-topic z. In order to generate documents w' that are as close to the input documents w, the decoder estimates the better probabilistic model $p(w'|\theta, z)$ according to the known $q(\theta|w)$. The generation process of θ needs to use backpropagation to calculate the gradient of the error function. Generally,

we want to minimize an expected cost by using gradient descent that requires computing the gradients. The generation process of θ adopts the reparametrization trick that is used in computing the gradients. The graphical model of the reparametrization trick is shown in Figure 5.

In Figure 5, the random variable $z \sim q(z|w)$ (at left) can be reparameterized as a Gaussian function and a random variable ε. More specifically, the reparametrization trick in this paper involves a function of a standard Gaussian variable $\varepsilon \sim \mathcal{N}(0, I)$ that is used to rewrite $z \sim \mathcal{N}(\mu, \sigma)$, such that $z = \mu + \sigma \odot \varepsilon$. Therefore, μ and σ only involve linear operation that can be optimized by stochastic gradient descent algorithm easily.

In this paper, the encoder has a 2-layer neural network to generate the variational approximate posterior distribution $q_\phi(\theta_t|w_t)$ of document-topic. The decoder estimates the better probabilistic model $p(w_t'|\theta_t, \beta_t)$ according to the known $q_\phi(\theta_t|w_t)$. According to the generation process of ST-TM, the marginal likelihood function of generating documents w' as a variational target is shown in equations (1) and (2):

TABLE 4: The generative process of ST-TM.

Steps	Description
1	For each topic $c \in C$ in time slice t:
2	Generate the multinomial topic distributions over vocabulary $\beta_{t,c}$ according to $\beta_t \sim \mathcal{N}(\mu_t, \sigma^2 I)$
3	End for
4	Draw $\alpha_t \sim \mathcal{N}(\mu_t, \sigma^2 I)$
5	For each document $d \in D_t$:
6	Generate the multinomial topic distributions over document $\theta_{t,d} \sim \mathcal{N}(\theta_t'|\alpha_t, a^2 I)$ according to $\theta_t \sim \mathcal{N}(\alpha_t, a^2 I)$ and the constraint matrix of document-topic θ_t'
7	For each term $W_{t,d,n} \in d$:
8	Generate topic $Z_{d,n} \sim \text{Multinomial}(\theta_{t,d})$
9	Generate term $W_{t,d,n} \sim \text{Multinomial}(\beta_{t,Z_{d,n}})$
10	End for
11	End for
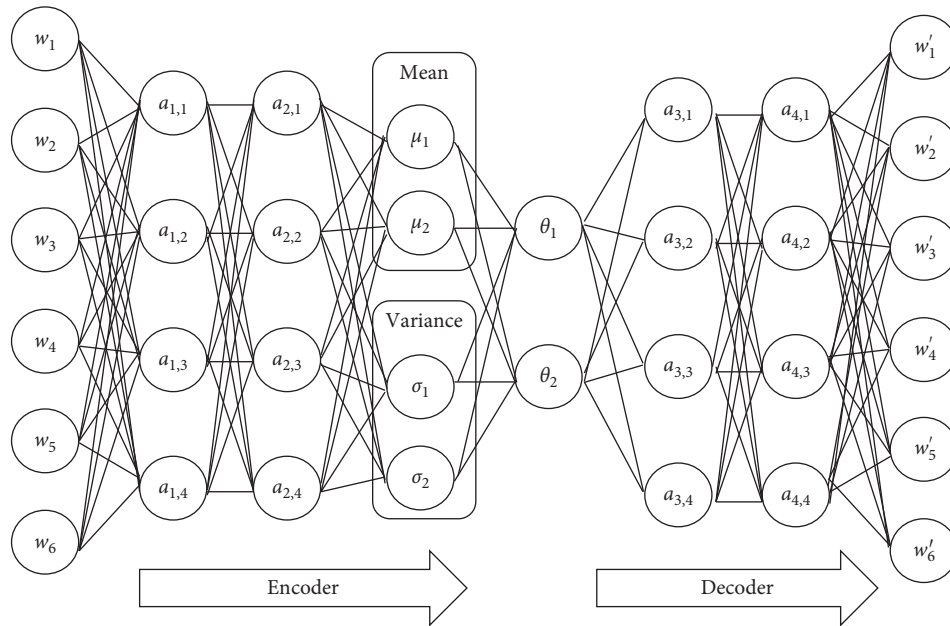
FIGURE 4: The network structure of variational autoencoder of the topic model.
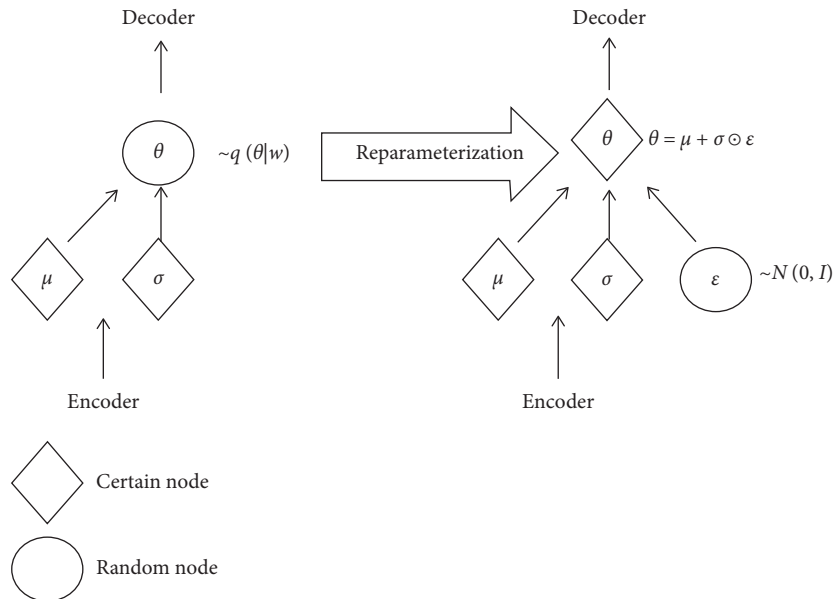
FIGURE 5: The graphical model of the reparametrization trick.

$$p(w'_t|\alpha_t, \beta_t, \theta'_t) = \int \left(\prod_{n=1}^{N} \sum_{z_{t,n}}^{K} p(w_{t,n}{}'|z_{t,n}, \beta_t) p(z_{t,n}|\theta_t) \right)$$
$$\cdot p(\theta_t|\alpha_t; \theta'_t) \mathrm{d}\theta_t, \tag{1}$$

$$p(w'_t|\alpha_t, \beta_t, \theta'_t) = \int \prod_{n=1}^{N} p(w_{t,n}{}'|\theta_t, \beta_t) p(\theta_t|\alpha_t; \theta'_t) \mathrm{d}\theta_t. \tag{2}$$

$$l(w'_t, \alpha_t, \beta_t, \theta'_t) = \log \int_{\theta_t} q_\phi(\theta_t|w_t, \theta'_t) \frac{p(w'_t|\theta_t, \beta_t) * p(\theta_t|\alpha_t; \theta'_t)}{q_\phi(\theta_t|w_t, \theta'_t)} \mathrm{d}\theta_t,$$

$$l(w'_t, \alpha_t, \beta_t, \theta'_t) \geq \int_{\theta_t} q_\phi(\theta_t|w_t, \theta'_t) * \log p(w'_t|\theta_t, \beta_t) \mathrm{d}\theta_t$$

$$+ \int_{\theta_t} q_\phi(\theta_t|w_t, \theta'_t) * \log \frac{p(\theta_t|\alpha_t; \theta'_t)}{q_\phi(\theta_t|w_t, \theta'_t)} \mathrm{d}\theta_t, \tag{3}$$

$$l(w'_t, \alpha_t, \beta_t, \theta'_t) \geq E_{q_\phi(\theta_t|w_t, \theta_t)} [\log p(w'_t|\theta_t, \beta_t)] - \mathrm{KL}\left[q_\phi(\theta_t|w_t, \theta'_t) \| p(\theta_t|\alpha_t; \theta'_t)\right],$$

$$E_{q_\phi(\theta_t|w_t, \theta_t)} [\log p(w'_t|\theta_t, \beta_t)] - \mathrm{KL}\left[q_\phi(\theta_t|w_t, \theta'_t) \| p(\theta_t|\alpha_t; \theta'_t)\right] \triangleq \mathcal{L}(q, \beta_t, \alpha_t; w_t, \theta'_t),$$

where $\mathcal{L}(q, \beta_t, \alpha_t; w_t, \theta'_t)$ is the variational lower bound for the marginal logarithmic likelihood function $l(w'_t, \alpha_t, \beta_t, \theta'_t)$, $l(w'_t, \alpha_t, \beta_t, \theta'_t) = \mathcal{L}(q, \beta_t, \alpha_t; w_t, \theta'_t)$ if and only if $q_\phi(\theta_t|w_t, \theta'_t) = p(\theta_t|\alpha_t; \theta'_t)$. In order to find the best $q_\phi(\theta_t|w_t, \theta'_t)$ which should be able to approximate the true posterior distribution $p(\theta_t|\alpha_t; \theta'_t)$, the most common approximate calculation of $l(w'_t, \alpha_t, \beta_t, \theta'_t)$ uses the Kullback–Leibler divergence (KL-divergence) between $q_\phi(\theta_t|w_t, \theta'_t)$ and $p(\theta_t|\alpha_t; \theta'_t)$. Our aim is to get $\arg\min \mathrm{KL} (q_\phi(\theta_t|w_t, \theta'_t) \| p(\theta_t|\alpha_t; \theta'_t))$ which is different from the precise KL-divergence $l(w'_t, \alpha_t, \beta_t, \theta'_t) - \mathcal{L}(q, \beta_t, \alpha_t; w_t, \theta'_t)$. Details are shown as follows:

$$\arg\min \mathrm{KL}\left(q_\phi(\theta_t|w_t, \theta'_t) \| p(\theta_t|\alpha_t; \theta'_t)\right)$$
$$= \arg\min\left(l(w'_t, \alpha_t, \beta_t, \theta'_t) - \mathcal{L}(q, \beta_t, \alpha_t; w_t, \theta'_t)\right), \tag{4}$$

$$\arg\min \mathrm{KL}\left(q_\phi(\theta_t|w_t, \theta'_t) \| p(\theta_t|\alpha_t; \theta'_t)\right)$$
$$= \arg\max \mathcal{L}(q, \beta_t, \alpha_t; w_t, \theta'_t), \tag{5}$$

$$\arg\min \mathrm{KL}\left(q_\phi(\theta_t|w_t, \theta'_t) \| p(\theta_t|\alpha_t; \theta'_t)\right)$$
$$= \arg\max\Big(E_{q_\phi(\theta_t|w_t, \theta_t)} [\log p(w'_t|\theta_t, \beta_t)] \tag{6}$$
$$- \mathrm{KL}\left[q_\phi(\theta_t|w_t, \theta'_t) \| p(\theta_t|\alpha_t; \theta'_t)\right]\Big),$$

where the first part is the expectation of reconstruction error and the second part is the KL-divergence of the approximate posterior distribution from the true posterior distribution. In t time slice, w'_t has N independent samples $(w_t^{1'}, w_t^{2'}, \ldots, w_t^{N'})$. As the first part of equation (6), $E_{q_\phi(\theta_t|w_t, \theta_t)}$

Unlike Table 3, n is the number of documents in time slice t. The variational functions $q_\phi(\theta_t|w_t, \theta'_t)$ are introduced for calculating $\log p(w'_t|\alpha_t, \beta_t, \theta'_t)$. According to Jensen's inequality, the marginal logarithmic likelihood function $l(w'_t, \alpha_t, \beta_t, \theta'_t)$ of generated document w'_t is as follows:

$[\log p(w'_t|\theta_t, \beta_t)]$ is generated by each sample with one point and then $E_{q_\phi(\theta_t|w_t^i, \theta_t)} [\log p(w_t^{i'}|\theta_t, \beta_t)] \approx \log p(w_t^{i'}|\theta_t, \beta_t)$ and obeys multivariate Bernoulli distribution. The expectation of reconstruction error can be measured by the cross entropy cost function, which is often used to measure the difference between the predicted value and actual value. In addition, the decrease of the learning rate of the loss function can be avoided in the process of gradient descent [26]. The logarithmic likelihood of $p(w'_t|\theta_t, \beta_t)$ uses the negative cross entropy denoted by L_1, where $L_1 = -w_t * \log p(w'_t|\theta_t, \beta_t)$.

The state variables β_t remove transitive dependency in order to simplify the complex reasoning process. Gaussian sampling is used to initialize the state variables β_t denoted by L_2, then $L_2 = \beta_t \sim \mathcal{N}(\mu_t, \sigma^2 I)$. In ST-TM, the encoder is to generate a variational approximate posterior distribution $q_\phi(\theta_t|w_t, \theta'_t)$ of the document-topic. We assume $q_\phi(\theta_t|w_t, \theta'_t)$ obeys diagonalization of the covariance matrices of the Gaussians distribution. Then, $\log q_\phi(\theta_t|w_t, \theta'_t) = \log \mathcal{N}(w_t; \mu'_t, \sigma'_t) \cdot \theta'_t$, where the mean μ'_t and the covariance σ'_t are generated by the neural network. Following the above reparametrization trick, we rewrite q_ϕ with an auxiliary parameter ε that makes $\theta_t = \sigma((\mu'_t + \sigma'^{(1/2)}_t * \varepsilon) \cdot \theta'_t)$. We apply this trick to obtain an estimator of the variational lower bound [27].

For the true posterior distribution, θ_t obeys a logarithmic Gaussian distribution with the mean μ_t and the variance σ_t, then $\log p(\theta_t|w_t, \theta'_t; \alpha_t) = \log \mathcal{N}(w_t; \mu_t, \sigma_t) \cdot \theta'_t$, where the covariance σ_t is the hyperparameters, and the mean $\mu_t = \alpha_t$ that is initialized by the Gaussian random sampling $\alpha_t \sim \mathcal{N}(1, \sigma_t^2 I)$ to take the place of $\alpha_t = 1$. According to the probability density function of multivariate Gaussian

distribution, the second part of (6) denoted by L_{kl} is inferred as follows:

$$
\begin{aligned}
\mathrm{KL}\left[q_\phi\left(\theta_t|w_t, \theta_t'\right)\| \, p\left(\theta_t|w_t, \theta_t'; \alpha_t\right)\right] &= E_{q_\phi}\left[\log q_\phi\left(\theta_t|w_t, \theta_t'\right) - \log p\left(\theta_t|w_t, \theta_t'; \alpha_t\right)\right] \\
&= \frac{1}{2}E_{q_\phi}\left[-\log|\sigma_t'| - (x - \mu_t')^T \sigma_t'^{(-1)}(x - \mu_t') + \log|\sigma_t| + (x - \mu_t)^T \sigma_t^{-1}(x - \mu_t)\right] \\
&= \frac{1}{2}\log\frac{|\sigma_t|}{|\sigma_t'|} + \frac{1}{2}E_{q_\phi}\left[-(x - \mu_t')^T \sigma_t'^{(-1)}(x - \mu_t') + (x - \mu_t)^T \sigma_t^{-1}(x - \mu_t)\right] \\
&= \frac{1}{2}\log\frac{|\sigma_t|}{|\sigma_t'|} + \frac{1}{2}E_{q_\phi}\left\{-tr\left[\sigma_t'^{(-1)}(x - \mu_t')(x - \mu_t')^T\right] + tr\left[\sigma_t^{-1}(x - \mu_t)(x - \mu_t)^T\right]\right\} \\
&= \frac{1}{2}\log\frac{|\sigma_t|}{|\sigma_t'|} + \frac{1}{2}E_{q_\phi}\left\{-tr\left[\sigma_t'^{(-1)}(x - \mu_t')(x - \mu_t')^T\right]\right\} + \frac{1}{2}E_{q_\phi}\left\{tr\left[\sigma_t^{-1}(x - \mu_t)(x - \mu_t)^T\right]\right\} \\
&= \frac{1}{2}\log\frac{|\sigma_t|}{|\sigma_t'|} - \frac{1}{2}tr\left\{E_{q_\phi}\left[\sigma_t'^{(-1)}(x - \mu_t')(x - \mu_t')^T\right]\right\} + \frac{1}{2}tr\left\{E_{q_\phi}\left[\sigma_t^{-1}(x - \mu_t)(x - \mu_t)^T\right]\right\} \\
&= \frac{1}{2}\log\frac{|\sigma_t|}{|\sigma_t'|} - \frac{1}{2}tr\left\{\sigma_t'^{(-1)}E_{q_\phi}\left[(x - \mu_t')(x - \mu_t')^T\right]\right\} \\
&\quad + \frac{1}{2}tr\left\{E_{q_\phi}\left[\sigma_t^{-1}\left(xx^T - \mu_t x^T - x\mu_t^T + \mu_t\mu_t^T\right)\right]\right\} \\
&= \frac{1}{2}\log\frac{|\sigma_t|}{|\sigma_t'|} - \frac{1}{2}tr\left\{\sigma_t'^{(-1)}\sigma_t'\right\} + \frac{1}{2}tr\left\{\sigma_t^{-1}E_{q_\phi}\left(xx^T - \mu_t x^T - x\mu_t^T + \mu_t\mu_t^T\right)\right\} \\
&= \frac{1}{2}\log\frac{|\sigma_t|}{|\sigma_t'|} - \frac{1}{2}|\theta_t'| + \frac{1}{2}tr\left\{\sigma_t^{-1}\left(\sigma_t' + \mu_t'\mu_t'^T - \mu_t\mu_t'^T - \mu_t'\mu_t^T + \mu_t\mu_t^T\right)\right\} \\
&= \frac{1}{2}\left\{\log\frac{|\sigma_t|}{|\sigma_t'|} - |\theta_t'| + tr\left(\sigma_t^{-1}\sigma_t'\right) + tr\left\{\sigma_t^{-1}\left(\mu_t'\mu_t'^T - \mu_t\mu_t'^T - \mu_t'\mu_t^T + \mu_t\mu_t^T\right)\right\}\right\} \\
&= \frac{1}{2}\left\{\log\frac{|\sigma_t|}{|\sigma_t'|} - |\theta_t'| + tr\left(\sigma_t^{-1}\sigma_t'\right) + tr\left(\sigma_t^{-1}\mu_t'\mu_t'^T - \sigma_t^{-1}\mu_t\mu_t'^T - \sigma_t^{-1}\mu_t'\mu_t^T + \sigma_t^{-1}\mu_t\mu_t^T\right)\right\} \\
&= \frac{1}{2}\left\{\log\frac{|\sigma_t|}{|\sigma_t'|} - |\theta_t'| + tr\left(\sigma_t^{-1}\sigma_t'\right) + tr\left(\mu_t'^T \sigma_t^{-1}\mu_t' - 2\mu_t'^T \sigma_t^{-1}\mu_t + \mu_t^T \sigma_t^{-1}\mu_t\right)\right\} \\
&= \frac{1}{2}\left\{\log\frac{|\sigma_t|}{|\sigma_t'|} - |\theta_t'| + tr\left(\sigma_t^{-1}\sigma_t'\right) + (\mu_t - \mu_t')^T \sigma_t^{-1}(\mu_t - \mu_t')\right\},
\end{aligned}
$$

$$(7)$$

where $|\theta_t'|$ is the column number of document-topic constraint matrix that is the number of topics in time slice t. In conclusion, the final variational objective can be represented by loss function $\mathscr{L}(\Theta)$ as follows:

$$\mathcal{L}(\Theta) = \arg\max\left(L_1 + L_2 - L_{kl}\right),$$

$$\mathcal{L}(\Theta) = \arg\min\left(-L_1 - L_2 + L_{kl}\right),$$

$$\mathcal{L}(\Theta) = \arg\min\{w_t * \log p(w_t'|\theta_t, \beta_t) + \beta_t + L_{kl}\},$$

$$\mathcal{L}(\Theta) = \arg\min\sum_i^D w_{t,i} * \log\left(\pi(\beta_t)\left(\pi(\mu_t' + \sigma_t' * \varepsilon) \cdot \theta_t'\right)\right) + \beta_t$$

$$+ \frac{1}{2}\left\{tr\left(\sigma_t^{-1}\sigma_t'\right) + (\mu_t - \mu_t')^T\sigma_t^{-1}(\mu_t - \mu_t') - |\theta_t'| + \log\frac{|\sigma_t|}{|\sigma_t'|}\right\},$$

$$(8)$$

where $\pi(\beta_t) = (\exp(\beta_{t,w})/\sum_w \exp(\beta_{t,w}))$, $\pi(\mu_t' + \sigma_t' * \varepsilon) = (\exp(\mu_{t,d}' + \sigma_{t,d}' * \varepsilon)/\sum_d \exp(\mu_{t,d}' + \sigma_{t,d}' * \varepsilon))$.

4. Experimental Results and Analysis

In this section, we conduct experiments to verify the efficiency of the proposed method on a real dataset. Our approach is compared with the state-of-the-art topic models.

4.1. Experiment Environment and Dataset.

The experiments were executed on a personal computer with an AMD FX-Series CPU FX-8350 @4.0 GHz, the processor having eight physical computing cores, 24 GB DDR3 RAM memory @1600 MHz. The machine was running on Windows 7 (64 bits) operating system, TensorFlow 1.4.0 with CPU support only, and Python 3.6.

To demonstrate the efficiency of our proposed approach, we use the paper corpus of SIGIR (International ACM SIGIR Conference on Research and Development in Information Retrieval) as the dataset. The main purpose of SIGIR is to show new technologies and achievements related to information retrieval. The dataset is constructed by abstracts of papers that are from the 2018, 2013, and 2009 annual conferences. Experimental data consisting of 564 papers with 76 topics and about 90000 words can be downloaded from the official SIGIR website. We divided the dataset into 3 time slices by year.

Our neural networks structure is composed of an encoder and decoder. The encoder has one input layer and two hidden layers, and the decoder has two hidden layers and one output layer. Input layer and output layer have the same number of neurons as the number of papers, which changes in different time slices. Every hidden layer has 100 hidden neurons according to the existing experimental conclusion [8]. Moreover, the learning rate is 0.002 and the number of iterations is 300 that can achieve convergence for our dataset.

4.2. Evaluation Metrics.

It is difficult to choose the appropriate evaluation method of the topic model. Although the traditional evaluating method uses perplexity to evaluate the qualitative of topics, it is generally not well suited to express the semantic property of topic terms [28]. Lau et al. [29] use Normalized Pointwise Mutual Information (NPMI) to evaluate the qualitative of topics, and NPMI is close to the ordinary understanding and judgment of people.

The experiment is evaluated by NPMI(k) that can express the coherence of topic k, and NPMI(k) is given as

$$\text{NPMI}(k) = \sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\frac{\log\left(p(w_i, w_j)/p(w_i)p(w_j)\right)}{-\log p(w_i, w_j)}, \quad (9)$$

where M is the number of topic terms in topic k. w_i and w_j are the ith topic term and the jth topic term, respectively. $p(w_i)$, $p(w_j)$, and $p(w_i, w_j)$ are the probability of the ith topic term, the probability of the jth topic term, and the cooccurrence probability of the ith and the jth topic term in topic k. Experiment results depend on the average topic coherence of K topics at last. The higher value the average of NPMI is, the better the topic model is.

4.3. Baseline Methods.

To validate the effectiveness of our method, three state-of-the-art topic model methods have been compared with ST-TM. The first baseline is collapsed Gibbs sampling on the LDA model denoted by SDTM Gibbs [30]. The second one is the static topic model based on autoencoding variational inference denoted by LDA VAE [31]. The last one uses a variational autoencoder to construct a time topic model denoted by DTM VAE [8]. The details of all methods are provided in Table 5. All baseline methods are conducted by the original paper of the corresponding methods.

4.4. Comparison with Baseline Methods.

We conduct a preliminary experiment. The convergence process of each method is shown in Figure 6. The convergence of Gibbs sampling depends on the perplexity value which is a measurement of stability for the topic model. The loss function is used to determine the convergence of other methods that are based on a variational autoencoder.

All methods on NPMI value are shown in Figure 7. Our method has the highest NPMI value at 0.316, while the other methods are at 0.083, 0.169, and 0.299, respectively. Comparing to the traditional Gibbs sampling method of constructing the topic model, the variational autoencoder method is more efficient. The supervised time topic model outperforms the static topic model and unsupervised time topic model on the semantic property of topic terms. The

TABLE 5: Details of all methods.

Method	Variational autoencoder	Time slices	Supervision
SDTM Gibbs	No	Yes	Yes
LDA VAE	Yes	No	No
DTM VAE	Yes	Yes	No
ST-TM	Yes	Yes	Yes

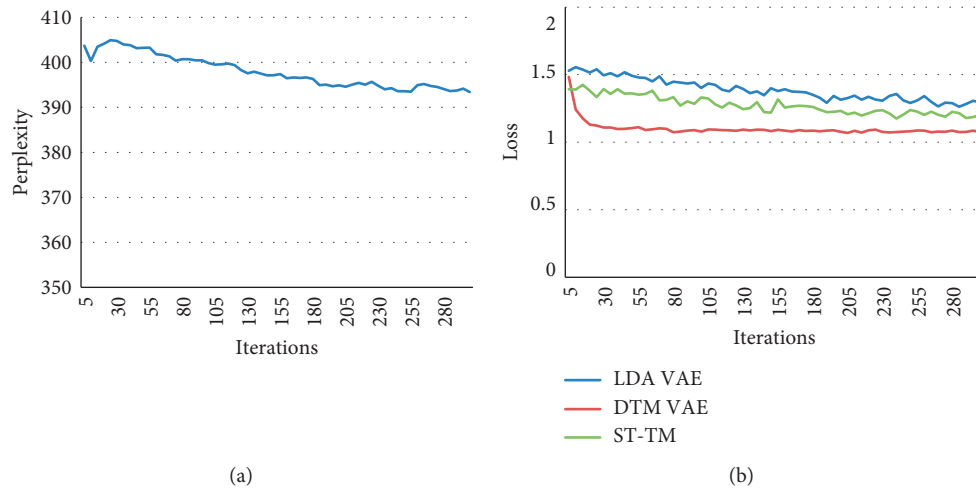(a)   (b)

FIGURE 6: The convergence process of all methods. (a) Gibbs sampling. (b) Other methods.
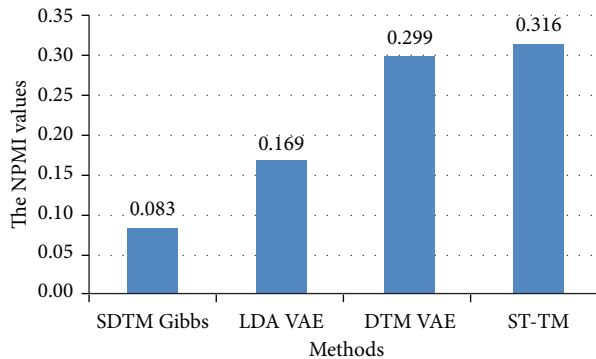
FIGURE 7: Comparison of all methods on NPMI.

result of the preliminary experiment shows that ST-TM has excellent performance at interpretation of the topic.

4.5. Discussion. The proposed method for constructing the supervised time topic model is verified through the public dataset, and the experimental results show that the average of NPMI is improved with time slices and supervision. The proposed method is easier to operate compared with the traditional Gibbs sampling method of constructing the topic model [30] and can achieve a good result. In addition, the proposed method is based on the basic structure of the variational autoencoder method [31] and introduces time slices and supervision for understanding the supervised time topic model. The largest advantage of this work is in solving the problem of constructing a supervised time topic model for some documents with multiple labels and which are time-stamped to achieve document-related tasks.

5. Conclusions

In this paper, we propose a method for constructing a supervised time topic model based on variational autoencoders. We reveal and discuss the limitations of the main current relevant works on the supervised topic models and time topic models, respectively. We further present the reasoning and constructing process of the supervised time topic model. Specifically, the graphical model of constructing process of supervised time topic model is proposed. The implementation of the method depends on a variational autoencoder that is composed of an encoder and decoder. An encoder is to generate a variational approximate

posterior distribution of the document-topic, and a decoder estimates the optimal generation probability. The process of reasoning of the marginal likelihood function of generating documents as a variational target is presented. In addition, we compare our method with baselines in the experiment for a preliminary evaluation. The results show that our method is more effective for the topic model than baseline methods. The contribution of this work is in solving the problem of constructing a supervised time topic model for some documents with multiple labels and time-stamped to achieve document-related tasks. The relevant theories of the topic model can be conducive for researchers to understand, reason, and use the supervised time topic model.

In future work, we will improve the method of constructing a supervised time topic model for further completion, refinement, and adaptation. Furthermore, to achieve better performance, we will also apply our constructed supervised time topic model in information retrieval, recommender systems, text classification, and other fields.

Data Availability

According to the funding policy of this work, data cannot be shared or made publicly available during the funding contract.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] T. N. Rubin, A. Chambers, P. Smyth et al., "Statistical topic models for multi-label document classifica-tion," *Machine Learning*, vol. 88, no. 1-2, pp. 157–208, 2012.

[3] P. Yang, Y. Yao, and H. Zhou, "Leveraging global and local topic popularities for LDA-based document clustering," *IEEE Access*, vol. 8, pp. 24734–24745, 2020.

[4] Y. Zhong, Z. Tang, X. Ding et al., "An improved LDA multi-document summarization model based on TensorFlow," in *Proceedings Of 29th IEEE International Conference On Tools With Artificial Intelligence*, pp. 255–259, Boston, MA, USA, November 2017.

[5] G. Doyle and C. Elkan, "Accounting for burstiness in topic models," in *Proceedings Of the 26th Annual International Conference On Machine Learning*, pp. 281–288, Quebec, Canada, June 2009.

[6] R. E. Madsen, D. Kauchak, and C. Elkan, "Modeling word burstiness using the Dirichlet distribution," in *Proceedings Of the 22nd International Conference On Machine Learning*, pp. 545–552, Bonn, Germany, August 2005.

[7] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings Of the Twenty-Third International Conference*, pp. 113–120, Pittsburgh, PA, USA, June 2006.

[8] Z. Gou, L. Han, L. Sun, J. Zhu, and H. Yan, "Constructing dynamic topic models based on variational autoencoder and factor graph," *IEEE Access*, vol. 6, pp. 53102–53111, 2018.

[9] D. Ramage, D. Hall, R. Nallapati et al., "Labeled LDA: a supervised topic model for credit attribution in multilabeled corpora," in *Proceedings Of the 2009 Conference On Empirical Methods In Natural Language Processing*, pp. 248–256, Singapore, August 2009.

[10] Z. Gou, Z. Huo, Y. Liu et al., "A method for constructing supervised topic model based on term frequency-inverse topic frequency," *Symmetry*, vol. 11, no. 12, 2019.

[11] F. Xue, J. Wang, S. Qian, T. Zhang, X. Liu, and C. Xu, "Multi-modal max-margin supervised topic model for social event analysis," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 141–160, 2019.

[12] J. Wang, L. Xia, X. Hu, and Y. Xiao, "Abnormal event detection with semi-supervised sparse topic model," *Neural Computing and Applications*, vol. 31, no. 5, pp. 1607–1617, 2019.

[13] H. Soleimani and D. J. Miller, "Exploiting the value of class labels on high-dimensional feature spaces: topic models for semi-supervised document classification," *Pattern Analysis and Applications*, vol. 22, no. 2, pp. 299–309, 2019.

[14] Y. Wu, S. Xi, Y. Yao, F. Xu, H. Tong, and J. Lu, "Guiding supervised topic modeling for content based tag recommendation," *Neurocomputing*, vol. 314, pp. 479–489, 2018.

[15] J. Han, Y. Huang, K. Kumar, and S. Bhattacharya, "Time-varying dynamic topic model," *Journal of Global Information Management*, vol. 26, no. 1, pp. 104–119, 2018.

[16] O. Isupova, D. Kuzin, and L. Mihaylova, "Learning methods for dynamic topic modeling in automated behavior analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 3980–3993, 2018.

[17] Z. Xu, L. Chen, Y. Dai, and G. Chen, "A dynamic topic model and matrix factorization-based travel recommendation method exploiting ubiquitous data," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1933–1945, 2017.

[18] F. Yao and Y. Wang, "Tracking urban geo-topics based on dynamic topic model," *Computers, Environment and Urban Systems*, vol. 79, 2020.

[19] S. Park, W. Lee, and I. Moon, "Supervised dynamic topic models for associative topic extraction with a numerical time series," in *Proceedings Of the 2015 Workshop On Topic Models: Post-Processing And Applications*, pp. 49–54, Melbourne, Australia, October 2015.

[20] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Proceedings Of the Twenty-First Annual Conference On Neural Information Processing Systems*, pp. 121–128, Vancouver, Canada, December 2006.

[21] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: discriminative learning for dimensionality reduction and classification," in *Proceedings Of the Twenty-Second Annual Conference On Neural Information Processing Systems*, pp. 897–904, Vancouver, Canada, December 2008.

[22] J. Zhu, A. Ahmed, and E. P. Xing, "M.L. D. A.: maximum margin supervised topic models," *Journal of Machine Learning Research*, vol. 13, pp. 2237–2278, 2012.

[23] D. Ramage, C. D. Manning, and S. Dumais, "Partially labeled topic models for interpretable text mining," in *Proceedings Of*

the 17th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, pp. 457–465, San Diego, CA, USA, August 2011.

[24] D. Kim, S. Kim, and A. Oh, "Dirichlet process with mixed random measures: a nonparametric topic model for labeled data," in Proceedings Of the 29th International Conference On Machine Learning, pp. 675–682, Edinburgh, UK, July 2012.

[25] E. Zosa and M. Granroth-Wilding, "Multilingual dynamic topic model," in Proceedings Of the International Conference On Recent Advances In Natural Language Processing, pp. 1388–1396, Varna, Bulgaria, September 2019.

[26] P. Mirowski, M. Ranzato, and Y. LeCun, "Dynamic auto-encoders for semantic indexing," in Proceedings of the 24th Annual Conference on Neural Information Processing Systems, pp. 1–9, Vancouver, Canada, December 2010.

[27] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," 2014, http://arxiv.org/abs/13126114.

[28] D. Newman, J. H. Lau, K. Grieser et al., "Automatic evaluation of topic coherence," in Proceedings Of the Conference Of the North American Chapter of the Association of Computational Linguistics, pp. 100–108, Los Angeles, CA, USA, June 2010.

[29] J. H. Lau, D. Newman, and 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, pp. 530–539, Gothenburg, Sweden, April 2014.

[30] A. McCallum, "MALLET version 0.4-2.0.8," 2018, http://mallet.cs.umass.edu/download.php.

[31] A. Srivastava and C. A. Sutton, "Autoencoding variational inference for topic models," in Proceedings Of the 5th International Conference On Learning Representations, Toulon, France, April 2017.