Xie, Qianqian; Tiwari, Prayag; Gupta, Deepak; Huang, Jimin; Peng, Min

Neural Variational Sparse Topic Model for Sparse Explainable Text Representation

Published in:
Information Processing and Management

DOI:
10.1016/j.ipm.2021.102614

Published: 01/09/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Xie, Q., Tiwari, P., Gupta, D., Huang, J., & Peng, M. (2021). Neural Variational Sparse Topic Model for Sparse Explainable Text Representation. *Information Processing and Management*, *58*(5), 1-15. Article 102614. https://doi.org/10.1016/j.ipm.2021.102614

Neural Variational Sparse Topic Model for Sparse Explainable Text Representation

Qianqian Xie[a], Prayag Tiwari[b,*], Deepak Gupta[c], Jimin Huang[d], Min Peng[d]

[a]*Department of Computer Science, University of Manchester, Manchester, United Kingdom*
[b]*Department of Computer Science, Aalto University, Espoo, Finland*
[c]*Maharaja Agrasen Institute of Technology, Delhi, India*
[d]*School of Computer Science, Wuhan University, Wuhan, China*

Abstract

Texts are the major information carrier for internet users, from which learning the latent representations has important research and practical value. Neural topic models have been proposed and have great performance in extracting interpretable latent topics and representations of texts. However, there remain two major limitations: 1) these methods generally ignore the contextual information of texts and have limited feature representation ability due to the shallow feed-forward network architecture, 2) Sparsity of the representations in topic semantic space is ignored. To address these issues, in this paper, we propose a semantic reinforcement neural variational sparse topic model (SR-NSTM) towards explainable and sparse latent text representation learning. Compared with existing neural topic models, SR-NSTM models the generative process of texts with probabilistic distributions parameterized with neural networks and incorporates Bi-directional LSTM to embed contextual information at the document level. It achieves sparse posterior representations over documents and words with zero-mean Laplace distribution and topics with sparsemax. Moreover, we propose a supervised extension of SR-NSTM via adding the max-margin posterior regularization to tackle the supervised tasks. The neural variational inference method is utilized to learn our models efficiently. Experimental results on Web Snippets, 20Newsgroups, BBC, and Biomedical datasets demonstrate that the contextual information and revisiting generative process can improve the performance, leading to the competitive performance of our models in learning coherent topics and explainable sparse representations for texts.

Keywords: Neural Sparse Topic Model, Neural Variational Inference, Explainable Text Representation

1. Introduction

On the internet, there are massive texts posted by active internet users, such as tweets, microblog texts, and news headlines, which carry a variety of valuable information, such as public opinions, social hot spots, and user interests. Learning the latent representations of texts has significant research and practical value. Deep neural networks including as convolutional neural networks (CNNs) [1] and recurrent neural networks (RNNs) [2], graph neural networks [3], have shown strengths in learning text representations. However, one well-known issue of them is that the learned representations of them are difficult to interpret [4]. On the other hand, topic modeling has been one of the most effective text analysis tools and able to generate interpretable topics. However, the application and extension of these methods are limited, since the subtle variant of the model increases the complexity of the probabilistic generation architecture, and requires the re-deduction of the whole inference process. Therefore, it is intuitive to combine both neural networks and

[*] Corresponding author

Email addresses: qianqian.xie@manchester.ac.uk (Qianqian Xie), prayag.tiwari@aalto.fi (Prayag Tiwari), deepakgupta@mait.ac.in (Deepak Gupta), huangjimin@whu.edu.cn (Jimin Huang), pengm@whu.edu.cn (Min Peng)

topic models to create advanced models that can be derived easily and be able to learn explainable neural representations.

To address this issue, a series of studies have been carried out in this line of research. There are many researches devoting to combine deep learning techniques with topic models. In these models [5, 4, 6, 7, 8, 9], the back propagation method is utilized to automatically update the parameters during the training, since the probabilistic mixtures in the generative process of traditional topic models are replaced by deep neural networks. Thus, they can be derived easily and extended flexibly with the neural network structures and the back propagation. Moreover, they also adopt pre-trained word embeddings which are proved to be effective in capturing the semantic information at the word level, achieving great performance on extracting meaningful latent topics and representation of texts. However, there are still major challenges for these neural topic models: 1) they ignore contextual information at the document level, limiting the express ability of the generated latent representations. With feed-forward neural networks, they are unable to model the sequence structure of words in the document. 2) they don't achieve sparsity of the representations in the topic semantic space. Yielding sparsity in representations has been proved to be effective in improving the discrimination and explainability of learned representations [10, 11]. In reality, it is intuitive that each document focuses on a few topics and each topic focuses on a few words.

To address these challenges, we propose a novel neural sparse topic model called semantic reinforcement neural variational sparse topic model (SR-NSTM), aiming to learn interpretable and more efficient latent representations of texts with sparse distribution prior. Different from previous approaches, SR-NSTM utilizes the neural networks to parameterize the prior distributions in the generative process of texts, rather than directly replacing the mixtures with neural networks. Our method further incorporates Bi-directional Long short-term memory (Bi-LSTM), to consider the sequential structure of words at the document level, which enriches the semantic information provided by texts. To achieve the sparsity enhancement in SR-NSTM, we utilize the parameterized zero-mean Laplace distribution to achieve sparse posterior representations over documents and words, and the sparsemax function to yield sparse representations over topics. We adopt the neural variational inference method to approximate the posterior distribution and reparameterize the sampling process with neural networks. Moreover, we further present a supervised extension of SR-NSTM to learn predictive representations with the max-margin posterior constraints, which can be directly utilized in supervised tasks. Our proposed method inherits the probabilistic characteristics of the sparse topic model, and inference in end-to-end style as the neural networks. Thus, it can be extended flexibly with additional hypothesis or regularization in the new scenario, while achieving explainable latent semantic representations of texts that humans can interpret.

1.1. Contribution

The main contributions of our paper can be summarized as follows:

- We propose a semantic reinforcement neural topic model SR-NSTM for sparse and explainable text representation. To learn more effective representations of texts, SR-NSTM revisits the generative process of sparse topic models and incorporates the contextual information with Bi-LSTM.
- We extend our model to supervise learning tasks with the max-margin posterior constraints and inference our models with the neural variational inference method.
- Experimental results on four text datasets demonstrate the superiority of our models in perplexity, topic coherence and text classification accuracy.

2. Related Work

Previous researches related to our work can be divided into three parts: traditional sparse topic models, neural topic models, and neural sparse topic models.

Sparse Topic Models. The sparsity of the representations in the semantic space is critical in improving the discriminating and explainability of representations [10, 12]. There were many sparsity-enhanced topic models, which aimed at extracting meaningful latent representations of texts and alleviating the issue of the sparse word co-occurrence information. Eisenstein et al. [13] proposed an alternative generative model called SAGE, in which each class label and the latent topic was endowed with a model of the deviation in log-frequency from a constant background distribution. Chen et al. [14] presented cFTM via leveraging contextual information about the author and document venue, in which the hierarchical beta process was employed to infer a focused set of topics associated with each author and venue. However, it can only control the sparsity of topics and ignored the sparsity at the document level. To achieve sparse representations in the document-topic and topic-term distributions, Williamson et al. [15] introduced a "Spike and Slab" prior to model the sparsity in finite and infinite latent topic structures of text. With the same purpose, Lin et al. [11] proposed a dual-sparse topic model that addressed the sparsity in both the topic mixtures and the word via applying the "Spike and Slab prior". Different from the above methods, Zhu et al. [10] presented sparse topical coding (STC) by utilizing the Laplacian prior to directly control the sparsity of inferred representations. Based on STC, Peng et al. [16] proposed a Bayesian Sparse Topical Coding (BSTC) by introducing sparse Bayesian learning to improve the modeling of the sparse structure of texts. However, the extension and application of these models are limited since their inference process is difficult due to the complex hierarchical structure.

Neural Topic Models. Deep learning techniques have shown great performance on various tasks, such as image classification, machine translation, and so on. The models based on deep neural networks can automatically update parameters during training via the back propagation method, thus can be trained in end-to-end style. They require no manual deriving and have high flexibility. To address the aforementioned issues, there were researches incorporating deep neural networks with topic models to improve the inference process. Larochelle et al. [5] proposed a neural network topic model inspired by the Replicated Softmax. Cao et al. [4] proposed a neural topic model (NTM) and presented a uniform framework where the representation of words and documents are efficiently and naturally combined. However, it didn't take the word order in texts into consideration. To deal with the problem, Tian et al. [17] proposed Sentence Level Recurrent Topic Model (SLRTM) to capture the sequence structure based on Recurrent Neural Networks (RNN).

Besides, there were attempts applying neural variational inference which can approximate the posterior distribution of a generative model with a variational parameterized by a neural network. Srivastava et al. [18] presented auto-encoding variational Bayes (AEVB) for topic models to improve the inference process. Miao et al. [19] provided an alternative neural approach in topic modeling based on parameterized distributions over topics. Inspired by [18], Card et al. [6] combined several variations of topic models with neural variational inference, including the supervise information and the sparse distribution. Cong et al. [20] presented TLASGR MCMC to learn simplex-constrained global parameters of all layers and topics simultaneously. Zhang et al. [21] developed Weibull hybrid autoencoding inference (WHAI) for topic models, with a hierarchy of gamma distributions in the generative network and a hierarchy of Weibull distributions in the inference network.

There were also previous researches focusing on incorporating word embeddings into topic models, which have been proved to be effective in capturing the contextual semantics of words via representation learning. Das et al. [22] modeled the document as a collection of word embeddings and topics as multivariate Gaussian distributions in the embedding space. However, the assumption that topics are unimodal in the embedding space is not appropriate, since topically related words can occur distantly from each other in the embedding space. Therefore, Hu et al. [23] proposed a latent concept topic model which introduced the concept as the distribution of word embeddings and modeled the topic as the distribution of concepts. Nguyen et al. [24] proposed to extend the LDA with word embeddings as latent features. Li et al. [25] combined the local information of word embeddings with the global information provided by LDA. Xun et al. [26] modeled each short document as a Gaussian topic over word embeddings in the vector space. Based on the same hypothesis, Xun et al. [27] further learned topic correlations among the continuous Gaussian topics. Batmanghelich et al. [28] adopted the von Mises-Fisher distribution to model the word embeddings in topic models. Bunk et al. [29] exchanged selected topic words via Gibbs sampling while estimated the topic distribution in the word embedding space. Xu et al. [30] adopted the Wasserstein distances with a distillation mechanism, to learn topics and word embeddings jointly. Dieng et al. [31] utilized the inner product between a word embedding and an embedding of the assigned topic to parameterize a categorical distribution as the word in topic models.

Nevertheless, most of the aforementioned works focused on the application of topic models in learning dense representations without considering the sparsity in the semantic space.

Neural Sparse Topic Models. There were also researches incorporating deep learning techniques with sparse topic models to improve the flexibility of these complex models and improve representation learning in texts when compared with neural topic models. Card et al. [6] produced a neural framework based on sparse additive generative models, to flexible incorporate the metadata of documents. It achieved strong performance on several metrics. Base on STC [10], Peng et al.[9] proposed neural sparse topical coding (NSTC) and its extensions to derive sparse representations of words and documents. It significantly improved the flexibility and efficiency of the original sparse topic model. Most recently, Lin et al. [7] proposed Neural SparseMax Document and Topic Models, which utilized sparsemax to directly control the sparsity of the topics. It outperformed previous neural sparse topic methods such as [6] in quality and stability. However, previous neural sparse topic models can not model the generative process of texts accurately based on only feed-forward networks. Compared with previous approaches, our models explicitly model the generative process of texts with sparse priors: the zero-mean Laplace prior distribution parameterized with neural networks, and also incorporates the context semantic information of documents to further improve the learning of document topic distribution.

3. Method

Before introducing our method, we make some definitions. We define that $D = \{1, ..., M\}$ is a document set with size M, $T = \{1, ..., K\}$ is a topic collection with K topics, $V = \{1, .., N\}$ is the vocabulary of the whole data set, and $w_d = \{w_{d,1}, .., w_{d,|I|}\}$ is a vector of terms representing a document d, where I is the index of words in document d, and $w_{d,n}(n \in I)$ is the frequency of word n in document d. We denote $\beta \in \mathbb{R}^{N \times K}$ as a global topic dictionary with K bases learned from the whole document set. Each column of it is a unigram distribution over V. $\vartheta_d \in \mathbb{R}^K$ is the latent representation of a document d in topic space, referred to the document code of d. $s_{d,n} \in \mathbb{R}^K$ is the latent representation of a word n in topic space, referred to the word code of n in d. To yield interpretable patterns, (ϑ, s, β) are constrained to be non-negative.

We start by illustrating the explainability of text representation. One well-known limitation of existing deep neural networks is the lack of interpretability. Interpretability of models is critical since it helps users to understand the overall strengths and weaknesses of the models. Generally, the text representation is deemed as interpretable, if each dimension of the representation corresponds to a fine-grained sense or a semantically coherent cluster [32]. In our model, the explainability of learned text representations, is offered by the topic structure and sparse mechanism. A topic model is a probabilistic generative model which represents a document as the mixtures of latent topics, while each topic is a probabilistic distribution over words. Based on the meaningful topics, the derived text representation by our topic model in the sparse topic semantic space makes more sense to humans compared to neural networks (embedding), in which each dimension of text representations is denoted as a coherent semantic concept namely: topic [33, 34, 35, 36]. To allow a user to better understand the topic structure, topics can be visually represented by word clusters (e.g. the top 10 or 20 most probable words), which can help the user understand the meaning of each topic and interpret each dimension of learned topics. Generally, the more topics are coherent, the more they are interpretable. Therefore, the interpretability of representations can be evaluated by the topic coherence, which is approximately calculated by the pointwise mutual information (PMI).

3.1. Neural Sparse Topical Coding

We then start by reviewing the traditional sparse topic model STC [10] and the latest neural sparse topic model NSTC[9]. STC is a non-probabilistic topic model, which aims to induce sparsity in generated representations with sparse regularization. In STC, word counts are assumed to be independent and can be reconstructed from the linear combination of a set of topic bases and the latent word code. To achieve sparsity, STC defines the prior distribution of word code as a super-Gaussian distribution with an isotropic

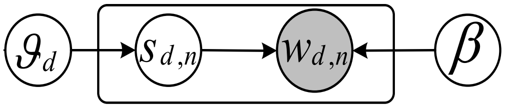


Figure 1: The graphical model of NSTM

Gaussian distribution and a Laplace distribution. However, the inference process is complex despite that there are closed-form coordinate descent equations for model parameters. Based on Sparse Topical Coding (STC)[10], [9] proposed an easily expandable neural spare topic model NSTC, by rebuilding STC with a neural network to simplify the inference process of the model and introduce word embeddings to benefit the learning of the topic dictionary. It combines the advantages of neural topic models and STC, with a flexible model structure that can learn sparse latent representation over the document and word level, and meaningful topic bases. After generating the topic dictionary from the neural network, NSTC follows the generative story below for each document d:

1. For each word n in document d:
   (a) Sample a latent variable word code $s_{d,n} \sim f_g(d,n)$.
   (b) Sample the observed word count $w_{d,n}$ from $p(w_{d,n}|s_{d,n},\beta_n) \sim Poisson(s_{d,n} * \beta_n)$

f_g is a feed-forward neural network to generate word code based on the word n in the document d and $p(w_{d,n}|s_{d,n},\beta_n)$ is the Poisson distribution for sampling observed word count. It collapses the document code from the generative process, and replaces the composite super-Gaussian prior of the word codes and the uniform distribution of the topic dictionary with the neural network. Despite the improvement in model flexibility, it cannot explicitly model the relations between random variables and the generative process as a non-probabilistic model. Moreover, it ignores the contextual information of words in the document with feed-forward neural networks.

3.2. Neural Variational Sparse Topic Model

To explicitly model the probabilistic distributions of latent variables in topic models, we devote to propose a novel neural sparse topic model NSTM. We revisit the generative process of NSTC with parameterized distributions. We consider NSTM as a deep generative model for observed data w that depends on a hierarchy of latent variables $\{\vartheta, s\}$, and follows the generative story below for each document d:

1. Sample the topic dictionary: $\beta = f_\beta(e)$
2. Sample a document code $\vartheta_d \sim p_\phi(\vartheta_d)$
3. For each word n in document d:
   (a) Sample a latent variable word code $s_{d,n} \sim p_\phi(s_{d,n}|\vartheta_d)$
   (b) Sample the observed word count $w_{d,n} \sim p_{\phi_{d,n}}(w_{d,n}|s_{d,n},\beta_n)$

ϕ refers to the parameter of prior distributions, and f_β is a neural network that deriving the topic dictionary from the representations e of all words mentioned in the corpus. The graphical representation of NSTM is depicted in Figure 1. In our model, we have several assumptions: 1) The document code is the document latent representation in topic space as the document topic distribution in LDA. Different from NSTC, we sample it from the zero-mean Laplacian distribution. Similarly, the word code is the word latent representation in topic space and sampled from the Gaussian distribution given the document code. 2) We hypothesis that each observed word count can be reconstructed from a linear combination between a set of topic bases and the word code as the coefficient vector. It is sampled from the Poisson distribution with mean parameter $s_{d,n}\beta_n$. 3) We consider the topic dictionary as a global variable. Rather than uniform distribution, we sample the topic dictionary from a topic dictionary neural network. Thus, we can introduce the external word embeddings into our model without increasing model complexity. Similar to NSTC, we can take the word embeddings as the input of the topic dictionary neural network rather than extra latent variables. The contextual information captured in word embeddings can help semantically similar words distributing in the same topic basis, resulting in better learning of word and document codes, without increasing the complexity of the model.

The topic dictionary neural network is comprised of the following layers:

Input layer ($e \in \mathbb{R}^{N\times 300}$): Supposing the word number of the vocabulary is N, each word is converted into a continuous representation with an embedding matrix e in this layer. Here, we adopt the pre-trained embeddings by GloVe based on a large Wikipedia dataset. [1].

Topic dictionary layer ($\beta \in \mathbb{R}^{N\times K}$): This is a fully connected layer that converts the word embeddings of all terms e to a topic dictionary with K topics: $\beta = e * W + b$, where $W \in \mathbb{R}^{300\times K}$ is a weight matrix and b is the bias. Here, we adopt the sparsemax transformation [37] on each topic basis for sparse and meaningful topic bases, in which the related words are focused while the unimportant ones are ignored with zero probability. Sparsemax can yield the Euclidean projection of the input vector via the probability simplex. For the closed-form expression of its Jacobian and a smooth convex loss function, Sparsemax can be directly incorporated in neural networks and trained with back-propagation algorithm. Therefore, we normalize each topic basis of the dictionary via Sparsemax as follow:

$$Sparsemax(\beta_{.k}) : argmin_{p\in\Delta^{N-1}}||p-\beta_{.k}||^2, \forall k \tag{1}$$

where p is a $N-1$ simplex.

The major difference between NSTM and our previous approach NSTC is the modeling of the document code θ_d and the latent word code$s_{d,n}$. As shown in the generative story, NSTM sample the document code and word code from prior distributions respectively, while NSTC collapses the document code and directly generates the word code from a feed-forward neural network. Therefore, NSTM can directly control the sparsity of document code and explicitly model the relations between latent variables to generate sparse and meaningful document and word representations. In general, NSTM inherits the probabilistic characteristics of traditional sparse topic models. It can effectively capture the correlation information in the texts with limited length and generate interpretable meaningful representations for words and documents when compared with neural topic models. However, the inference is more complicated than NSTC since the sampling of the mixtures is considered in the model. Thus, we adopt the neural variational inference to approximate the posterior distribution and automatically update the parameters with back-propagation method, which will be further introduced in the following section.

3.3. Neural Variational Inference for NSTM

3.3.1. The Variational Bound

In our model, the posterior inference over parameters is intractable. The general solution is the variational inference via introducing the variational approximation optimized to the true posterior distribution. We aim to maximize the probability of word count w in document under the generative process:

$$\begin{aligned} logp_\phi(w) &= log\int p_\phi(w|s,\beta)p_\phi(s|\vartheta)p_\phi(\vartheta) \\ &\geqq \int q_\theta(s,\vartheta|w)log\frac{p_\phi(w|s,\beta)p_\phi(s|\vartheta)p_\phi(\vartheta)}{q_\theta(s,\vartheta|w)} \\ &\geqq -D_{KL}[q_\theta(s,\vartheta|w)||p_\phi(s,\vartheta)] + \mathbb{E}_{q_\theta(s,\vartheta|w)}(log\int p_\phi(w|s,\beta)) \end{aligned} \tag{2}$$

The above equation deduces a lower bound to the marginal log likelihood, named *evidence lower bound (ELBO)*. The first term of ELBO is a regularizer that constraints the Kullback-Leibler divergence between the variational posterior distribution and the prior distribution of the latent variables. The second term of ELBO is the reconstruction loss.

[1] http://nlp.stanford.edu/projects/glove/

The variational distribution $q_\theta(s,\vartheta)$ is introduced to approximate the true posterior distribution $p_\phi(s,\vartheta)$, where

$$\begin{aligned}
p_\phi(s,\vartheta) &= p_\phi(s|\vartheta)p_\phi(\vartheta)\\
p_\phi(s_{d,n}|\vartheta_d) &= N(s_{d,n};\vartheta_d,\sigma^2_{s_{d,n}}(\vartheta_d;\phi))\\
p_\phi(\vartheta_d) &= L(\vartheta_d;0,I)\\
q_\theta(s,\vartheta) &= q_\theta(s)q_\theta(\vartheta|s)\\
q_\theta(\vartheta_d|\tilde{s}_d) &= L(\vartheta_d;\tilde{s}_d,\sigma_{\vartheta_d}(\tilde{s}_d;\theta))\\
q_\theta(s_{d,n}|w_{d,n}) &= N(s_{d,n};0,\sigma^2_{s_{d,n}}(w_{d,n};\theta))
\end{aligned}$$

To carry out the neural variational inference, we focus on parametrizing the above various distributions in ELBO with neural network, which allows the ELBO to be optimized by the back propagation method.

3.3.2. ***The Neural Parameterizing***

For the document codes, we construct the inference network to parametrize the approximate posterior $q_\theta(\vartheta|s)$, which takes input the average word codes $\tilde{s}_d = \frac{1}{|I_d|}\sum_{n\in I_d} s_{d,n}$ of the document d to output the latent variable ϑ with the variational parameters θ. We adopt the variational posterior $L(\vartheta_d;\tilde{s}_d,\sigma_d(\tilde{s}_d))$ to approximate the prior $L(\vartheta_d;0,I)$ for the sparse document codes. A zero-mean Laplace prior has the same effect as L_1 regularization. In the variational posterior, the location parameter μ_d is equal to the average word code $\tilde{s}_d$, which aims to make the document code be close to the averaging aggregation of its individual word codes. It is worth noting that although the variational posterior $q_\theta(\vartheta|s)$ is not zero-mean Laplace, most of terms in the location parameter $\tilde{s}_d$ will tend to zero for enough sparse word codes. Then, the sparse document codes ϑ_d will encourage the sparsity of word codes in turn, because we consider each word codes is generated by the Gaussian distribution with mean ϑ_d in generative process. The scale parameter σ_d is parametrized as follow:

$$\pi_\vartheta = f(\tilde{s}_d;\theta), \quad \sigma_{\vartheta_d} = Softplus(W_\sigma\pi_\vartheta + b_\sigma) \tag{3}$$

where $f(\tilde{s}_d)$ is a multilayer perceptron acting on the average word code of document d. As in [38], the Softplus is used to ensure positive scale parameters.

For the word codes, in the generative process, we consider each of them is generated by the Gaussian distribution with mean ϑ_d. We aim to make each word code $s_{d,n}$ in document d be close to the aggregation center ϑ_d. As for the variance in $p_\phi(s_{d,n})$, it is parameterized with the generative network as follow:

$$\pi_s = f(\vartheta_d;\phi), \quad \sigma^2_{s_{d,n}} = Softplus(W_\sigma\pi_s + b_\sigma) \tag{4}$$

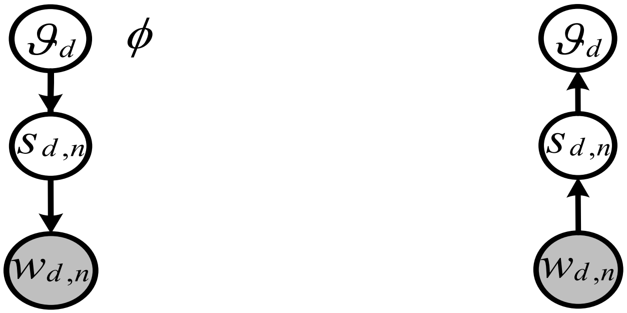
In the inference process, we consider the distribution $q_\theta(s_{d,n}|w_{d,n})$ is Gaussian with zero mean and the variance are parametrized with the inference networks:

$$\pi_s = f(w_{d,n};\theta), \quad \sigma^2_{s_{d,n}} = Softplus(W_\sigma\pi_s + b_\sigma) \tag{5}$$

As in [39], the zero-mean Laplace distribution $L(s;0,\sigma)$ is equivalent to a two-level hierarchical-Bayes model: zero-mean Gaussian priors with independent $N(s;0,\tau)$, exponentially distributed variances $e(\tau;\sigma)$. When τ approaches zero, we can induce corresponding sparse s. To achieve sparse word codes, we apply the sparse mechanism which is similar to the hierarchical Laplace prior. In our inference process, based on the zero mean Gaussian distribution, the word code $s_{d,n}$ is also zero when the variance $\sigma^2_{s_{d,n}}$ is zero.

3.3.3. ***The Reparameterization Trick***

We devote to differentiate and optimize the lower bound above with stochastic gradient decent (SGD). To reduce the variance of the stochastic gradients, we make a differentiable transformation, called reparameterization trick according to [40]. We can reparameterize the variational distribution $q(s)$ and $q(\vartheta)$ from



(a) generative model

(b) inference model

Figure 2: Generative model and inference model of NSTM with two stochastic layers.

simple samples ϵ and ε as follows:

$$\begin{aligned} s_{d,n} &\sim N(0, \sigma^2_{s_{d,n}}) \rightarrow s_{d,n} = \sigma_{s_{d,n}} \odot \epsilon, \epsilon \sim N(0, I) \\ \vartheta_d &\sim L(\tilde{s}_d, \sigma_{\vartheta_d}) \rightarrow \vartheta_d = \tilde{s}_d + \sigma_{\vartheta_d} \odot \varepsilon, \varepsilon \sim L(0, I) \end{aligned} \tag{6}$$

Through reparameterization, we can deem $s_{d,n}$ and ϑ_d as a function with the parameter $\mu_{s_{d,n}}$, $\sigma_{s_{d,n}}$, and σ_{ϑ_d} deriving from the inference networks. It allows the reconstruction error to flow through the whole network. Figure 2 presents the complete VAE generative and inference process for NSTM. Moreover, in order to achieve more interpretable document and word codes [10], we constrain s and ϑ to be non-negative, and apply the ReLU activation function after the transformation. After applying the reparameterization trick to the variational lower bound and obtain the topic dictionary β with topic dictionary neural network, we can yield:

$$\begin{aligned} -D_{KL}[q_\theta(s)||p_\phi(s)] = &\sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k=1}^{K}(\frac{1}{2} - \frac{\sigma^2_{s_{d,nk}}(\phi) + \vartheta^2_{d,k}}{2\sigma^2_{s_{d,nk}(\theta)}} - \\ &\frac{\sigma_{s_{d,nk}(\theta)}}{\sigma_{s_{d,nk}}(\phi)}) \\ -D_{KL}[q_\theta(\vartheta)||p_\phi(\vartheta)] = &\sum_{d=1}^{D}\sum_{k=1}^{K}(-ln\sigma_{\vartheta_{d,k}}(\theta) - \frac{\tilde{s}_d}{\sigma_{\vartheta_{d,k}}(\theta)} - 1) \end{aligned} \tag{7}$$

$$\begin{aligned} L(\theta, \phi) = &\mathbb{E}_{\epsilon \sim N(0,I), \varepsilon \sim L(0,I)} \sum_{d=1}^{D}\sum_{n=1}^{N}(\sum_{k=1}^{K} s_{d,nk}\beta_{nk} \\ &- w_{d,n} ln(\sum_{k=1}^{K} s_{d,nk}\beta_{nk})) - D_{KL}[q_\theta(s)||p_\phi(s)] \\ &- D_{KL}[q_\theta(\vartheta)||p_\phi(\vartheta)] \end{aligned} \tag{8}$$

3.4. Semantic Reinforcement Neural Variational Sparse Topic Model

In the previous section, we introduce our proposed model NSTM which inherits the probabilistic characteristics of the sparse topic models in neural topic model to generate more sparse and meaningful representations for documents and words. However, the method still ignores the semantic structure of words in the documents and only leverages the global co-occurrence information of words in the corpus. It is clear that the shallow structure in topic models is unable to model the sequential and contextual information of words in

Algorithm 1 Training Algorithm for NSTM

```
Require: initialize $\theta, \phi; W$
repeat
    $w^M \leftarrow$ Random mini-batch of $M$ word counts from full datasets
    $\varepsilon \leftarrow$ Random samples from noise distribution $p(\varepsilon)$
    $\epsilon \leftarrow$ Random samples from noise distribution $p(\epsilon)$
    $g \leftarrow \bigtriangledown_{\theta,\phi,W} L(\theta, \phi; w^M, \varepsilon, \epsilon)$
    $\theta, \phi, W \leftarrow$ Update parameters using SGD
until convergence
```

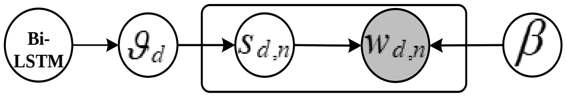


Figure 3: The graphical model of SR-NSTM

the document. Recently, Long Short Term Memory (LSTM) is effective to capture long-range dependencies between words and has been widely used in the document-level semantic encoding. Therefore, we propose Semantic Reinforcement Neural Variational Sparse Topic Model (SR-NSTM), in which the Bi-directional LSTM is incorporated to enrich the semantic feature space of texts via modeling the sequence of the words in the documents. Similar to NSTM, it follows the generative story bellow for each document:

1. Sample the topic dictionary: $\beta = f_\beta(e)$
2. Sample a document code ϑ_d =Bi-LSTM(e_d)
3. For each word n in document d:
    (a) Sample a latent variable word code $s_{d,n} \sim p_\phi(s_{d,n}|\vartheta_d)$
    (b) Sample the observed word count $w_{d,n} \sim p_{\phi_{d,n}}(w_{d,n}|s_{d,n}, \beta_n)$

The major difference between SR-NSTM and NSTM is that the document code is generated from a Bi-LSTM rather than a prior distribution which takes the sequence of the words in the document as the input. The detailed structure of Bi-LSTM encoder is below:

Word embedding layer ($e_d \in \mathbb{R}^{n_d \times 300}$): Supposing the word number of word sequence x_d^{seq} in the document d is n_d, each word in the word sequence of document d is converted into a continuous representation with a pre-trained word embedding in this layer. Here we adopt GloVe word embeddings based on Wikipedia dataset whose dimension is 300.

Bi-LSTM hidden layer : Given the input word embedding in current t word e_{dt}, the output of the forward and backward LSTM hidden layer unit at the last $t-1$ time h_{t-1}^f, and at last $t+1$ time h_{t+1}^b, we have the output of the forward and backward hidden layer units at the current word:

$$h_t^f = H(e_{dt}, h_{t-1}^f, c_{t-1}, b_{t-1})$$
$$h_t^b = H(e_{dt}, h_{t+11}^b, c_{t-1}, b_{t-1})$$

Bi-LSTM output layer ($g \in \mathbb{R}^{1\times K}$): It aims to connect forward and backward two LSTM hidden layer units at current time: $g_t = \sigma(W_h^f h_t^f + W_h^b h_t^b + b_g)$.

Average pooling layer: For the document-level semantic is relate to every word in it, we apply an average pooling on outputs of all time: $pool(g) = \sum_{t=1}^{n'_d} \frac{g_t}{n_d}$.

Document code layer ($\vartheta \in \mathbb{R}^{1\times K}$): Given the output of pooling layer, we aim to output the desired sparse document code with sparsemax: $\vartheta_d = sparsemax(pool(g))$.

To induce the sparsity in the document code, we introduce the sparsemax as the transformation of the output of the Bi-LSTM. Sparsemax can help the model to generate sparse and more interpretable document codes which only focus on several relevant topics and filters out other irrelevant topics.

3.4.1. ***Neural Variational Inference for SR-NSTM***

Similar to NSTM, we aim to maximize the probability of word count w in document under the generative process:

$$log \int p_{\phi}(w) \geqq -D_{KL}[q_{\theta}(s|w)||p_{\phi}(s|\vartheta)] + \mathbb{E}_{q_{\theta}(s|w)}(log \int p_{\phi}(w|s,\beta)) \tag{9}$$

The variational distribution $q_{\theta}(s|w)$ is introduced to approximate the true posterior distribution $p_{\phi}(s|\vartheta)$, where

$$\begin{aligned} p_{\phi}(s_{d,n}|\vartheta_d) &= N(s_{d,n};\vartheta_d,\sigma^2_{s_{d,n}}(\vartheta_d;\phi)) \\ q_{\theta}(s_{d,n}|w_{d,n}) &= N(s_{d,n};0,\sigma^2_{s_{d,n}}(w_{d,n};\theta)) \end{aligned}$$

With the same neural parameterizing method and reparameterization trick as in NSTM, after obtaining the document code ϑ with Bi-LSTM encoder and topic dictionary β with topic dictionary neural network, we can yield

$$\begin{aligned} -D_{KL}[q_{\theta}(s)||p_{\phi}(s)] = &\sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k=1}^{K}(\frac{1}{2} - \frac{\sigma^2_{s_{d,nk}}(\phi)+\vartheta^2_{d,k}}{2\sigma^2_{s_{d,nk}(\theta)}} \\ &- \frac{\sigma_{s_{d,nk}(\theta)}}{\sigma_{s_{d,nk}}(\phi)}) \end{aligned} \tag{10}$$

$$\begin{aligned} L(\theta,\phi) = \mathbb{E}_{\varepsilon\sim N(0,I)} &\sum_{d=1}^{D}\sum_{n=1}^{N}(\sum_{k=1}^{K} s_{d,nk}\beta_{nk} \\ &- w_{d,n}ln(\sum_{k=1}^{K} s_{d,nk}\beta_{nk})) - D_{KL}[q_{\theta}(s)||p_{\phi}(s)] \end{aligned} \tag{11}$$

3.5. ***Supervised Extension of NSTM***

To further demonstrate the flexibility of our model provided by the neural variational inference, we present a supervised extension of the proposed NSTM model called sNSTM, which incorporates the label information to guide the generation of the topic dictionary and representations. sNSTM model has the same neural model structure as NSTM, and the label information is introduced with the max-margin based discriminant regularization. Compared with traditional topic models, the proposed NSTM model inherits the high flexibility offered by the neural network structure. Therefore, we can derive the supervised extension of the proposed model without changing the model structure and re-deduced mathematical inference. With the training data pairs $D = \{(w_d, y_d)\}_{d=1}^{D}$, we consider the multi-class classification problem, where y_d takes value from the set $C = \{1, ..., M\}$. To include the supervision information, we define a deep neural max-margin classifier on the latent document code. Given the latent document code, we define the linear discriminant function:

$$F(y,\vartheta,\eta;w) = \eta_y^{\top}\vartheta \tag{12}$$

where $\eta_y \in \mathbb{R}^K$ is a class-specific K-dimensional parameter vector associated with class y. To connect the classifier with previous Bayesian inference, we also treat η as a latent variable. Thus, to consider the uncertainty of latent variable η, ϑ, we take a expectation over $q(\eta,\vartheta)$ and define the corresponding expected multi-class hinge loss:

$$\mathcal{R}_h(\{\vartheta_d\},\eta;\{w_d\}) = \sum_{d=1}^{D}\max_{y\in C}(\Delta l(y_d,y) - \mathbb{E}_{q(\eta,\vartheta)}[\eta_y^{\top}\Delta f_d(y)]) \tag{13}$$

where $\Delta l(y_d, y)$ is a non-negative cost function that measures how different the prediction y is from the true class label y_d, $\Delta f_d(y) = F(y, \vartheta_d) - F(y_d, \vartheta_d)$ is the difference of the feature vectors. Therefore, the supervised NSTM solves the following Reg-Bayes problem:

$$\min_{\theta,\phi,q(\eta)} \mathbb{E}_{q(\eta)}(\mathcal{L}(\theta, \phi; w)) + KL(q(\eta)||p(\eta)) + c_h\mathcal{R}_h \tag{14}$$

where $\mathcal{L}(\theta, \phi; w)$ is the original variational objective of NSTM as in Eq.8, and c_h is the positive regularization parameter.

We make a structured mean-field assumption that $q(\vartheta, \eta) = q(\vartheta)q(\eta)$. According to [41], we consider the normal distribution for η: given $p(\eta) = N(0, \sigma^2 I)$, we have $q(\eta) = N(\lambda, \sigma^2 I)$, where $\lambda = \sigma^2 \sum_{d,y} \varpi^y \mathbb{E}_q[\Delta f_d(y)]$, ϖ^y is the the Lagrange multiplier. In this case, ,the objective function can be rewritten as:

$$\min_{\theta,\phi,\lambda} \mathcal{L}(\theta, \phi; w) + \sum_{d=1}^{D} \frac{\lambda_d}{2\sigma^2_d} + \sum_{d=1}^{D} \max_{y \in C}(\Delta l(y_d, y) - \lambda_d^\top \mathbb{E}_{q(\vartheta_d)}[\Delta f_d(y)]) \tag{15}$$

Thus, λ is only related to the last two terms, and the whole objective function can be optimized with SGD. Rather than the conditions of the deep generative model to describe the inputs, we deem the labels as side information to guide the generation of topic models. Therefore we can introduce the supervision in the training process with label information without coupling with the label information and introducing extra latent variables, when compared with conditional deep generative models. This fully proves the effectiveness of our method in variation and extension.

4. Experiments

4.1. Data and Setting

To evaluate the performance of our models, we present a series of experiments below. The objectives of the experiments include: perplexity, topic coherence, classification accuracy, the quality and interpretability of extracted topics and document representations. Our evaluation is based on the four datasets: 1) **20Newsgroups**: The classic 20 newsgroups dataset, which is comprised of 18775 newsgroup articles with 20 categories, and contains 60698 unique words.[2]. 2) **Web Snippet**: The web snippet dataset, which includes 12340 Web search snippets in 8 categories.[3]. 3) **BBC**: It consists of 2225 BBC news articles from 2004-2005 with 5 classes. We only use the title and headline of each article.[4]. 4) **Biomedical**: It consists of 20000 paper titles from 20 different MeSH in BioASQ's official website.[5]. For all four datasets, we remove the stop words, words with fewer than 3 characters, and words which are mentioned less than 3 times in the corpus. Statistics on the four datasets after preprocessing is reported in Table 1.

Table 1: Statistics on the four datasets.

Dataset	Label	Docs	Words	Vocab
20NG	20	18775	135	60698
Snippet	8	12265	10.72	5581
BBC	5	2225	11.97	2453
Bio	20	19989	7.95	6887

We compare our model with follow models: 1) **LDA** [42]. A classical probabilistic topic model [6]. .We set the iteration number $n = 2000$, the Dirichlet parameter for distribution over topics $\alpha = 0.1$ and the

[2] http://www.qwone.com/ jason/20Newsgroups/

[3] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz

[4] http://mlg.ucd.ie/datasets/bbc.html

[5] http://participants-area.bioasq.org/

[6] https://pypi.python.org/pypi/lda

Dirichlet parameter for distribution over words $\eta = 0.01$. 2) **STC** [10]. A sparsity-enhanced topic model [7]. .We set the regularization constants as $\lambda = 0.2, \rho = 0.001$ and the maximum number of iterations of hierarchical sparse coding, dictionary learning as 100. 3) **DocNADE** [5]. An unsupervised neural network topic model of documents [8] . We choose the sigmoid activate function, the hidden size is 50, the learning rate is 0.01 , the bath size is 64 and the max training number is 1000. 4) **GaussianLDA** [22]. A topic model introducing word embedding [9] . We use default values for the parameters. 5) **NVDM** [43]. A neural variational document model [10] . We use default values for the parameters. 6) **AVITM** [18]. An autoencoder variational inference model for LDA [11] . We set the learning rate is 0.02, the hidden size is 100, the bath size is 200 and the max training number is 100. 7) **NSTC** [9]. A neural extension of STC model. 8) **NFTM** [6]. A neural sparse topic model [12] . We use default values for the parameters.

Our model is implemented in Python via TensorFlow. For four datasets, we utilize the pre-trained 300-dimensional word embeddings from Wikipedia by GloVe, which is fixed during training. For each out-of-vocabulary word, we sample a random vector from a normal distribution in the interval $[0, 1]$. We adopted AdaM optimizer for weight updating with an initial learning rate of $4e - 4$ for four datasets. All weight matrices are initialized with the Xavier initialization. The generative, inference, and classifier networks in supervised NSTM (sNSTM) are all implemented with three fully connected layers. We set the size of the output layer to the number of topics in generative and inference network, and to the classes number in classifier network, the size of the other two hidden layers to 250, the size of Bi-LSTM hidden layer to 250, and the regularization weight c_h in sNSTM is 10. We also perform cross-validation in training data for four datasets respectively to determine hyperparameters in our methods and baselines.

4.2. PMI and Perplexity

Point-wise Mutual Information(PMI)[44] is the most commonly used automatic evaluation index for topic semantic coherence. A higher PMI value indicates a stronger topic semantic coherence and interpretability. For a topic proportion ψ_k, it is calculated as : $PMI(\psi_k) = \frac{2}{V(V-1)} \sum_{1<i,j<V} \frac{p(w_i,w_j)}{p(w_i)p(w_j)}$,where V is the vocabulary size, $p(w_i, w_j)$ is the joint is the joint probability of words w_i and w_j co-occurring in the test document, and $p(w_i)$ is the marginal probability of word w_i appearing in a test document. We select top-10 words to calculate the average relatedness of each pair as the PMI score of each topic. Another widely used index to evaluate the generalization ability of the topic model is perplexity [42]. It is the geometric mean of word likelihood in the test documents and defined as: $perplexity = exp(-\frac{\sum_i \log p(w_i^{test})}{\sum_i N_i^{test}})$, where w_i^{test} is the word in test document i, N_i^{test} is the total word counts in document i. We show the test document PMI and perplexity on the 20NewsGroups with 50 and 200 topic numbers in Table 2.

Table 2: PMI and Perplexity on test dataset of 20NG.

Measure	K	LDA	STC	DocNADE	NVDM	AVITM	NFTM	NSTC	NSTM	SR-NSTM
PPL	50	1091	611	896	836	665	641	517	515	493
	200	1058	587	862	852	711	639	523	528	486
PMI	50	0.17	0.21	0.12	0.08	0.24	0.17	0.18	0.19	0.21
	200	0.14	0.24	0.14	0.06	0.19	0.19	0.19	0.21	0.23

We notice that topic models with sparse enhancement such as STC, NFTM, NSTC, NSTM and SR-NSTM are better than other models. It proves that the sparse enhancement can improve the quality of topics. Our proposed methods NSTM and SR-NSTM yield competitive results compared with other neural

[7] http://bigml.cs.tsinghua.edu.cn/ jun/stc.shtml/

[8] https://github.com/AYLIEN/docnade

[9] https://github.com/rajarshd/Gaussian_LDA

[10] https://github.com/ysmiao/nvdm

[11] https://github.com/akashgit/autoencoding_vi_for_topic_models

[12] https://github.com/dallascard/neural_topic_models

Table 3: Classification accuracy of different models on Web snippet and 20NG.

Dataset	**Snippet**					**20NG**				
k	**50**	**75**	**100**	**125**	**150**	**50**	**100**	**150**	**200**	**250**
LDA	0.682	0.615	0.592	0.583	0.573	0.545	0.615	0.607	0.613	0.623
STC	0.678	0.686	0.699	0.724	0.701	0.602	0.631	0.647	0.652	0.654
DocNADE	0.656	0.656	0.645	0.646	0.647	**0.682**	0.670	0.646	0.583	0.573
GLDA	0.669	0.689	0.675	0.670	0.623	0.367	0.438	0.465	0.496	0.526
NVDM	0.614	0.628	0.640	0.654	0.669	0.578	0.593	0.601	0.613	0.621
NSTC	0.734	0.756	0.791	0.793	0.789	0.634	0.671	0.682	0.690	0.72
NSTMR	0.653	0.658	0.661	0.665	0.663	0.567	0.578	0.589	0.601	0.603
NSTM	**0.792**	0.808	**0.822**	0.805	0.818	0.654	0.671	0.692	0.720	**0.740**
SR-NSTM	0.723	**0.815**	0.795	**0.841**	**0.846**	0.667	**0.691**	**0.711**	**0.723**	0.734

Table 4: Classification accuracy of different models on BBC and Biomedical.

Dataset	**BBC**					**Biomedical**				
k	**20**	**30**	**40**	**50**	**60**	**50**	**100**	**150**	**200**	**250**
LDA	0.784	0.774	0.796	0.762	0.758	0.536	0.534	0.547	0.534	0.541
STC	0.602	0.593	0.599	0.634	0.604	0.351	0.405	0.439	0.464	0.494
DocNADE	**0.793**	**0.839**	0.832	0.834	**0.819**	**0.597**	0.588	0.588	0.583	0.582
GLDA	0.609	0.566	0.573	0.564	0.567	0.482	0.515	0.497	0.483	0.513
NVDM	0.707	0.711	0.723	0.719	0.713	0.504	0.516	0.527	0.521	0.534
NSTC	0.648	0.646	0.691	0.712	0.715	0.531	0.533	0.547	0.519	0.546
NSTMR	0.597	0.601	0.603	0.611	0.607	0.473	0.497	0.511	0.519	0.523
NSTM	0.783	0.835	0.833	0.836	0.813	0.567	0.623	0.645	**0.671**	0.664
SR-NSTM	0.785	0.831	**0.835**	**0.836**	0.816	0.579	**0.629**	**0.651**	0.699	**0.671**

sparse topic methods such as NFTM and NSTC, it demonstrates that the sparse mechanism based on explicitly sparse prior parameterized by neural networks is helpful to generate more coherent topics compared with pure feed-forward neural networks. SR-NSTM with semantic reinforcement shows a significant improvement compared with NSTM. It demonstrates that our proposed method can extract meaningful topics with improved flexibility in extension and variation, and improve the learning of topics with semantic reinforcement for enriching the feature space of semantic information.

4.3. Classification Accuracy

To evaluate the effectiveness of the representation of documents learned by NSTM and SR-NSTM, we further perform text classification tasks on web snippet, 20NG, BBC and Biomedical using the document codes learned by topic models as the feature representation in a multi-class SVM. We make the partition of training and testing as previous methods [10, 25]. On web snippet, we utilize 80% of documents for training and 20% for testing. On 20NG, BBC and Biomedical, we keep 60% of documents for training and 40% for testing. Same as previous methods, we use all the data to learn the parameters of unsupervised methods, and perform cross-validation in the training documents to select hyper-parameters for regularization and classifier. Table 3 and Table 4 report the classification accuracy under different methods with different settings on the number of topics among four datasets.

It clearly denotes that: 1) Our model generally yields the best performance overall datasets, especially in short text datasets. This is because our model can generate sparse and distinct document codes with the sparse prior. Additionally, the introduced word embeddings and deep semantic structure can improve the overall performance further, thus making a better performance of NSTM and SR-NSTM. 2) For NSTM and NSTMR, in which the topic dictionary is randomly initialized without introducing word embeddings, we can see that NSTM outperforms NSTMR in all datasets. It further proves the efficiency of word embeddings

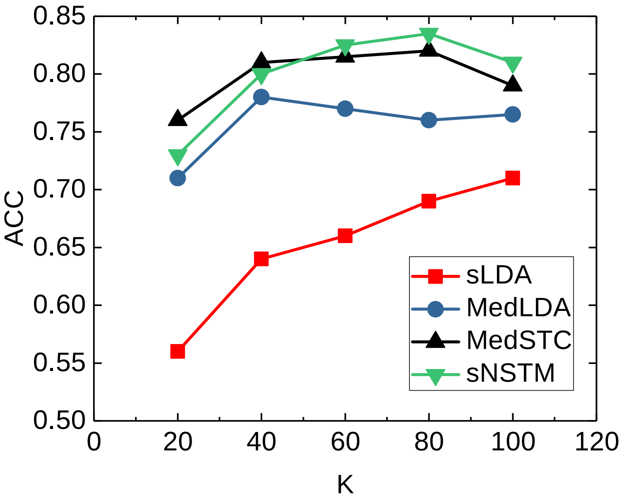
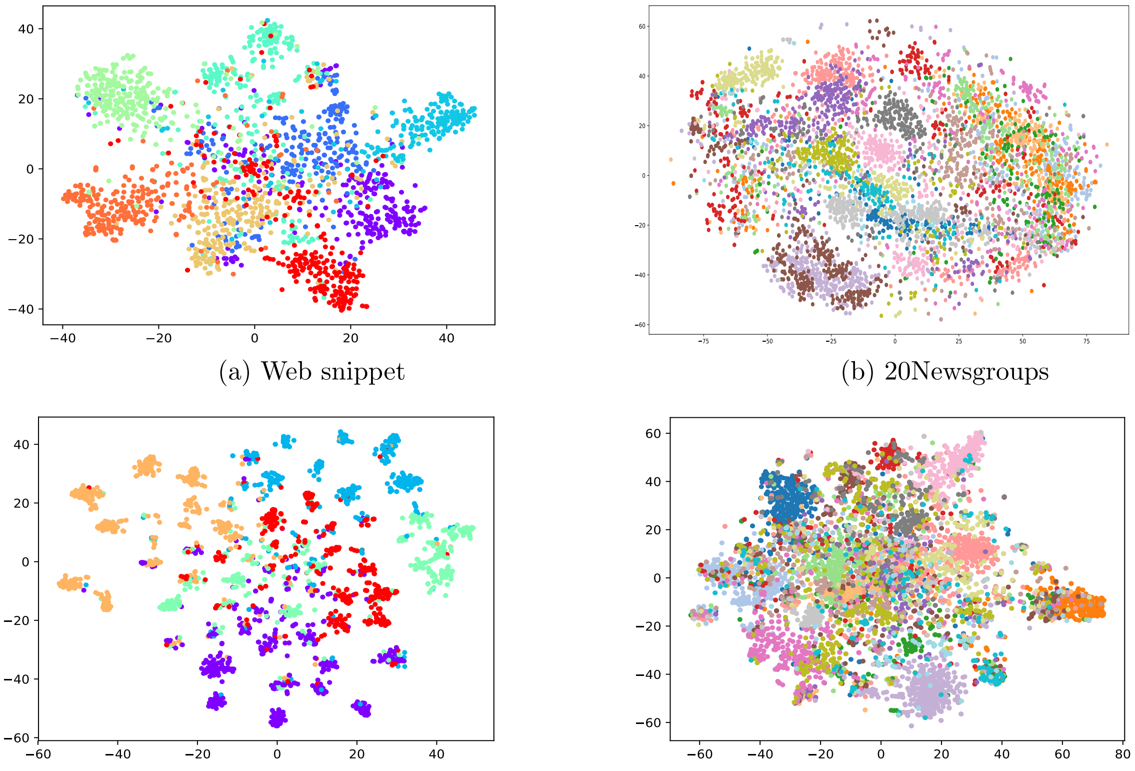


Figure 4: Classification accuracy on supervised models.



(a) Web snippet

(b) 20Newsgroups

(c) BBC

(d) Biomedical

Figure 5: t-SNE projection of the estimated document codes.

Table 5: The ablation study on Web snippet and 20NG.

Dataset	**Snippet**					**20NG**				
k	**50**	**75**	**100**	**125**	**150**	**50**	**100**	**150**	**200**	**250**
W/O NN	0.677	0.687	0.700	0.723	0.702	0.601	0.633	0.649	0.651	0.655
W/O Sparse	0.708	0.711	0.724	0.720	0.712	0.506	0.518	0.530	0.551	0.543
W/O Bi-LSTM	**0.792**	0.808	**0.822**	0.805	0.818	0.654	0.671	0.692	0.720	**0.740**
Full Model	0.723	**0.815**	0.795	**0.841**	**0.846**	0.667	**0.691**	**0.711**	**0.723**	0.734

Table 6: The topics discovered by NSTM

.

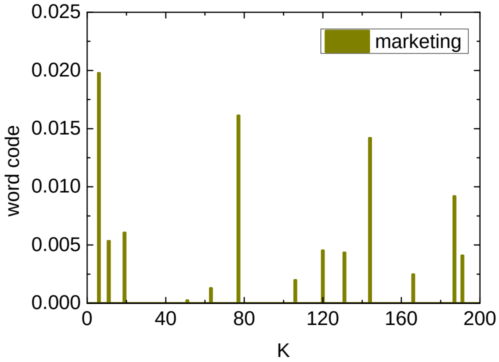
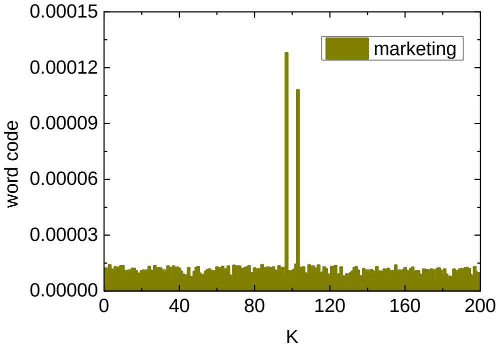
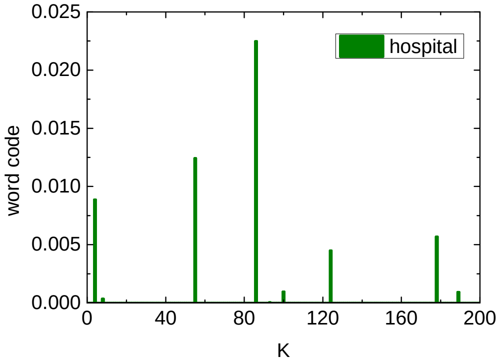
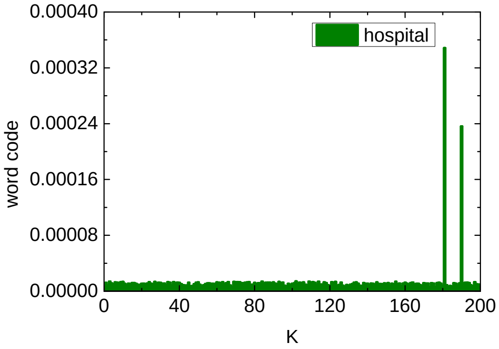
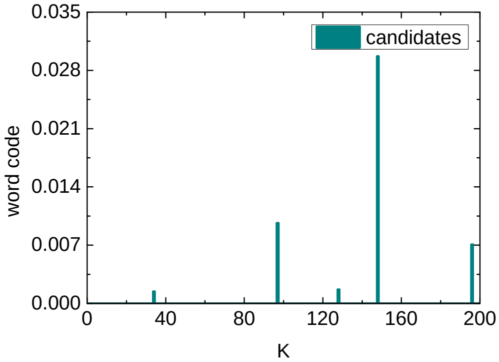
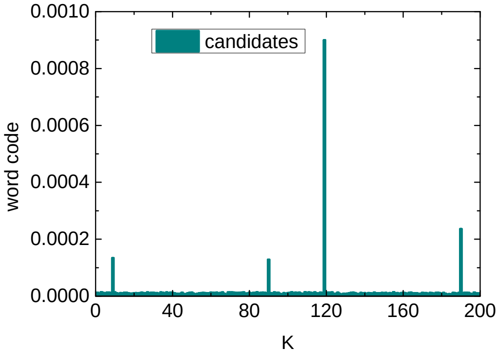
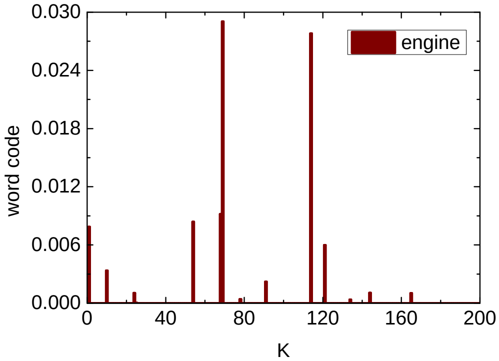
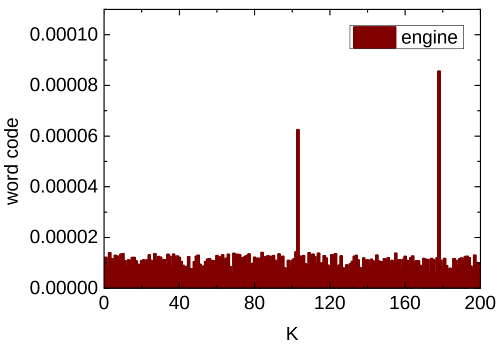
Category	**Topic**
Business	T67: investing ratneshwar investments investment investors invest equity niddk income 0.14999
	T133: products source product quality premium csail content manufacture socialsciences 0.13412
	T144: development serv ecommerce develope innovation developers business market projects 0.12130
Computers	T112: firefox mozilla netscape macintosh linux windows adobe verizon zdnet 0.13780
	T118: systems system control security controls remote automatic monitoring automation 0.13361
	T121: msn yahoo firefox aol gmail java algorithm algorithms signonsandiego 0.12191
Engineering	T90: factory inc steel searchsmb chrome ford socialsciences wieeless ltd 0.12512
	T100: device cancertopics devices modem cable died wireless semiconductor connection 0.11989
	T114: gasoline diesel fuel petrol engines engine emissions gas combustion 0.11213
Health	T86: hospital webobjects home nutritionsource clinic homes nursing emedicinehealth center 0.18371
	T124: disease cancer lung flu cancers infection influenza arthritis infections 0.17572
	T142: vitamins foods herbal diet alcohol supplements vitamin oils nutritional 0.16215

in improving document representations. 3) For GLDA which introduces word embeddings as well, it yields much poor performance in 20NG and Bio whose category numbers are large. It turns out that the document topic proportions learned by it are almost uniform and non-discriminative. Similarly, it happened for models without considering the sparsity such as LDA and DocNADE. The accuracy decays when the topic number increasing. On the contrary, our model along with STC can learn sparse and semantic enriched representations for addressing the sparseness issue. 4) As for sparse enhanced model STC, it performs considerably inferior to all other methods in BBC and biomedical. To learn interpretable document codes, STC tends to remain high frequency words and ignore the low frequency ones in the whole dataset during learning. However, in these two datasets, we notice that the word frequency distribution is nearly uniform, resulting in poor learning of STC. In our model, the introduced word embeddings can improve the word frequency distribution, thus benefit the performance. 5) Compared with NSTC which only uses the feed-forward neural networks to model the generative process, our model NSTCM and SR-NSTCM have better performance due to model the generative process with sparse prior distribution explicitly and capture the context information with Bi-LSTM.

We also perform an ablation study on our method to verify the effectiveness of each module. We compare our model with its variants by removing one of the components neural network parameterization, sparse distribution and Bi-LSTM encoder respectively, as shown in Table 5. From which we can see that each component makes a certain contribution to the overall performance. In the case of removing the neural network parameterization (W/O NN) module, our model consists of the Bi-LSTM encoder and feed forward neural networks, in which it degenerates into the context information enhanced NSTC model. As in the case of removing the spare distribution (W/O Sparse), our model degenerates into the context information enhanced NVDM model. In the case of removing Bi-LSTM encoder (W/O Bi-LSTM), the model declines into the NSTM model. We can see that missing the sparse distribution has a more significant negative influence than missing the Bi-LSTM based encoder module, illustrating the relative effectiveness of the sparse prior in improving the discrimination of latent representations.

In Figure 4, we present the results of supervised topic models on 20Newsgroups. We compare our

Table 7: The word codes of representative words for different categories discovered by NSTM and LDA.

Category	NSTM	LDA
Business		
Health		
Politics-society		
Engineer		

methods with the following supervised methods: 1) **sLDA** [45]. Supervised extension of LDA which models the class labels. 2) **MedLDA** [41]. Integrating max-margin regularization with LDA. 3) **MedSTC** [10]. Integrating max-margin regularization with STC. Generally, sNSTM and MedSTC outperform MedLDA and sLDA for their sparse property. sNSTM, MedSTC, MedLDA have better results than sLDA for the discriminative ability provided by max-margin posterior regularization. As for sNSTM and MedSTC, the feature representation ability provided by deep structural and extra word semantic information makes its better performance.

4.4. Characteristics of Code Representation

In this part, we quantitatively investigate the word codes and documents codes learned by our model NSTM.

Word code: We compute the average word code as [10]. Table 7 shows the average word codes of some representative words learned by NSTM and LDA in 4 categories of web snippet. For each category, we also present the topics learned by NSTM in Table 6. We list top-9 words according to their probabilities under each topic, and top-4 topics according to their PMI score in top-15 words. In Table 7, the results illustrate that the codes discovered by NSTM are apparently much sparser than those discovered by LDA. It tends to focus on a narrow spectrum of topics and obtains discriminative and sparse representations of words. In contrast, LDA generates word codes with many non-zeros, leading to a confusing topic distribution. Besides, in NSTM, it is clear that each non-zero element in the word codes represents the topical meaning of words in the corresponding position. The weights of these elements express their relationship with the topics. Noticed that there are words (e.g. candidates) having only a small range of topical meanings, indicating a narrow usage of those terms. While other words (e.g. marketing) tend to have a broad spectrum of topical meanings, denoting a general usage of those terms.

Document code: To demonstrate the quality of the learned representations by our model, we produce a t-SNE projection with for the document codes of two datasets learned by our model in Figure 5. For 20newsgroups and Biomedical, we sample 30% of the whole document codes. As for BBC, we present the whole document codes. It is obvious to see that all documents are clustered into distinct categories, which is equal to the ground truth number of categories in the four datasets. It proves the semantic effect of the document codes learned by our model. We can also notice that the data sets with fewer categories (web snippet, BBC) have better results of clustering than the other two datasets with more classes.

5. Conclusion

In this paper, we propose novel neural sparse topic modeling approaches, which explicitly model the probabilistic mixtures in sparse topic models with neural sparse prior, focusing on generating meaningful, sparse and explainable representations for texts. The Bi-LSTM is further adopted in our method to capture the sequential structure of words in the documents and enrich the semantic feature space. Moreover, we incorporate the max-margin posterior in our methods to utilize the label information in supervised tasks without extra latent variables, which shows the flexibility of our proposed methods. The evaluation results demonstrate the effectiveness of our models. Future work can include introducing other available information such as common knowledge to further improve the performance, since we only consider the contextual information and label information in this paper.

Acknowledgments

This research is supported by the CSC Scholarship offered by China Scholarship Council. This work is also partially supported by the Academy of Finland (grants 336033, 315896), Business Finland (grant 884/31/2018), and EU H2020 (grant 101016775).

References

[1] Y. Kim, Convolutional neural networks for sentence classification, in: EMNLP, 2014, pp. 1746–1751.

[2] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[3] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: ICLR, 2017.

[4] Z. Cao, S. Li, Y. Liu, W. Li, H. Ji, A novel neural topic model and its supervised extension., in: AAAI, 2015, pp. 2210–2216.

[5] H. Larochelle, S. Lauly, A neural autoregressive topic model, in: Advances in Neural Information Processing Systems, 2012, pp. 2708–2716.

[6] D. Card, C. Tan, N. A. Smith, A neural framework for generalized topic models, arXiv preprint arXiv:1705.09296.

[7] T. Lin, Z. Hu, X. Guo, Sparsemax and relaxed wasserstein for topic sparsity, Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining.

[8] P. Tiwari, H. Zhu, H. M. Pandey, Dapath: Distance-aware knowledge graph reasoning based on deep reinforcement learning, Neural Networks 135 1–12.

[9] M. Peng, Q. Xie, H. Wang, Y. Zhang, X. Zhang, J. Huang, G. Tian, Neural sparse topical coding, in: ACL, 2018.

[10] J. Zhu, E. P. Xing, Sparse topical coding, in: Proceedings of the 27th International Conference on Uncertainty in Artificial Intelligence, Vol. 831, 2011, p. 838.

[11] 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 conference on World wide web, ACM, 2014, pp. 539–550.

[12] T. Lin, S. Zhang, H. Cheng, Understanding sparse topical structure of short text via stochastic variational-gibbs inference, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, ACM, 2016, pp. 407–416.

[13] J. Eisenstein, A. Ahmed, E. P. Xing, Sparse additive generative models of text.

[14] X. Chen, M. Zhou, L. Carin, The contextual focused topic model, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 96–104.

[15] S. Williamson, C. Wang, K. A. Heller, D. M. Blei, The ibp compound dirichlet process and its application to focused topic modeling, in: Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 1151–1158.

[16] M. Peng, Q. Xie, H. Wang, Y. Zhang, G. Tian, Bayesian sparse topical coding, IEEE Transactions on Knowledge and Data Engineering 31 (2019) 1080–1093.

[17] F. Tian, B. Gao, D. He, T.-Y. Liu, Sentence level recurrent topic model: Letting topics speak for themselves, ArXiv abs/1604.02038.

[18] A. Srivastava, C. Sutton, Neural variational inference for topic models, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017.

[19] Y. Miao, E. Grefenstette, P. Blunsom, Discovering discrete latent topics with neural variational inference, arXiv preprint arXiv:1706.00359.

[20] Y. Cong, B. Chen, H. Liu, M. Zhou, Deep latent dirichlet allocation with topic-layer-adaptive stochastic gradient riemannian mcmc, ArXiv abs/1706.01724.

[21] H. Zhang, B. Chen, D. Guo, M. Zhou, Whai: Weibull hybrid autoencoding inference for deep topic modeling, in: ICLR, 2018.

[22] R. Das, M. Zaheer, C. Dyer, Gaussian lda for topic models with word embeddings., in: ACL (1), 2015, pp. 795–804.

[23] W. Hu, J. Tsujii, A latent concept topic model for robust topic inference using word embeddings, in: The 54th Annual Meeting of the Association for Computational Linguistics, 2016, p. 380.

[24] D. Q. Nguyen, R. Billingsley, L. Du, M. Johnson, Improving topic models with latent feature word representations, Transactions of the Association for Computational Linguistics 3 (2015) 299–313.

[25] S. Li, T.-S. Chua, J. Zhu, C. Miao, Generative topic embedding: a continuous representation of documents, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, 2016, pp. 666–675.

[26] G. Xun, V. Gopalakrishnan, F. Ma, Y. Li, J. Gao, A. Zhang, Topic discovery for short texts using word embeddings, 2016 IEEE 16th International Conference on Data Mining (ICDM) (2016) 1299–1304.

[27] G. Xun, Y. Li, W. X. Zhao, J. Gao, A. Zhang, A correlated topic model using word embeddings, in: IJCAI, 2017.

[28] N. Batmanghelich, A. Saeedi, K. Narasimhan, S. J. Gershman, Nonparametric spherical topic modeling with word embeddings, Proceedings of the conference. Association for Computational Linguistics. Meeting 2016 (2016) 537–542.

[29] S. Bunk, R. Krestel, Welda: Enhancing topic models by incorporating local word context, in: JCDL '18, 2018.

[30] H. Xu, W. Wang, W. Liu, L. Carin, Distilled wasserstein learning for word embedding and topic modeling, in: NeurIPS, 2018.

[31] A. B. Dieng, F. J. R. Ruiz, D. M. Blei, Topic modeling in embedding spaces, ArXiv abs/1907.04907.

[32] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (3) (2018) 31–57.

[33] Y. Chai, W. Li, Towards deep learning interpretability: A topic modeling approach.

[34] A. Panigrahi, H. V. Simhadri, C. Bhattacharyya, Word2sense: sparse interpretable word embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5692–5705.

[35] B. Peters, V. Niculae, A. F. Martins, Interpretable structure induction via sparse attention, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018, pp. 365–367.

[36] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.

[37] A. Martins, R. Astudillo, From softmax to sparsemax: A sparse model of attention and multi-label classification, in: International Conference on Machine Learning, 2016, pp. 1614–1623.

[38] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, O. Winther, Ladder variational autoencoders, in: Advances in neural information processing systems, 2016, pp. 3738–3746.

[39] M. A. Figueiredo, et al., Adaptive sparseness for supervised learning, IEEE transactions on pattern analysis and machine intelligence 25 (9) (2003) 1150–1159.

[40] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114.

[41] J. Zhu, A. Ahmed, E. P. Xing, Medlda: maximum margin supervised topic models, Journal of Machine Learning Research 13 (Aug) (2012) 2237–2278.

[42] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.

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

[44] D. Newman, J. H. Lau, K. Grieser, T. Baldwin, Automatic evaluation of topic coherence, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 100–108.

[45] J. D. Mcauliffe, D. M. Blei, Supervised topic models, in: Advances in neural information processing systems, 2008, pp. 121–128.