NEURAL VARIATIONAL SPARSE TOPIC MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Effectively inferring discriminative and coherent latent topics of short texts is a critical task for many real world applications. Nevertheless, the task has been proven to be a great challenge due to the data sparsity problem induced by the characteristics of short texts. A large number of topic models have been proposed to deal with the data sparsity problem in short texts. However, the complex and rigorous inference algorithm become a bottleneck for these traditional models to rapidly explore variations. In this paper, we propose a novel model called Neural Variational Sparse Topic Model (NVSTM) based on a popular model sparsity-enhanced topic model named Sparse Topical Coding (STC). In the model, the auxiliary word embeddings are utilized to improve the generation of representations. The Variational Autoencoder (VAE) approach is applied to inference the model efficiently, which makes the model easy to explore extensions for its black-box inference process. Experimental results on Web Snippets and 20NewsGroups datasets show the effectiveness and efficiency of the model.

1 INTRODUCTION

With the great popularity of social networks and Q&A networks, short texts have been the prevalent information format on the Internet. Uncovering latent topics from huge volume of short texts is fundamental to many real world applications such as emergencies detection (Sakaki et al., 2010), user interest modeling (Sasaki et al., 2014), and automatic query-reply (Peng et al., 2016). However, short texts are characteristic of short document length, a very large vocabulary, a broad range of topics, and snarled noise, leading to much sparse word co-occurrence information. Thus, the task has been proven to be a great challenge.

Based on the observation that the latent representations to be learned are highly sparse (i.e. each document focuses on a few topics, and each topic focuses on a few words), many works have been devoted to tackle the data sparsity problem of standard topic models in modeling short texts. One strategy is introducing the sparse prior into classical probabilistic topic models, to achieve sparse representation in the document-topic and topic-term distributions (Lin et al., 2014; Williamson et al., 2010; Blei et al., 2003a). Another strategy is aiming at extracting focused topics or focused words in text by imposing sparsity inducing regularizers on non-probabilistic coding or matrix factorization (Zhu & Xing, 2011; Deerwester et al., 1990; Heiler & Schnörr, 2006). Both strategies yield a better performance on short texts, leveraging the highly sparse information. Yet these models general rely on computationally expensive inference procedures like Markov Chain Monte Carlo, which makes them hard to rapidly explore extensions. Even if there are only minor changes to model assumptions, the inference algorithms need to be re-deduced. It is mathematic challenging and time consuming, thus limiting the applications and extensions of these models.

With the advent of deep neural networks, the neural variational inference has emerged as a powerful approach to unsupervised learning of complicated distributions (Kingma & Welling, 2013; Rezende et al., 2014; Mnih & Gregor, 2014). It approximates the posterior of a generative model with a variational distribution parameterized by a neural network, which allows back-propagation based function approximations in generative models. The variational autoencoder (VAE) (Kingma & Welling, 2013), one of the most popular deep generative models, has shown great promise in modeling complicated data. It is a particular de facto choice for topic models to be trained with back-propagation without complicated mathematical inference.

Motivated by the promising potential of VAE in building generative models with black-box inference process, we propose a Neural Variational Sparse Topic Model (NVSTM) based on a popular model sparsity-enhanced topic model STC for short texts. The model is parameterized with neural networks and trained with VAE, which is easy and natural to be employed in other tasks. It still follows the probabilistic characteristics of STC. Thus, the model inherits the advantages of both topic models and deep neural networks. Additionally, we exploit the auxiliary word embeddings to improve the generation of short text representations.

To summarize, the main contributions of this paper are as follows:

1. We propose a novel Neural Variational Sparse Topic Model (NVSTM) to learn sparse representations of short texts. The VAE is utilized to inference the model effectively. To our best knowledge, this is the first work to incorporate a sparsity-enhanced topic model with the VAE inference method.
2. The general word semantic information is introduced to improve the sparse representations of short texts via word embeddings.
3. We conduct experiments on 20 Newsgroups and Web Snippet datasets. Experimental results on demonstrate our model's superiority in topic coherence and text classification accuracy.

The rest of this paper is organized as follows. First, we reviews related work. Then, we present the details of the proposed NVSTM, followed by the experimental results. Finally, we draw our conclusions.

2 Related Work

Topic models. Traditional topic models and their extensions (Archambeau et al., 2015; Blei et al., 2003b; Mcauliffe & Blei, 2008) have been widely applied to many tasks such as information retrieval, document classification, and so on. These models work well on long texts which have abundant word co-occurrence information for learning, but get stuck in short texts. There have been many efforts to address the data sparsity problem of short texts. To achieve sparse representations in the document-topic and topic-term distributions, Williamson et al. (2010) introduced a Spike and Slab prior to model the sparsity in finite and infinite latent topic structures of text. Similarly, Lin et al. (2014) proposed a dual-sparse topic model that addresses the sparsity in both the topic mixtures and the word usage. These models are inspired by the effect of the variation of the Dirichlet prior on the probabilistic topic models. There are also some non-probabilistic sparse topic models aiming at extracting focused topics and words by imposing various sparsity constraints. Heiler & Schnörr (2006) formalized topic modeling as a problem of minimizing loss function regularized by lasso. Subsequently, Zhu & Xing (2011) presented sparse topical coding (STC) by utilizing the Laplacian prior to directly control the sparsity of inferred representations. However, over complicated inference procedure of these sparse topic models has limited their applications and extensions.

Neural Variational Inference for topic models. Neural variational inference is capable of approximating the posterior of a generative model with a variational distribution parameterized by a neural network (Kingma & Welling, 2013; Rezende et al., 2014; Mnih & Gregor, 2014). The variational autoencoder (VAE), as one of the most popular neural variational inference approach, has shown great promise in building generative models with black-box inference process (Kingma & Welling, 2013). To break the bottleneck of over complicated inference procedure in topic models, there are previous efforts devoting to inference topic models with VAE. Srivastava & Sutton (2017) presents auto-encoding variational Bayes (AEVB) based inference method for latent Dirichlet allocation (LDA), tackling the problems caused by the Dirichlet prior and component collapsing in AEVB. Miao et al. (2017) presents alternative neural approaches in topic modeling by providing parameterized distributions over topics. It allows training the topic model via back-propagation under the framework of neural variational inference. Card et al. (2017) combines certain motivating ideas behind variations on topic models with modern techniques for variational inference to produce a flexible framework for topic modeling that allows for rapid exploration of different models. Nevertheless, aforementioned works are based on traditional LDA, thus bypass the sparsity problem of short texts. Drawing inspiration from the analysis, we propose a novel neural variational sparse topic model NVSTM based on VAE for short texts, which combines the merits of neural networks and sparsity-enhanced topic models.

3 Neural Variational Sparse Topic Model

In this section, we start from describing Sparse Topical Coding (STC). Based on it, we further propose Neural Variational Sparse Topic Model (NVSTM). Later, we focus on the discussion of the inference process for NVSTM.

3.1 Sparse Topical Coding

In standard STC, each document and each word is represented as a low-dimensional code in topic space. Based on the topic dictionary β with K topic bases sampled from a uniform distribution, the generative process is described as follows:

1. Sample the document code θ from a prior $p(\theta) \sim Laplace(\lambda_1)$.
2. For each observed word n:
   (a) Sample the word code s_n from a conditional distribution $p(s_n|\theta) \sim supergaussian(\theta, \lambda_2)$.
   (b) Sample the observed word count w_n from a distribution $p(w_n|{s_n}^T\beta_n) \sim Poisson({s_n}^T\beta_n)$

STC reconstructs each observed word count from a linear combination of a set of topic bases, where the word code is utilized as the coefficient vector. According to the above generative process, we have the joint distribution:

$$p(\theta, s, w|\beta) = \prod_d p(\theta_d) \prod_n p(s_n|theta)p(w_n|s_n, \beta) \quad (1)$$

To simplify the calculation, the document code can be collapsed and later obtained via an aggregation of the individual word codes of all its terms. Although STC has closed form coordinate descent equations for parameters (θ, s, β), it is inflexible for its complex inference process.

3.2 Generative Process for NVSTM

To address the aforementioned issue, we introduce black box inference methods into STC. We present NVSTM based on VAE via the reparameterization trick and introduces word embeddings. In our methods, we also collapse the document code to simplify the model structure as well as STC. Analogous to the generative process in STC, our model follows the generative story below for each document d:

1. For each word n in document d:
   (a) Sample a latent variable word code $s_n \sim Laplace(0, b_n)$.
   (b) Sample the observed word count w_n from $p(w_n|{s_n}^T\beta_n) \sim Poisson({s_n}^T\beta_n)$

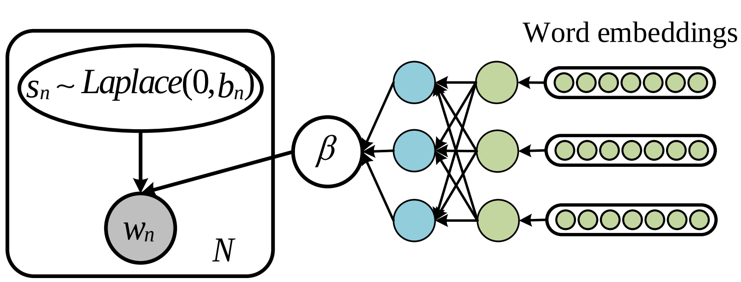


Figure 1: The graphical model of NVSTM.

The graphical representation of NVSTM is depicted in Figure 1. In the above generative process, each word code vector is generated from Laplacian distributions to achieve word-level sparse representations, and the observed word count is sampled from Poisson distribution. Different from traditional STC, we replace the uniform distribution of the topic dictionary with a topic dictionary neural network. In the topic dictionary neural network, we introduce the word semantic information via word embeddings to enrich our model for short texts. The topic dictionary neural network is comprised of two layers:

Word embedding layer ($WE \in \mathbb{R}^{N \times 300}$): Supposing the word number of the vocabulary is N, this layer devotes to transform each word to a distributed embedding representation. Here, we adopt the pre-trained embeddings by Glob2Vec based on a large Wikipedia dataset [1]. Given a word embedding matrix WE, we map each word to a 300-dimensional embedding vector, which can capture subtle semantic relationships between words.

Topic dictionary layer ($\beta \in \mathbb{R}^{N \times K}$): This layer aims at converting WE to a topic dictionary similar to the one in STC.

$$\beta(n) = relu(WE \times W) \tag{2}$$

where $W \in \mathbb{R}^{300 \times K}$ is the weight matrix between the word embedding layer and the topic dictionary layer. To conform to the framework of STC, we also make a simplex projection among the output of topic dictionary neural network.

Based the above generative process, the traditional variational inference for the model is to minimize the follow optimization problem:

$$L(\gamma|\beta) = D_{KL}[q(s|\gamma)||p(s|w, \beta, b)] - logp(w|s, \beta) \tag{3}$$

where $q(s|\gamma)$ is approximate variational posterior, and γ is the variational parameter.

3.3 Variational Autoencoder for NVSTM

In this paper, we employ the VAE to carry out neural variational inference for our model. Variational Autoencoder (VAE) is one of the most popular deep generative network. It is a black-box variational method which bridges the conceptual and language gap of neural networks and probability generative models. From neural network perspective, a variational autoencoder consists of an encoder network, a decoder network, and a loss function. In our model, the encoder network is to parametrize the approximate posterior $q_\theta(s|w)$, which takes input as the observed word count to output the latent variable s with the viriational parameters θ. The decoder network outputs the observed data w with given s and the generative parameters phi, which is denoted as $p_\phi(w|s)$. Based on VAE, we rewrite the ELBO as:

$$L(\theta, \phi|\beta) = -D_{KL}[q_\theta(s|w)||p(s)] + E_{q_\theta(s|w)}(logp_\phi(w|s, \beta)) \tag{4}$$

The first term is a regularizer that constraints the Kullback-Leibler divergence between the encoder's distribution distribution and the prior of the latent variables. The second term is the reconstruction loss, which encourages the decoder to reconstruct the data in minimum cost.

3.4 The Reparameterization Trick and Optimizing

We devote to differentiate and optimize the lower bound above with stochastic gradient decent (SGD). However, the gradient of the lower bound is tricky since the error is unable to back propagate through a random drawn variable s, which is a non-continuous and has no gradient. Similar to the standard VAE, we make a differentiable transformation, called reparameterization trick. We approximate s with an auxiliary noise variable $\varepsilon \sim U(0,1)$:

$$s_n \sim Laplace(0, b_n) \rightarrow s_n = -b_n sign(\varepsilon) ln(1 - 2|\varepsilon|), \varepsilon \sim U(0,1) \tag{5}$$

Through reparametrization, we can take s as a function with the parameter b deriving from the encoder network. It allows the reconstruction error to flow through the whole network. Figure 2 presents the complete VAE inference process for NVSTM. After apply the reparameterization trick to the variational lower bound, we can yield

$$L(\Theta) = \sum_{i=1}^{d}\sum_{j=1}^{N}(1 + log2b_{ij}) + \sum_{i=1}^{d}\frac{1}{N}\sum_{j=1}^{N} logp(w_{ij}|s_{ij}, \beta_j) \tag{6}$$

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

where $s_n = -b_n sign(\varepsilon) ln(1 - 2|\varepsilon|), \varepsilon \sim U(0,1)$, and Θ represents the set of all the model. As explained above, the decoding term $logp(w_{ij}|s_{ij}, \beta_i)$ is the Poisson distribution, and β is generated by a topic dictionary neural network. After the differentiable transformation, the variation objective function can be computed in closed form and efficiently solved with SGD. The detailed algorithm is shown in Algorithm 1.

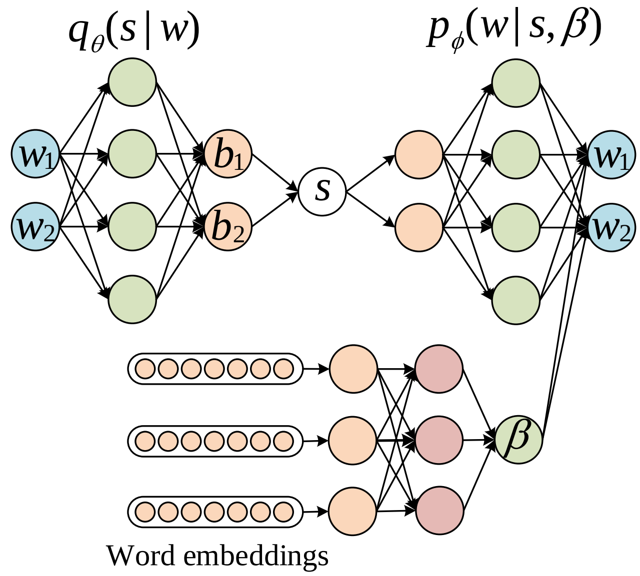


Figure 2: The VAE inference for NVSTM, the VAE is implemented as a feedforward neural network.

Algorithm 1 Training Algorithm for NVSTM

Input: initialize θ, ϕ, W
1: **repeat**
2: $w^M \leftarrow$ Random mini-batch of M word counts from full datasets
3: $\varepsilon \leftarrow$ Random samples from noise distribution $p(\varepsilon)$
4: $g \leftarrow \bigtriangledown_{\theta,\phi,W} L(\theta, \phi; w^M, \varepsilon)$
5: $\theta, \phi, W \leftarrow$ Update parameters using SGD
6: **until** convergence

4 Experiments

4.1 Data and Setting

To evaluate the performance of our model, we present a series of experiments below. The objectives of the experiments include: (1) the qualitative evaluations: classification accuracy of documents and sparse ratio of latent representations; (2) the qualitative inspection: the quality of extracted topics. Our evaluation is based on two datasets:

- **20Newsgroups**: The classic 20 newsgroups dataset, which is comprised of 18775 newsgroup articles with 20 categories, and contains 60698 unique words. [2].
- **Web Snippet**: The web snippet dataset, which includes 12340 Web search snippets in 8 categories. We remove the words with fewer than 3 characters or whose document frequency less than 3 in the dataset. After the preprocessing, it contains 5581 unique words.[3].

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

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

Table 1: Statistics on the two datasets. Label: the number of ground truth labels or categories; Docs: the total number of documents; Words: average number of words per document.

Dataset	Label	Docs	Words	Vocabulary
Web Snippet	8	12265	10.72	5581
20Newsgroups	20	18775	135	60698

Statistics on the two datasets after preprocessing is reported in Table 1. Although documents in Web Snippet are much shorter, its vocabulary is much smaller than 20Newsgroups. Therefore, the data sparsity of both datasets is in the same degree.

We compare our model with three topic models:

- **LDA** (Blei et al., 2003b). A classical probabilistic topic model. We use the open source LDA implemented by Gibbs sampling. [4]
- **STC** (Zhu & Xing, 2011). A sparsity-enhanced topic model which has been proven to perform better than many existing models. We adopt the implementation of STC released by its authors. [5]
- **NTM** (Cao et al., 2015). A recently proposed neural network based topic model, which has been reported to outperform the Replicated Softmax model. We employ the implement released by its authors. [6]

Our model is implemented in Python via TensorFlow. For both 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 interval $[0, 1]$. We adopted ADAM optimizer for weight updating with an initial learning rate of $1e-4$ for both dataset. All weight matrices are initialized with a uniform distribution in interval $[0, 1e-5]$. In practice, we found that our model is stable with the size of hidden layer, and set it to 500.

4.2 Characteristics of Code Representation

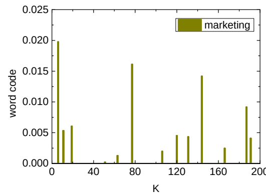
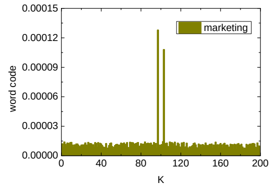
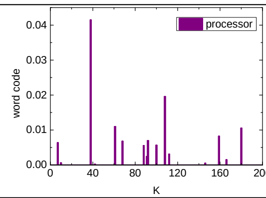
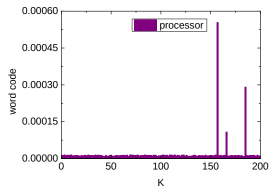
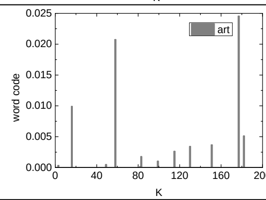
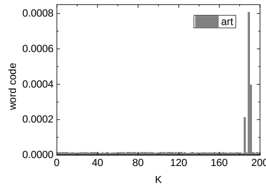
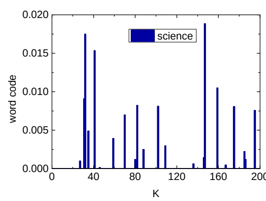
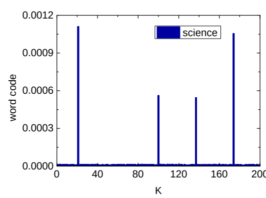
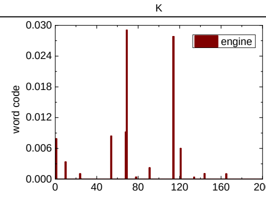
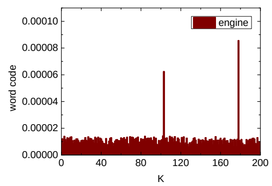
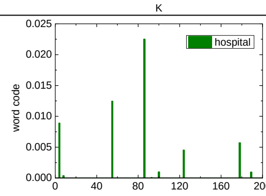
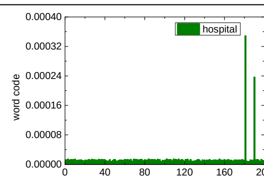
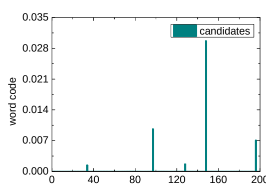
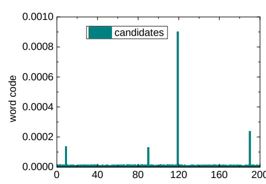
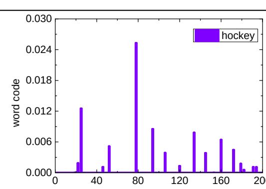
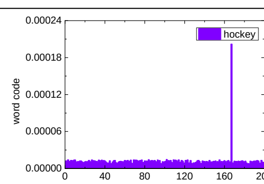
In this part, we quantitatively investigate the word codes and documents codes learned by our model.

Word code: We compute the average word code as $\overline{s_n} = \frac{1}{D_n}\sum_{d \in D_n} s_{d,n}$ over all documents that word n appears in. Table 2 shows the average word codes of some representative words learned by NVSTM and LDA in 8 categories of web snippet. To save limited space, we omit the results of STC and DocNADE, which are similar to the results of NVSTM and LDA respectively. For each category, we also present the topics learned by NVSTM. We list top-k words according to their probabilities under each topic. The results illustrate that the codes discovered by NVSTM are apparently much sparser than those discovered by LDA. It tends to focus on narrow spectrum of topics and obtains discriminative and sparse representations of word. In contrast, LDA generates word codes with many non-zeros due to the data sparsity, leading to a confused topic distribution. Besides, in NVSTM, it is clear that each non-zero element in the word codes represents the topical meaning of words in corresponding position. The weights of these elements express their relationship with the topics. Noticed that there are words (e.g. candidates and engine) have only a small range of topical meanings, indicating a narrow usage of those terms. While other words (e.g. hockey and science) tend to have a broad spectrum of topical meanings, denoting a general usage of those terms.

Document code: Here, each document code is calculated as $\theta_d = \sum_{d=1}^{D}\sum_{n=1}^{N_d} s_{d,nk}\,\beta_{kn} / \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K} s_{d,nk}\,\beta_{kn}$ (Bai et al., 2013). Note that in LDA, NTM and NVSTM, the inferred latent representations of documents are admixture proportions, while in STC, the latent representations of documents are unnormalized code vectors. To demonstrate the quality of the learned representations by our model, we produce a t-SNE projection with 3000 perplexity for the document code of each document learned by our model with 200 topics in Figure 3. It

[4] https://pypi.python.org/pypi/lda
[5] http://bigml.cs.tsinghua.edu.cn/ jun/stc.shtml/
[6] https://github.com/elbamos/NeuralTopicModels

Table 2: The word codes of representative words for different categories discovered by NVSTM and LDA.

Category	NVSTM	LDA	Topic
Business			T6: marketing parascope development business sustainable partnerships movieactors developing partnership T63: finance loans equity loan mortgage financing banking investment mortgages T67: investing ratneshwar investments investment investors invest equity niddk income T133: products source product quality premium csail content manufacture socialsciences T144: development sciserv ecommerce developing innovation developers business marketing projects T176: trade trading markets commodities commodity stocks market parascope currencies
Computers			T38: processor microprocessor processors llnl signonsandiego cpu microprocessors intel cores T108: memory laptop computer computers processor nutritionsource laptops intel disk T112: firefox mozilla netscape macintosh linux windows adobe verizon zdnet T118: systems system control security controls remote automatic monitoring automation T121: msn yahoo firefox aol gmail java algorithm algorithms signonsandiego T159: quantum computing space nasa cpu computational computers astrophysics physics
culture -arts- entertainment			T3: ocos parascope space socialsciences living world academyawards planet intradoc T5: film films indie filmmaker filmmakers movie comedy screening filmmaking T10: sound audio voice acoustic recordings recording listening bass song T16: photography poetry poems prose poet writing getthejob poem photographer T58: sculptor painter artist sculpture sculptures paintings artists artwork surrealist T177: art sculpture socialsciences sculptures painter paintings sculptor painting pcguide
education - science			T41: mathematics physics maths professors students undergraduate science teachers ncidod T59: undergraduate degree student undergraduates faculty students acts particles mathematics T82: teaching school mathforum english teacher mathematics education college schools T102: lecture book lectures papers essay journal seminar conference books T109: topics mathforum essays lectures articles journals emedicinehealth literature syllabus T147: science scientific research journals published theories sciences publications articles
Engineering			T7: machine machines software printer llnl pcguide converter freeware kurose T69: engine engines fuel diesel cylinder gearbox piston motor petrol T90: factory inc steel searchsmb chrome ford socialsciences wieeless ltd T100: device cancertopics devices modem cable died wireless semiconductor connection T114: gasoline diesel fuel petrol engines engine emissions gas combustion T141: salary cancertopics engineer nutritionsource engineering engineers jobs job talent
Health			T55: therapists medical physicians therapy nurses pediatric therapist clinics doctors T56: calories nutritional vitamins calorie diet carbohydrates fats nutrition foods T86: hospital webobjects home nutritionsource clinic homes nursing emedicinehealth center T124: disease cancer lung flu cancers infection influenza arthritis infections T142: vitamins foods herbal diet alcohol supplements vitamin oils nutritional T156: drugs drug medications medication tobacco smoking alcohol prescription cancer
Politics-society			T27: culture democracy capitalism politics socialism society socialist ideology democratic T103: secretary party fhlb highfat parliamentary managementhelp parties elections election T128: candidates candidate election elections presidential ballot electoral vote voting T148: scandal election president senator minister senate presidency elections chairman T155: participatory debates debate democracy activism democratic pluralism grassroots encourage T168: party parties election elections electoral parliamentary parliament presidential political
Sports			T72: player socialsciences players essortment game games straight button bottom T79: hockey basketball football soccer volleyball sports baseball tennis sport T95: resort village resorts town mountain peaceful hotel valley coastal T119: tennis volleyball emnlp swimming tournaments cricket hockey softball skiing T125: poker tournaments games tournament game betting gaming competitions football T127: football league soccer rugby championship basketball championships playoffs leagues

is obvious to see that all documents are clustered into 8 distinct categories, which is equal to the ground truth number of categories in the web snippet. It proves the semantic effectiveness of the documents codes learned by our model. Noticed that a few of the document codes in different categories are mixed together, which can be caused by the dimensionality reduction.

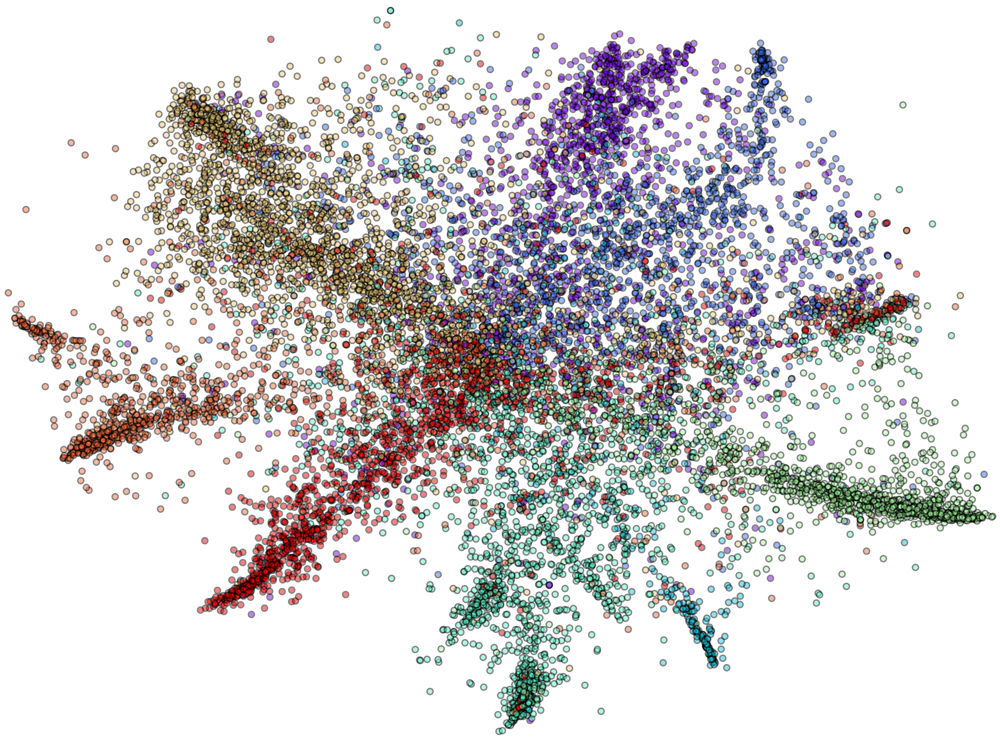


Figure 3: t-SNE projection of the estimated document codes from Web snippet. The vectors are learned by the NVSTM model with 200 topics, each color represents one category from the 8 different categories of the dataset.

4.3 Classification Accuracy

To further evaluate the effectiveness of the representation of documents learned by NVSTM, we perform text classification tasks on web snippet and 20NG, using the document codes of documents

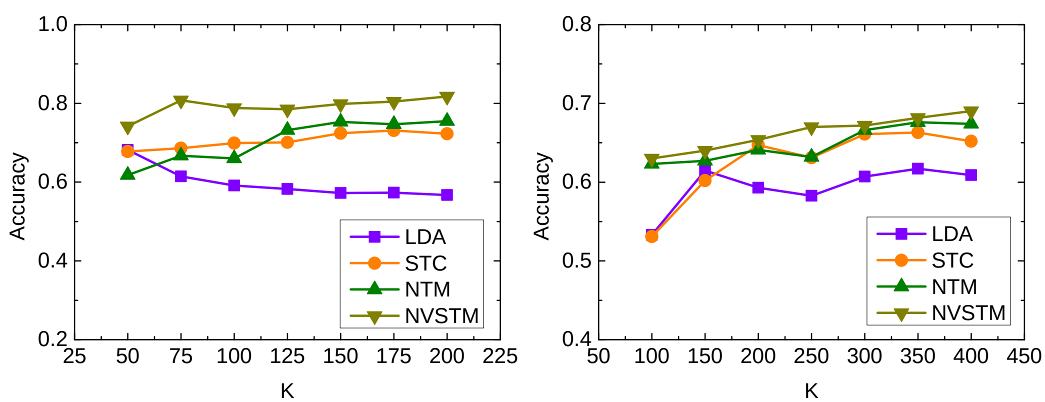


Figure 4: Classification accuracy of different models on Web snippet and 20NG, with different number of topic K settings.

as the feature representation in a multi-class SVM. On web snippet, we utilize 80% documents for training and 20% for testing. On the 20NG data set, we keep 60% documents for training and 40% for testing, which is the same configuration as in (Lin et al., 2014). Figure 4 report the classification accuracy under different methods with different settings on the number of topics among web snippet and 20newsgroup. It clearly denotes that 1) The NVSTM yields the highest accuracy, followed by NTM and STC which all outperform the LDA. 2) The neural network based NVSTM and NTM generate better document representations than STC and LDA, demonstrating the representative advantage of neural networks in distributed representations. 3) Sparse models (NVSTM and STC) are superior to non-sparse models NTM and LDA separately. It indicates that sparse topic models are more capable to extract topics from short documents. 4) The NTM, STC without word embeddings generate similar performance, which affirms both word embeddings and sparse penalty are contributing to learning the document representations with clear semantic explanations.

5 Conclusion

We propose a neural sparsity-enhanced topic model NVSTM, which is the first effort in introducing effective VAE inference algorithm to STC as far as we know. We take advantage of VAE to simplify the inference process, which require no model-specific algorithm derivations. With the employing of word embeddings and neural network framework, NVSTM is able to generate clearer and semantic-enriched representations for short texts. The evaluation results demonstrate the effectiveness and efficiency of our model. Future work can include extending our model with other deep generative models, such as generative adversarial network (GAN).

References

Cedric Archambeau, Balaji Lakshminarayanan, and Guillaume Bouchard. Latent ibp compound dirichlet allocation. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):321–333, 2015.

Lu Bai, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. Group sparse topical coding: from code to topic. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 315–324. ACM, 2013.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2003a.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003b.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. In *AAAI*, pp. 2210–2216, 2015.

Dallas Card, Chenhao Tan, and Noah A Smith. A neural framework for generalized topic models. *arXiv preprint arXiv:1705.09296*, 2017.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41 (6):391, 1990.

Matthias Heiler and Christoph Schnörr. Learning sparse representations by non-negative matrix factorization and sequential cone programming. *Journal of Machine Learning Research*, 7(Jul): 1385–1407, 2006.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong 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*, pp. 539–550. ACM, 2014.

Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pp. 121–128, 2008.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. *arXiv preprint arXiv:1706.00359*, 2017.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

Min Peng, Binlong Gao, Jiahui Zhu, Jiajia Huang, Mengting Yuan, and Fei Li. High quality information extraction and query-oriented summarization for automatic query-reply in social network. *Expert Systems with Applications*, 44:92–101, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM, 2010.

Kentaro Sasaki, Tomohiro Yoshikawa, and Takeshi Furuhashi. Online topic model for twitter considering dynamics of user interests and topic trends. In *EMNLP*, pp. 1977–1985, 2014.

Akash Srivastava and Charles Sutton. Neural variational inference for topic models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Sinead Williamson, Chong Wang, Katherine A Heller, and David 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)*, pp. 1151–1158, 2010.

Jun Zhu and Eric P. Xing. Sparse topical coding. *CoRR*, abs/1202.3778, 2011.