




Extracting nonlinear neural topics with neural variational bayes

Yiming Wang^{1,2} · Ximing Li^{1,2}  · Jihong Ouyang^{1,2} · Zeqi Guo^{1,2} · Yimeng Wang^{1,2}

Received: 21 July 2020 / Revised: 14 August 2021 / Accepted: 1 October 2021 /
Published online: 20 November 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Recently, topic modeling has been upgraded by neural variational inference, which simultaneously allows the model structures deeper and proposes efficient update rules with the reparameterization trick. We formally call this recent new art as *neural topic model*. In this paper, we investigate a problem of neural topic models, where they formulate topic embeddings and measure the word weights within topics by linear transformation between topic and word embeddings, resulting in redundant and inaccurate topic representations. To solve this problem, we propose a novel neural topic model, namely *Generative Model with Nonlinear Neural Topics* (GMNNT). The insight of GMNNT is to replace the topic embeddings with neural networks of topics, named *neural topic*, so as to capture nonlinear relationships between words in the embedding space, enabling to induce more accurate topic representations. We derive the inference process of GMNNT under the framework of neural variational inference. Extensive empirical studies have been conducted on several widely used collections of documents, including datasets of both short texts and normal long texts. The experimental results validate that GMNNT can output more semantically coherent topics compared with traditional topic models and neural topic models.

Keywords Variational auto-encoders · Topic modeling · Neural topic · Reparameterization

✉ Ximing Li
liximing86@gmail.com

Yiming Wang
yimingw17@gmail.com

Jihong Ouyang
ouyj@jlu.edu.cn

Zeqi Guo
guozeqi95@gmail.com

Yimeng Wang
wangyimeng116@gmail.com

¹ College of Computer Science and Technology, Jilin University, Jilin, China

² Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Jilin, China

1 Introduction

Nowadays, the volume of text data becomes increasingly large everyday, *e.g.*, online news and reports generated by a variety of daily web services [33]. Automatically mining the latent theme information from them with unsupervised learning is a significant and challenging research subject. Topic models [2], *e.g.*, Latent Dirichlet Allocation (LDA) [6] and Hierarchical Dirichlet Processes (HDP) [48], have become one of the great successful unsupervised techniques for inducing latent topics from text documents. In the past decades, topic models surveyed by [7] have been applied to numbers of fields, *e.g.*, sociology, marketing, and political science, to name just a few.

Most traditional topic models, built on the spirit of LDA [6], are probabilistic generative models of documents, basically supposing that each document is represented by a topic proportion and each topic is a multinomial distribution over words. With conjugacy designs, *e.g.*, Dirichlet-Multinomial distributions, the posterior distributions associated with the topics can be efficiently inferred by either variational inference [3, 6, 24] or sampling methods, *e.g.*, collapsed Gibbs sampling [18]. Generally, the expressiveness of topic models grows with more complicated model structures, however, this also results in intractable inference problems. Recently, topic modeling has been upgraded by neural variational inference [26, 35, 39, 45], which approximates the posterior distribution of a generative model with a variational distribution parameterized by a neural network [34]. This simultaneously allows the model structures deeper and proposes efficient update rules with the reparameterization trick [26, 49], developing a new trend of topic modeling, formally referred to as **neural topic model**, *i.e.*, the art that marries topic modeling with deep neural networks.

The previous research literatures have introduced dozens of neural topic models [8, 10, 12, 13, 19, 22, 33–35, 41, 43, 47, 52–54]. From the perspective of model inference by variational Bayes, in neural topic models the neural network serves as a variational distribution to the target distribution, *i.e.*, often the posterior of latent variables. Or they can be read as Variational Auto-Encoders (VAE) kind of models [26]. As an example, the Neural Variational Document Model (NVDM) [35] involves two halves, *i.e.*, an encoding network for latent topics and a generative decoding model for document reconstruction from topics. The subsequent study [34] normalizes the latent topics of NVDM for achieving distribution expressions of topics, and proposes three versions of neural topic models with different neural structures. However, the aforementioned models suffer from a shared weakness: they formulate topics as embeddings, *i.e.*, distributed representations in the word embedding space, and measure the word weights within topics by linear transformation between topic and word embeddings. This results in the problem of redundant and inaccurate topic representations, which is going to be discussed in the following part.

1.1 Problem, motivation and contribution

To deeply discuss the prior neural topic models, we briefly introduce a standard generative formulation of documents [13, 34]. Specifically, the generative process of a document $\{w_{dn}\}_{n=1}^{N_d}$ can be described as follows:

$$\begin{aligned}\theta_d &\sim \mathbf{G}(\mu_0, \sigma_0), \\ z_{dn} &\sim \mathbf{Multinomial}(\theta_d), \quad n \in [N_d], \\ w_{dn} &\sim \mathbf{Multinomial}(\phi_{z_{dn}}), \quad n \in [N_d],\end{aligned}\tag{1}$$

where θ_d denotes the topic proportion drawn from $\mathbf{G}(\mu_0, \sigma_0)$, *i.e.*, a neural network conditioned on an isotropic Gaussian $\mathcal{N}(\mu_0, \sigma_0)$, *e.g.*, logistic-normal distribution [13], Gaussian softmax distribution, and Gaussian stick breaking distribution [34]; z_{dn} the topic assignment drawn from θ_d ; and ϕ_t the topic distribution over words, constructed by the softmax function of the product of word embeddings ρ and topic embedding β_t :

$$\phi_t = \mathbf{softmax}(\rho^\top \beta_t) \tag{2}$$

For simplicity, we now by no means introduce the notations too much, which will be detailedly described in the latter section (also see Table 1).

Referring to (2), we notice that the word weights within topics are actually computed by the inner product distance between topic and word embeddings. In this situation, neighboring words tend to share similar weights in the same topic, resulting in potentially redundant top topical words, also observed in the early models [1, 11, 28]. To visualize this problem, Figure 1 shows two examples (*i.e.*, topic embeddings and embeddings of top words) learnt by the prior model [13] across *NewYorkTimes*. We can observe that many of the top word lists contain many similar words, resulting in redundancy.

In this paper, we introduce the proposed generative model that extracts nonlinear neural topics in embedding spaces, namely **Generative Model with Nonlinear Neural Topics (GMNNT)**. In GMNNT we replace the topic embeddings with neural networks of topics, formally referred to as **neural topic**, which can capture nonlinear relationships between words in the embedding space. Therefore, even similar words, *i.e.*, neighbors measured by word embeddings, are allowed with totally different probabilities in the same topic distribution, leading to more accurate topic representations.

The main contributions of this paper are described below:

- We investigate the problem of prior neural topic models, where they may output inaccurate topics with redundant top topical words.
- We develop a new GMNNT model that uses neural topics, describing nonlinear relationships between word embeddings.
- Empirical studies show that GMNNT can generate semantically coherent topics in contrast to traditional neural topic models.

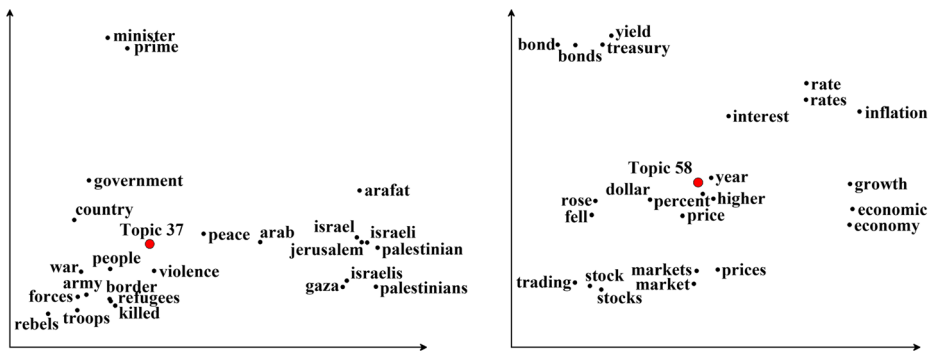


Figure 1 Top topical word redundancy in ETM [13]. Each topic is visualized by its 24 top words on a 2-dimensional space compacted by principle component analysis. The topic embeddings and word embeddings are represented by red and black points, respectively

The rest of this paper is organized as follows: In Section 2, we introduce the most related works. We present the proposed GMNNT and its inference process in Section 3. The empirical studies are shown in Section 4. Finally, we conclude this work in Section 5.

2 Related work

In this section, we briefly introduce the previous studies on traditional topic models and neural topic models.

2.1 Traditional topic model

On referring to reviews [2, 7], traditional topic modeling, such as LDA [6] and its nonparametric version (*i.e.*, HDP) [48], has been well studied as probabilistic generative models of documents. Generally speaking, they are probabilistic generative models of documents. For example, in the context of LDA, it considers the corpus as a mixture of K topics and each document corresponds with a topic proportion θ_d , drawn from the Dirichlet prior α . Each topic presents a multinomial distribution over the vocabulary ϕ , drawn from the Dirichlet prior β . Specifically, the generative process of a document collection can be described as follows:

- For each topic $k \in [K]$
 - Sample a topic $\phi_k \sim \mathbf{Dirichlet}(\beta)$
- For each document $w_d, d \in [D]$
 - Sample a topic proportion $\theta_d \sim \mathbf{Dirichlet}(\alpha)$
 - For each word token $w_{dn}, n \in [N_d]$
 - Sample a topic assignment $z_{dn} \sim \mathbf{Multinomial}(\theta_d)$
 - Sample a word $w_{dn} \sim \mathbf{Multinomial}(\phi_{z_{dn}})$

where z_{dn} represents the topic assignment for each word token.

In the past decades, researchers have developed many extensions built on the above formulation, and they have been successfully applied to deal with various problems as well as various kinds of text data, *e.g.*, topic correlations [5, 23, 27, 30], dynamic topics varying over time [4, 51], and sparse topics within short texts [9, 16, 31, 44], *etc.* Commonly, the popular model inference methods include variational inference often with mean-field approximations [6, 24], Gibbs sampling [18], and hybrid methods [29, 38]. However, to maintain model inference efficient, traditional topic models are more willing to be designed as shallow structures with conjugate priors, which somehow limits expressiveness.

2.2 Neural topic model

Recently, a new trend of topic modeling, *i.e.*, neural topic model, has raised lots of concerns [8, 10, 12, 13, 19, 22, 33–35, 41, 47, 53, 54]. For ease of understanding, we briefly introduce these models from the eye of VAE, *i.e.*, the encoder-decoder perspective of documents. *Encoding*: In this situation, the original document representations, *e.g.*, bag-of-words, are

encoded as (unnormalized) topic proportions by variational neural networks¹. *Decoding*: The reconstructed document representations, *i.e.*, probabilities of the vocabulary, are computed by a LDA-like generative process [13, 34], which starts from topic proportions with topic-word distributions specified by topic embeddings. Specially, the variational distributions are often fixed as Gaussians for efficiently applying the reparameterization trick [26, 39, 45]. Generally speaking, several typical neural topic models include NVDM [35], Gaussian Softmax Model (GSM) [34], LDA with Products of Experts (ProdLDA) [47], Neural Variational LDA (NVLDA) [47], and Embedded Topic Model (ETM) [13], *etc.*

Due to the reparameterization property of the Gaussian distribution, many existing neural topic models [13, 34, 35] employ it as the prior distribution for efficient inference. NVDM [35] is an early attempt modeling topics with Gaussian prior, in which a two-layer multi-layer perception is applied for encoding. NVDM can achieve rather superior perplexity scores and meanwhile generates incoherent topics in many cases, as reported in [13]. Inspired by the context embedding [37], ETM [13] incorporates word and topic embeddings into neural topic models, assuming that each topic is presented in a word embedding space. ETM generates topic and word embeddings simultaneously during training and has a variant with pre-trained word embeddings. Due to the inner product of the topic and word embeddings in the decoder of ETM, words with high semantic correlations are more likely to gather in the same topic embedding space, leading to topical words redundancy. Besides, to capture topic correlations, [33] proposed Neural Variational Correlated Topic Model (NVCTM), which incorporates Centralized Transformation Flow (CTF), enabling to model Gaussian distributions with covariance matrix.

Another group of attempts approximate the Dirichlet distribution as prior for neural topic models as which in LDA, for lacking the intuitional non-central differentiable reparameterizations for Dirichlet distribution under neural variational inference [8]. ProdLDA [47] explores the Laplace approximation for the Dirichlet prior. A relatively high learning rate and batch normalization prevent ProdLDA from component collapsing. [40] approaches the Dirichlet prior by a rejection sampler on Gamma distribution. The proposed Rejection Sampling Variational Inference (RSVI) creates an elegant and extensible way for solving the reparameterization challenge and studies the approximations of Gamma distribution and Dirichlet distribution. Based on RSVI, [54] generates an approximation of Gamma distribution utilizing Weibull distribution since reparameterization trick is available on Weibull distribution. [8] proposes Dirichlet Variational Autoencoder (DVAE) and decouples sparsity and smoothness in the Dirichlet distribution.

There also exist studies about variants of VAE-based topic models. Adversarial-neural Topic Model (ATM) [52] generates an approach adapting Generative Adversarial Nets (GANs) with the Dirichlet prior to topic modeling. [41] broadens the Wasserstein Autoencoder [50] and proposes W-LDA which is capable of matching aggregated posteriors to priors utilizing the Maximum Mean Discrepancy (MMD).

3 Model

In this section, we introduce the proposed generative model that extracts nonlinear neural topics in embedding spaces, namely **Generative Model with Nonlinear Neural Topics (GMNNT)**.

¹We will clarify the definition of variational neural network in Section 3.2.

3.1 Model description

Consider a corpus of D documents, *i.e.*, denoted by $\{w_d\}_{d=1}^D$, with a fixed vocabulary of V words. Each document of N_d word tokens is represented by $\{w_{dn}\}_{n=1}^{N_d}$. With any existing embedding technique, *e.g.*, *Word2Vec* [36] and *GloVe* [42], the pre-trained L -dimensional word embeddings almost covering the current vocabulary are available, *i.e.*, denoted by $\rho \in \mathbb{R}^{L \times V}$, where each column ρ_v is the corresponding embedding of word v .

In the context of traditional neural topic models [13, 34], the documents are generated from distributions associated with topics. The topics are represented by L -dimensional topic embeddings, and referring to (2), the word weights within topics are measured by linear transformation between topic and word embeddings. However, such resulting topic-word distributions may contain many redundant top topical words, *i.e.*, outputting inaccurate topic representations, which has been explained before (see Figure 1). To solve this problem, in GMNNT we replace the topic embeddings with neural networks of topics (*i.e.*, neural topics), which can capture nonlinear relationships between words in the embedding space [32]. Therefore, even similar words, *i.e.*, neighbors measured by word embeddings, are allowed with totally different probabilities in the same topic distribution, leading to more accurate topic representations. Specifically, we define that $\mathbf{NT}(\rho|\varphi_t)$ denotes the neural topic t with word embeddings ρ , formulated as follows:

$$\phi_{tv} \triangleq \mathbf{NT}(\rho|\varphi_t) \Big|_v = \mathbf{softmax} \left([f(\rho_1|\varphi_t), \dots, f(\rho_V|\varphi_t)]^\top \right) \Big|_v \quad (3)$$

where $f(\rho_v|\varphi_t)$ is a neural network parameterized by φ_t with the input of one word embedding ρ_v and the output of the corresponding untransformed word weight.

Overall speaking, the model structure of GMNNT is under the framework of traditional neural topic models, described in (1). For clarity, we now formally introduce the generative process of GMNNT as follows: Suppose that there are totally T neural topics ϕ (*i.e.*, (3)), representing multinomial distributions over words. For each document w_d , GMNNT first draws a topic proportion θ_d from a neural network kind of prior $\mathbf{G}(\mu_0, \sigma_0)$, named *topic proportion generator*. Then, it draws a topic assignment z_{dn} from θ_d , and then draws a word w_{dn} from $\phi_{z_{dn}}$. Repeat this process N_d times for N_d word tokens. In summary, the generative process of GMNNT is described below:

- For each document w_d , $d \in [D]$
 - Sample a topic proportion $\theta_d \sim \mathbf{G}(\mu_0, \sigma_0)$
 - For each word token w_{dn} , $n \in [N_d]$
 - Sample a topic assignment $z_{dn} \sim \mathbf{Multinomial}(\theta_d)$
 - Sample a word $w_{dn} \sim \mathbf{Multinomial}(\phi_{z_{dn}}) \triangleq \mathbf{NT}(\rho|\varphi_{z_{dn}})$

In this work, we specify the topic proportion generator $\mathbf{G}(\mu_0, \sigma_0)$ as the Gaussian softmax distribution [34]. For each document w_d it first generates an untransformed topic proportion δ_d from a isotropic Gaussian $\mathcal{N}(\mu_0, \sigma_0)$ and then applies the softmax function to compute the final θ_d :

$$\delta_d \sim \mathcal{N}(\mu_0, \sigma_0), \quad \theta_d = \mathbf{softmax}(W^\top \delta_d), \quad (4)$$

where $W \in \mathbb{R}^{T \times T}$ is the linear transformation matrix. We would like to note that our GMNNT is feasible to apply more complex topic proportion generator, leading to more practical variants of GMNNT. The important notations of this paper are shown in Table 1 for convenience.

Table 1 Descriptions of important notations

| Notation | Description |
|------------------------------------|--|
| D | number of documents |
| N_d | number of word tokens in document d |
| V | number of words |
| T | number of topics |
| L | dimension of word embeddings |
| $\rho \in \mathbb{R}^{L \times V}$ | word embeddings |
| θ_d | topic proportion of document d |
| z_{dn} | topic assignment for word token w_{dn} |
| ϕ_t | topic distribution over words of topic t |
| $\beta_t \in \mathbb{R}^L$ | topic embedding of topic t defined in [13, 34] |
| φ_t | parameter of neural topic t defined by our GMNNT |

3.2 Inference

From the perspective of topic modeling, the model parameters of GMNNT include the neural topic parameter φ and the Gaussian hyper-parameters $\{\mu_0, \sigma_0\}$, while the latent variables include the untransformed topic proportion δ and topic assignment z . In our situation, $\{\mu_0, \sigma_0\}$ are fixed as known priors and z can be analytically integrated out. Therefore, the inference problem refers to finding the optimum of $\{\varphi, \delta\}$ by fitting GMNNT given a collection of documents \mathcal{D} and word embeddings ρ .

Commonly, the inference problem of GMNNT is intractable to compute, therefore we resort to approximating inference by neural variational inference [34, 35] with the reparameterization trick [26]. With the spirit of amortized inference [17], we posit the following variational distributions over the untransformed topic proportion δ :

$$q(\delta_d | w_d, \lambda) = \mathcal{N}(\delta_d | \mu_d, \sigma_d), \quad \{\mu_d, \sigma_d\} = g(w_d | \lambda), \quad d \in [D], \quad (5)$$

where $g(w_d | \lambda)$ is the *variational neural network* parameterized by λ (*i.e.*, considered as the variational parameter). That is, the network ingests w_d and outputs $\{\mu_d, \sigma_d\}$. Following [13], we form the input w_d by normalizing its bag-of-word representation by the number of word tokens N_d . Applying these variational distributions, we can formulate the following variational objective, *i.e.*, Evidence Lower Bound (ELBO), with respect to $\{\varphi, \lambda\}$:

$$\mathcal{L}(\varphi, \lambda) = \sum_{d=1}^D \mathbb{E}_q [\log p(w_d | \delta_d, \rho, \varphi)] - \sum_{d=1}^D \text{KL}(q(\delta_d | w_d, \lambda) || p(\delta_d | \mu_0, \sigma_0)), \quad (6)$$

The likelihood of each document in (6) is given by:

$$p(w_d | \delta_d, \rho, \varphi) = \prod_{n=1}^{N_d} \sum_{t=1}^T \theta_{dt} \phi_{t w_{dn}}, \quad (7)$$

where ϕ and θ are obtained by (3) and (4), respectively.

Since variational distributions are Gaussians, we can replace the variational objective of (6) with its Monte Carlo approximation by using the reparameterization trick [26]:

$$\begin{aligned}
 \mathcal{L}(\varphi, \lambda) &= \sum_{d=1}^D \mathbb{E}_q [\log p(w_d | \delta_d, \rho, \varphi)] - \sum_{d=1}^D \text{KL}(q(\delta_d | w_d, \lambda) \parallel p(\delta_d | \mu_0, \sigma_0)) \\
 &= \sum_{d=1}^D \mathbb{E}_q [\log p(w_d | \delta_d, \rho, \varphi) + \log p(\delta_d | \mu_0, \sigma_0) - \log q(\delta_d | w_d, \lambda)] \\
 &\approx \frac{1}{S} \sum_{d=1}^D \sum_{s=1}^S \log p(w_d | h(\delta_d^{(s)}), \rho, \varphi) + \log p(h(\delta_d^{(s)}) | \mu_0, \sigma_0) - \log q(h(\delta_d^{(s)}) | w_d, \lambda) \\
 &\qquad\qquad\qquad \delta_d^{(s)} \sim \mathcal{N}(0, \mathbf{I}), \quad d \in [D], s \in [S],
 \end{aligned} \tag{8}$$

where S is the number of Monte Carlo samples² and $h(\delta_d^{(s)}) = \delta_d^{(s)} \sigma_d + \mu_d$ the mapping function.

Algorithm 1 Approximating inference for GMNNT.

- 1: *Initialize* neural topic parameter φ and variational neural network parameter λ
 - 2: *Compute* the normalized bag-of-word representations of documents
 - 3: *Draw* S Monte Carlo samples from the standard Gaussian to form the approximating variational objective, referring to (8)
 - 4: **While** $\{\varphi, \lambda\}$ almost unchanged **Do**
 - 5: *Compute* the topic distributions ϕ using (3)
 - 6: *Draw* a small subset of documents
 - 7: **For** each document w_d in the subset draw **Do**
 - 8: *Compute* $\{\mu_d, \sigma_d\}$ with the variational neural network, referring to (5)
 - 9: *Draw* a sample δ_d from the variational distribution, referring to (5)
 - 10: *Compute* the topic proportion θ_d by (4)
 - 11: **End For**
 - 12: *Update* $\{\varphi, \lambda\}$ with their gradients, and the adaptive learning rate method can be used
 - 13: **End While**
-

Given this approximating variational objective, we can form its gradients with respect to $\{\varphi, \lambda\}$, where the subgradients of neural networks (*i.e.*, variational neural networks and neural topics) can be computed by backpropagation. We then update $\{\varphi, \lambda\}$ with their gradients under numbers of updating cycles until $\{\varphi, \lambda\}$ are almost unchanged. To efficiently deal with corpora of massive documents, the data subsampling methodology from [20, 21] can be also applied. Finally, for fast and safe updating processes, we can adopt any adaptive learning rate method, *e.g.*, Adagrad [15], Adam [25], and Nadam [14], *etc.* The full inference procedure of GMNNT is briefly shown in *Algorithm 1*. Specially, we would like to note that we can reform the approximating variational objective of (8) by drawing new Monte Carlo samples during the updating cycles. We omit this detail in *Algorithm 1* for concise expression.

²In this work, we fix S to 1 as suggested in [26].

4 Experiment

Corpora The experiments have been conducted across 5 publicly available datasets, whose statistics are summarized in Table 2. Specifically, they include *Trec*³, *StackOverflow*⁴, *Abstract*⁵, *Tmc2007*⁶, and *NewYorkTimes*⁷. For each dataset, we removed the standard stop words and infrequent words occurred in less than 5 documents. We randomly selected 85% instances as the training dataset, 10% as test dataset and 5% as validation dataset.

Besides, we employed the pre-trained *GloVe*⁸ word embeddings [42], *i.e.*, the 300-dimensional version trained on Wikipedia2014 and Gigaword5. We randomly generated the embeddings of words that have not been covered by GloVe embeddings.

Comparing Topic Models We compare the performance of GMNNT with four existing topic models, including three neural topic models, *i.e.*, ETM, NVDM and ProdLDA, and also the standard Online LDA (OLDA) model. Details of all comparing models are described below:

- **LDA** [6, 20, 21] is the standard LDA model trained by stochastic variational inference. Here, the Dirichlet priors of document-topic proportions and topics are set to 0.1 and 0.01, respectively. The code is available on the net⁹.
- **NVDM** [35] is an unnormalized neural topic model with Gaussian prior. Following [14], the encoder used in NVDM defines a fully connected network with 2 layers and 500 hidden neurons. It involves an inner iteration for optimizing the encoder. The code is provided by its authors.¹⁰
- **ProdLDA** [47] is a neural topic model with Dirichlet prior, solved by Laplace approximation. Following [47], we define its encoder as a fully connected network with 3 layers and 100 hidden neurons, where the tricks of batch normalization and 0.2 dropout are also used. The code is provided by its authors¹¹.
- **GSM** [34] is a neural topic model using Gaussian Softmax which constructs a finite topic distribution. We inherit the same encoder from NVDM and the number of hidden neurons and word vectors dimension are both set as 500. We adopt the document model version. The code is available on the net.¹²
- **ETM** [13] is a neural topic model with Gaussian prior. Following [13], the encoder used in ETM defines a fully connected network with 3 layers and 800 hidden neurons. The code is provided by its authors.¹³
- **GMNNT** is our proposed neural topic model with nonlinear neural topics. We use the same encoder as ETM. The neural topics are designed as fully-connected networks with λ_d layers and λ_w hidden neurons, and the batch normalization is applied. The

³<http://cogcomp.cs.illinois.edu/Data/QA/QC/>

⁴<https://github.com/jacoxu/STC2>

⁵dataset of paper abstracts

⁶<http://mulan.sourceforge.net/datasets-mlc.html>

⁷<https://bitbucket.org/franrruiz/data-nyt-largev-6/src/master/>

⁸<https://nlp.stanford.edu/projects/glove/>

⁹<https://github.com/blei-lab/lda-c>

¹⁰<https://github.com/ysmiao/nvdm>

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

¹²<https://github.com/linkstrife/NVDM-GSM>

¹³<https://github.com/adjidieng/ETM>

Table 2 Summary of dataset statistics. “#Validation” and “AvgL” denote the validation set size and the average document length, respectively

| Dataset | #Train | #Test | #Validation | #Word | AvgL |
|---------------|--------|-------|-------------|-------|-------|
| Trec | 4972 | 492 | 291 | 1036 | 3.0 |
| StackOverflow | 16662 | 1798 | 980 | 2195 | 3.7 |
| Abstract | 21481 | 2526 | 1266 | 25955 | 74.5 |
| Tmc2007 | 24305 | 2859 | 1431 | 24593 | 59.4 |
| NewYorkTimes | 85000 | 10000 | 5000 | 32746 | 290.9 |

parameters of neural topics are tuned over the following ranges: $\lambda_d \in \{2, 3, 4, 5, 6, 7, 8\}$ and $\lambda_w \in \{60, 80, 100, 120\}$. We will analyze these parameters later.

For all comparing models, the mini-batch size is set to 1000. The learning rate and epoch for GMNNT are set as 0.0001 and 2000, respectively. Besides, the Adam method is used for adaptively tuning the learning rate under the following settings: $\beta_1 = 0.9$, $\beta_2 = 0.999$. Specifically, we study two variants of ETM and GMNNT following [13] where the one applies the pre-trained word embeddings, and the other leaves the word embeddings as trainable parameters. The versions with pre-trained word embeddings are called p -ETM and p -GMNNT, respectively.

4.1 Qualitative study

The first concern is whether GMNNT enables to alleviate the problem of redundant top topical words that have been observed in previous neural topic models. To answer it, as shown in Table 3 we list 10 same topics learned by p -ETM and p -GMNNT across the NewYorkTimes dataset. We may find that p -ETM severely suffers from the topical words redundancy, where 17 pairs of words with same etyma exist in 10 topics, *i.e.*, “restaurant” and “restaurants”. As for p -GMNNT, there only exist 4 pairs in 4 topics which indicates that our method significantly relieves the topical words redundancy.

More specifically, ETM and p -ETM apply an inner product decoder, which gathers closer words in the word embedding space and generate topics closest to these gathered words. These close words are mostly semantically related and result in topical words redundancy. Our GMNNT and p -GMNNT learn word weights for each topics independently in a non-linear manner, which weakens the words gathering in the embedding space. Furthermore, different topics in ETM and p -ETM exist in the same embedding space and therefore words under different topics may be duplicated when the word embeddings are not discriminative enough, especially for the version without pre-trained embeddings. The topic uniqueness results in Section 4.2.3 coincide with this analysis.

We further compare topic quality over all methods and present topics about “aircraft fire” from Tmc2007 in Table 4. We find that NVDM and ProdLDA generate many identical topics, which severely suffer from topic redundancy. Comparing with baselines, GMNNT and p -GMNNT only generate one topic with more meaningful words, corresponding with the high topic uniqueness in Section 4.2.3. There seems to exist a trade-off between the number of topics and topic uniqueness scores on NVDM and ETM. Actually, topical words

Table 3 Top topical words of 10 same NewYorkTimes topics generated by ETM and GMNNT, respectively. Redundant words are given in boldface

| Method | Topical words |
|-----------------|--|
| <i>p</i> -GMNNT | street city people york town manhattan east place restaurant local neighborhood bar dr health drug medical disease patients doctors aids cancer study research drugs care health hospital york services program people city state center hospitals energy space power nuclear dr environmental scientists air gas plant research earth fashion ms clothes black designer dress white women wear design collection style bush republican campaign president clinton party senator democratic democrats computer internet technology company software web system companies microsoft editor article news page times newspaper york paper writer magazine daily press court case judge justice federal trial lawyers law lawyer charges jury state attorney president white house clinton office committee investigation told asked report |
| <i>p</i> -ETM | restaurant wine food restaurants sauce menu dinner dishes cheese wines dining health medical drug patients care hospital doctors cancer disease treatment drugs people home middle years high low families class year homes community live oil energy power car cars gas production fuel ford plant auto vehicles vehicle fashion clothes hair dress wear designer dresses wearing black leather pants republican campaign democratic party senate election senator democrats vote computer technology internet information software computers web system editor news times magazine newspaper press paper editorial writer newspapers court law judge case federal justice rights legal supreme state decision states investigation report information documents officials department records official |

generated by NVDM on Tmc2007 are quite infrequent and even less repeated, leading to higher topic uniqueness scores, which will be analyzed below. As for ETM, it generates plenty of identical words among different topics, e.g., “aircrafts”, which severely harm the topic uniqueness.

We also notice that NVDM and ProdLDA generate “low-quality” topics on larger datasets. For smaller datasets, e.g., Trec and StackOverflow, the two methods generate “high-quality” topics with meaningful words. However when facing the larger Abstract and Tmc2007, they begin to generate some infrequent words which harm the topic quality. As for the large NewYorkTimes, they generate rather poor topics with numerous infrequent words, e.g., names of persons and places. For validating the guess that NVDM and ProdLDA suffer from topic quality descending with the increment of dataset scale, we examine NVDM and ProdLDA on different truncation sizes of NewYorkTimes and the results show the same case.

4.2 Quantitative study

We quantitatively evaluate the proposed GMNNT model by three tasks of held-out likelihood, topic coherence and topic uniqueness.

Table 4 Topical words on topics about aircraft fire generated by GMNNT and other baselines on Tmc2007 dataset

| Method | Topical words |
|-----------------|---|
| LDA | smoke engine fire light auxiliarypowerunit start aircraft maintain electric report |
| NVDM | enginefire firebottle firewarning flame firebell firelight enginefirewarning firehandle fanblade enginefire firewarning firewarninglight aircraftrescuefirefightingequipment odor smell fume acrid smoke dissipate goggle electric electricalfire tennis |
| ProdLDA | gcy fire evacuate smoke engineoilpressure oilpressure smell oilquantity declare fire overweight extinguish evacuate oilpressure qrh declare smoke cabinaltitude gcy fire oilquantity oilpressure engineoilpressure evacuate smoke overweight declare fire overheat aircondition recirculation cardiac acrid overwingexit cabin electric pack fire enginefire declare overweight extinguish smoke oilquantity engineindication |
| GSM | smell smoke evacuate paramedic odor fume deplane doctor maincabin fire rescue medicalpersonnel paramedic defibrillator fume dissipate minorinjury smell evacuate |
| ETM | emergency declare engine land cabin flightattendant checklist passenger smoke fire |
| <i>p</i> -ETM | door smoke sit seat back fire window smell touch inside burn floor front button nose |
| GMNNT | oilpressure enginefirechecklist oilquantity fireindication lowoilpressure firelight |
| <i>p</i> -GMNNT | emergency flightattendant declare cabin passenger gate engine cockpit smoke |

4.2.1 Evaluation on held-out likelihood

Perplexity The perplexity is a widely used metric for measuring the held-out likelihood. Considering a test dataset $\widehat{W} = \{\widehat{w}_d\}_{d=1}^{\widehat{D}}$, its perplexity can be computed as follows:

$$\text{Perplexity}(\widehat{W}) = \exp\left(-\frac{\sum_{d=1}^{\widehat{D}} \log(p(\widehat{w}_d))}{\sum_{d=1}^{\widehat{D}} N_d}\right),$$

where $\log(\widehat{w}_d)$ represents the log probability of document \widehat{w}_d . Following [35], we use the variational lower bound to approximate the perplexity.

Results We show the results in Table 5. Overall speaking, our GMNNT with trainable word embeddings achieves the best perplexities among all baselines and the *p*-GMNNT with pre-trained word embeddings ranks the second. According to the average ranks, the performance order of perplexity is given by GMNNT > *p*-GMNNT > NVDM > ETM \approx *p*-ETM > LDA > ProdLDA > GSM. More observations and discussions are detailed below.

Our GMNNT achieves the best perplexities and the two versions both perform well among all settings. GSM has the worst results on short text datasets, *i.e.*, Trec and Stack-Overflow, and LDA also performs badly on these two datasets which coincides with the fact that LDA fails to handle short texts due to the lack of word patterns. As for long text datasets, *i.e.*, Abstract, Tmc2007 and NewYorkTimes, LDA lies in middle position and performs better than ProdLDA and GSM, which gain the worst perplexities on long text datasets. NVDM performs well on long text datasets and the performance gaps comparing with our GMNNT are quite small. NVDM achieves the best perplexity on NewYorkTimes, corresponding with the best topic coherence and topic uniqueness which will be analyzed later. ProdLDA has the worst perplexities on most settings, especially for long text datasets.

Table 5 Experimental results of Perplexity. The lower score means better performance, and the best scores are in boldface

| Dataset | T | LDA | NVDM | ProdLDA | GSM | ETM | p -ETM | GMNNT | p -GMNNT |
|---------------|-----|--------|---------------|---------|--------|--------|----------|---------------|---------------|
| Trec | 25 | 590.3 | 507.4 | 578.9 | 745.7 | 406.7 | 322.6 | 188.0 | 193.5 |
| | 50 | 696.5 | 506.3 | 556.2 | 801.8 | 444.3 | 321.9 | 201.4 | 211.5 |
| StackOverflow | 25 | 1147.0 | 862.6 | 987.0 | 1267.9 | 799.8 | 598.8 | 345.9 | 372.9 |
| | 50 | 1347.2 | 862.9 | 896.5 | 1248.3 | 798.6 | 577.5 | 381.8 | 431.5 |
| Abstract | 25 | 2199.6 | 1824.2 | 3260.6 | 2607.2 | 2159.4 | 2235.5 | 1927.1 | 1742.7 |
| | 50 | 2330.8 | 1828.2 | 3253.2 | 2660.5 | 2246.4 | 2078.1 | 2064.4 | 1928.8 |
| Tmc2007 | 25 | 1507.8 | 1293.0 | 2269.2 | 1693.3 | 1406.5 | 1784.7 | 1195.9 | 1217.7 |
| | 50 | 1701.8 | 1321.0 | 2179.0 | 1774.1 | 1480.3 | 1647.2 | 1298.0 | 1468.3 |
| NewYorkTimes | 100 | 3615.5 | 3040.7 | 10466.7 | 4757.0 | 4382.0 | 4232.1 | 3338.8 | 3278.0 |
| AvgRank | – | 5.00 | 2.63 | 5.82 | 5.91 | 3.55 | 3.55 | 1.36 | 1.63 |

The cause for the extreme high perplexity goes to the high KL-divergence for approximating the Dirichlet prior in the encoder, and these observations are consistent with [8]. We have the same observations on the objective values in Section 4.3.

4.2.2 Evaluation on topic coherence

Topic coherence Broadly speaking, the Topic Coherence (TC) measures the quality of topics by counting the co-occurrences of their top words. In the experiment, we compute the score of topic coherence by using the publicly available project *Palmetto*¹⁴ developed by the previous study [46]. We employ the version of C_V suggested by [46].

Results Table 6 illustrates results of topic coherence. Our p -GMNNT and GMNNT achieve significant improvements over other methods, especially on Trec and Abstract. The performance improvements are up to 0.025 and 0.059 on Trec and Abstract when $K = 25$ and $K = 50$, respectively. Our GMNNT achieves much higher topic coherence scores than ETM, especially on Trec, Abstract and NewYorkTimes, which indicates that our method generates more coherent topics with less redundant words, *e.g.*, semantically related words, which may seldomly co-occur and therefore harm the topic coherence. This observation confirms the effectiveness of our motivation. Besides, both the two versions of GMNNT perform steady results for different topic numbers T . LDA suffers from the sparsity problem on short text datasets and therefore results in the bad performance on Trec and StackOverflow and meanwhile beats most neural topic models on Abstract and Tmc2007.

Meanwhile, we find that high topic coherence scores may not always correspond with higher topic qualities. For NewYorkTimes, NVDM achieves the best topic coherence up to 0.557. Back to the aforementioned topic quality decline problem in Section 4.1, when facing larger datasets, NVDM and ProdLDA tend to generate more infrequent words for each topic, *e.g.*, names of persons or places. These infrequent words may strongly co-occur in external mass corpus however leading to less meaningful topics.

¹⁴<https://github.com/dice-group/Palmetto/wiki/Coherences>

Table 6 Experimental results of topic coherence. The higher score means better performance, and the best scores are in boldface

| Dataset | T | LDA | NVDM | ProdLDA | GSM | ETM | p -ETM | GMNNT | p -GMNNT |
|---------------|-----|-------|-------|---------|--------------|--------------|----------|--------------|--------------|
| Trec | 25 | 0.384 | 0.400 | 0.391 | 0.405 | 0.379 | 0.396 | 0.419 | 0.430 |
| | 50 | 0.378 | 0.391 | 0.389 | 0.381 | 0.374 | 0.396 | 0.399 | 0.418 |
| StackOverflow | 25 | 0.422 | 0.417 | 0.419 | 0.403 | 0.497 | 0.410 | 0.453 | 0.442 |
| | 50 | 0.416 | 0.427 | 0.402 | 0.402 | 0.466 | 0.407 | 0.464 | 0.450 |
| Abstract | 25 | 0.403 | 0.368 | 0.354 | 0.381 | 0.418 | 0.377 | 0.462 | 0.302 |
| | 50 | 0.397 | 0.386 | 0.354 | 0.388 | 0.420 | 0.367 | 0.431 | 0.307 |
| Tmc2007 | 25 | 0.362 | 0.308 | 0.328 | 0.385 | 0.367 | 0.364 | 0.331 | 0.307 |
| | 50 | 0.369 | 0.309 | 0.330 | 0.363 | 0.375 | 0.373 | 0.330 | 0.291 |
| NewYorkTimes | 100 | 0.432 | 0.557 | 0.462 | 0.528 | 0.392 | 0.419 | 0.510 | 0.518 |
| AvgRank | – | 3.82 | 4.00 | 4.82 | 3.55 | 3.00 | 4.00 | 2.18 | 3.91 |

4.2.3 Evaluation on topic uniqueness

Topic Uniqueness Topic Uniqueness (TU) measures the redundancy of top- M words of topics [41]. Given learned top- M words of all topics $\{\Omega_t\}_{t=1}^T$, TU is computed as follows:

$$\mathbf{TU} = \frac{1}{TM} \sum_{t=1}^T \sum_{w_i \in \Omega_t} \frac{1}{\mathbf{cnt}(w_i)}, \quad (9)$$

where $\mathbf{cnt}(w_i)$ denotes the number of times that the word w_i appears in top- M word lists of all topics.

Results Table 7 illustrates the results of topic uniqueness. Overall speaking, our p -GMNNT achieves the best topic uniqueness in most settings and ranks the first. Our p -GMNNT beats p -ETM 7 times on 11 settings and GMNNT with trainable word embeddings beats ETM in all settings, which indicates the effectiveness of topic-specific generation decoder.

Table 7 Experimental results of topic uniqueness. The higher score means better performance, and the best scores are in boldface

| Dataset | T | LDA | NVDM | ProdLDA | GSM | ETM | p -ETM | GMNNT | p -GMNNT |
|---------------|-----|-------|--------------|---------|-------|-------|--------------|-------|--------------|
| Trec | 25 | 0.692 | 0.774 | 0.852 | 0.608 | 0.422 | 0.935 | 0.877 | 0.956 |
| | 50 | 0.624 | 0.646 | 0.723 | 0.614 | 0.239 | 0.888 | 0.520 | 0.729 |
| StackOverflow | 25 | 0.778 | 0.790 | 0.886 | 0.480 | 0.214 | 0.918 | 0.629 | 0.976 |
| | 50 | 0.678 | 0.702 | 0.898 | 0.474 | 0.190 | 0.924 | 0.334 | 0.667 |
| Abstract | 25 | 0.647 | 0.957 | 0.934 | 0.600 | 0.339 | 0.833 | 0.902 | 0.946 |
| | 50 | 0.603 | 0.877 | 0.885 | 0.522 | 0.238 | 0.813 | 0.591 | 0.916 |
| Tmc2007 | 25 | 0.500 | 0.885 | 0.545 | 0.806 | 0.310 | 0.765 | 0.697 | 0.962 |
| | 50 | 0.424 | 0.749 | 0.512 | 0.679 | 0.187 | 0.631 | 0.365 | 0.902 |
| NewYorkTimes | 100 | 0.714 | 0.967 | 0.566 | 0.940 | 0.323 | 0.774 | 0.682 | 0.953 |
| AvgRank | – | 4.45 | 2.27 | 3.18 | 4.45 | 8.00 | 2.45 | 4.64 | 1.45 |

p -ETM performs well on short text datasets while NVDM generates higher topic uniqueness scores on long text datasets. GSM applies the topic uniqueness regularization and achieves 0.940 on NewYorkTimes, while the regularization plays a weak part and has limited effects in reducing topic redundancy. Furthermore, we find that GSM tends to collapse to identical topics with longer training process, and this observation is rather similar with the component collapsing mentioned in [8], from which the VAE family mostly suffer. The reason for component collapsing is that the KL-divergence in objective tends to converge much faster than the reconstruction loss and the model falls into local optima in early training. We indeed have the same observations, where the topic uniqueness scores have the tendency of first rising and then descending. For our GMNNT and p -GMNNT, we apply a relatively high learning rate at 0.0001 and Batch Normalization is employed before the reparameterization trick for avoiding the component collapsing according to [8]. These settings provide gradually rising topic uniqueness values during the entire training process.

4.3 Convergence analysis

In this section, we present the convergence analysis of neural topic models and plot the objective, *i.e.*, negative evidence lower bound (NELBO) values across Trec, StackOverflow, Abstract and Tmc2007 in Figure 2. Overall, our GMNNT and p -GMNNT converge fast within 40 epochs. Due to the smaller learning rate comparing with our baselines, our GMNNT and p -GMNNT converge a little slower than other methods. ProdLDA converges hard due to the approximation of Dirichlet prior. In summary, GMNNT is more practical in real applications due to its fast convergence.

4.4 Sensitivity analysis of parameters

In this section, we examine the impacts of number of layers λ_d and number of hidden neurons λ_w in neural topics by perplexity, topic coherence and topic uniqueness on p -GMNNT. We study the sensitivity experiment on Trec, StackOverflow, Abstract and Tmc2007 for memory limitation.

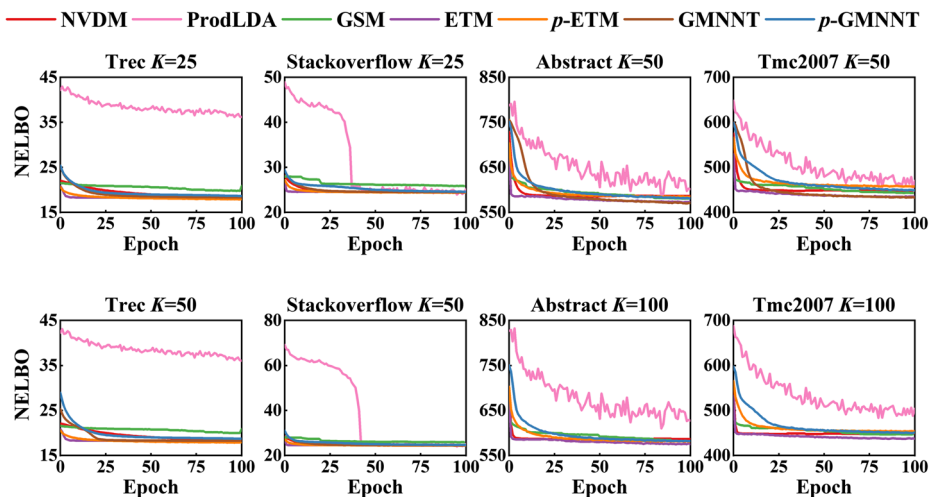


Figure 2 Convergence curves of NELBO values on Trec, StackOverflow, Abstract and Tmc2007

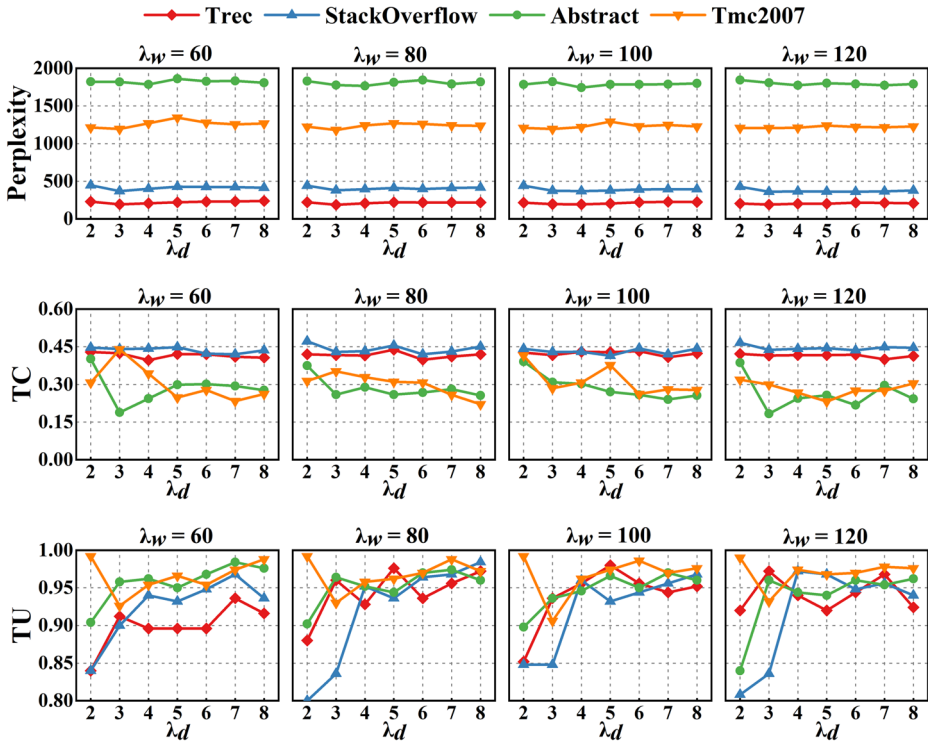


Figure 3 Perplexity, topic coherence and topic uniqueness performance varying λ_d on Trec, StackOverflow, Abstract and Tmc2007

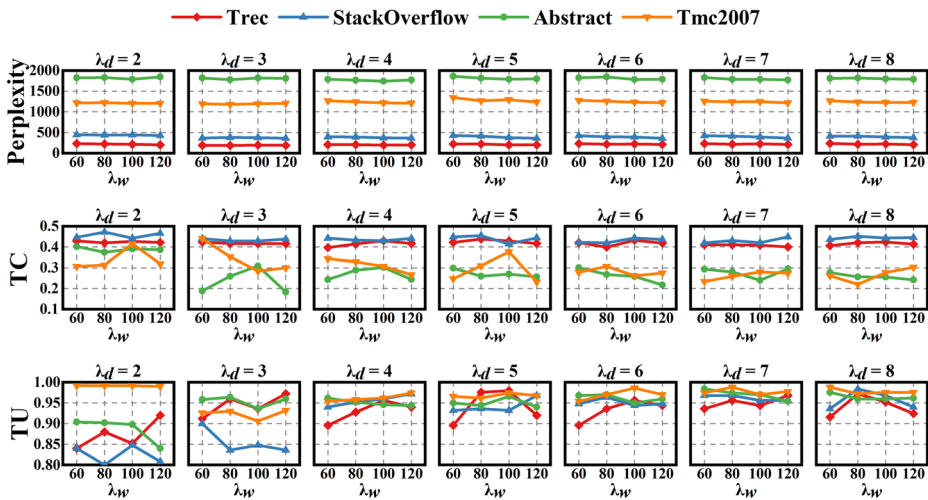


Figure 4 Perplexity, topic coherence and topic uniqueness performance varying λ_w on Trec, StackOverflow, Abstract and Tmc2007

For λ_d , we varied it from $\{2, 3, 4, 5, 6, 7, 8\}$ and present results in Figure 3. In general, a more complicated network leads to a worse perplexity under same settings. We find that for all four datasets, perplexities of p -GMNNT show little change and a slight rising trend. As for topic coherence, p -GMNNT shows insensitivity on Trec and StackOverflow and meanwhile shows slight decrements for Abstract and Tmc2007 when facing deeper networks. As for topic uniqueness, a deeper network mostly leads to better topic uniqueness scores for most datasets, which coincides with the fact that topic coherence and topic uniqueness are more likely to have opposite tendencies. In a word, GMNNT is insensitive to the number of layers in neural topic, making the model more robust in real scenarios. For λ_w , we varied it from $\{60, 80, 100, 120\}$ and plot results in Figure 4. In keeping with λ_d , perplexities shows small fluctuations on λ_w . Topic coherence scores on Trec and StackOverflow are stable and insensitive to λ_w and a wider network leads to higher topic uniqueness scores. In brief, our GMNNT shows great robust on λ_d and λ_w , which makes it practical in real-world applications.

5 Conclusion

In this paper, we aim at alleviating the problem that the existing neural topic models often output redundant and inaccurate topic representations. To this end, we suggest a novel GMNNT model by replacing the topic embeddings with neural topics, enabling to capture nonlinear relationships between words in the embedding space. With this design, even similar words are allowed with different probabilities in the same topic distribution, leading to more accurate topic representations. By employing the spirit of neural variational inference, we can efficiently train GMNNT with the reparameterization trick. We conduct numbers of experiments to empirically compare GMNNT against several existing neural topic models and also the standard LDA model. Experimental results show that GMNNT is on a par with the existing baseline models over the evaluations of held-out likelihood, topic coherence and topic uniqueness. Specifically, the topics learned by GMNNT are more semantically coherent both qualitatively and quantitatively.

Acknowledgements This work was supported by the National Natural Science Foundation of China (NSFC) [No.61876071] and Scientific and Technological Developing Scheme of Jilin Province [No.20180201003SF, No.20190701031GH] and Energy Administration of Jilin Province [No.3D516L921421].

References

1. Batmanghelich, K., Saedi, A., Narasimhan, K., Gershman, S.: Nonparametric spherical topic modeling with word embeddings. In: Annual Meeting of the Association for Computational Linguistics, pp. 537–542 (2016)
2. Blei, D.M.: Probabilistic topic models. *Communications of The ACM* **55**(4), 77–84 (2012)
3. Blei, D.M., Kucukelbir, A., Mcaliffe, J.: Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**(518), 859–877 (2017)
4. Blei, D.M., Lafferty, J.: Dynamic topic models. In: International Conference on Machine Learning, pp. 113–120 (2006)
5. Blei, D.M., Lafferty, J.: A correlated topic model of science. *The Annals of Applied Statistics* **1**(1), 17–35 (2007)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**(Jan), 993–1022 (2003)

7. Boyd-Graber, J., Hu, Y., Mimno, D.: Applications of topic models. *Foundations and Trends in Information Retrieval* **11**(2-3), 143–296 (2017)
8. Burkhardt, S., Kramer, S.: Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.* **20**(131), 1–27 (2019)
9. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. *IEEE Trans. Knowl. Data Eng.* **26**(12), 2928–2941 (2014)
10. Cong, Y., Chen, B., Liu, H., M.Z.: Deep latent dirichlet allocation with topic-layer-adaptive stochastic gradient riemannian MCMC. In: *International Conference on Machine Learning*, pp. 864–873 (2017)
11. Das, R., Zaheer, M., Dyer, C.: Gaussian LDA for topic models with word embeddings. In: *International Joint Conference on Natural Language Processing*, pp. 795–804 (2015)
12. Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: The dynamic embedded topic model. [arXiv:1907.05545](https://arxiv.org/abs/1907.05545) (2019)
13. Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* **8**, 439–453 (2020)
14. Dozat, T.: Incorporating nesterov momentum into adam. In: *International Conference on Learning Representations Workshop* (2016)
15. Duchi, J.C., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**(7), 2121–2159 (2011)
16. Feng, J., Rao, Y., Xie, H., Wang, F., Li, Q.: User group based emotion detection and topic discovery over short text. *World Wide Web* **23**, 1553–1587 (2020)
17. Gershman, S., Goodman, N.D.: Amortized inference in probabilistic reasoning. In: *Annual Meeting of the Cognitive Science Society* (2014)
18. Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proc. Natl. Acad. Sci. U.S.A.* **101**(suppl 1), 5228–5235 (2004)
19. Gui, L., Leng, J., Pergola, G., Zhou, Y., Xu, R., He, Y.: Neural topic model with reinforcement learning. In: *Conference on Empirical Methods in Natural Language Processing*, pp. 3478–3483 (2019)
20. Hoffman, M.D., Bach, F., Blei, D.M.: Online learning for latent dirichlet allocation. In: *Neural Information Processing Systems*, pp. 856–864 (2010)
21. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. *J. Mach. Learn. Res.* **14**(1), 1303–1347 (2013)
22. Isonuma, M., Mori, J., Bollegala, D., Sakata, I.: Tree-structured neural topic model. In: *Annual Meeting of the Association for Computational Linguistics*, pp. 800–806 (2020)
23. Jiang, H., Zhou, R., Zhang, L., Wang, H., Zhang, Y.: Sentence level topic models for associated topics extraction. *World Wide Web* **22**(6), 2545–2560 (2019)
24. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. *Mach. Learn.* **37**(2), 105–161 (1999)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: *International Conference on Learning Representations* (2015)
26. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: *International Conference on Learning Representations* (2014)
27. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: *International Conference on Machine Learning*, pp. 577–584 (2006)
28. Li, X., Chi, J., Li, C., Ouyang, J., Fu, B.: Integrating topic modeling with word embeddings by mixtures of VMFs. In: *International Conference on Computational Linguistics*, pp. 151–160 (2016)
29. Li, X., Ouyang, J., Zhou, X.: Sparse hybrid variational-gibbs algorithm for latent dirichlet allocation. In: *SIAM International Conference on Data Mining*, pp. 729–737 (2016)
30. Li, X., Zhang, A., Li, C., Ouyang, J., Cai, Y.: Exploring coherent topics by topic modeling with term weighting. *Inform. Process. Manage.* **54**(6), 1345–1358 (2018)
31. Li, X., Zhang, J., Ouyang, J.: Dirichlet multinomial mixture with variational manifold regularization: Topic modeling over short texts. In: *AAAI Conference on Artificial Intelligence*, pp. 7884–7891 (2019)
32. Li, Z., Wang, X., Li, J., Zhang, Q.: Deep attributed network representation learning of complex coupling and interaction. *Knowl.-Based Syst.* **212**, 106,618 (2021)
33. Liu, L., Huang, H., Gao, Y., Zhang, Y., Wei, X.: Neural variational correlated topic modeling. In: *The Web Conference*, pp. 1142–1152 (2019)
34. Miao, Y., Grefenstette, E., Blunsom, P.: Discovering discrete latent topics with neural variational inference. In: *International Conference on Machine Learning*, pp. 2410–2419 (2017)
35. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: *International Conference on Machine Learning*, pp. 1727–1736 (2016)
36. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: *International Conference on Learning Representations* (2013)

37. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Neural Information Processing Systems*, pp. 3111–3119 (2013)
38. Mimno, D.M., Hoffman, M.D., Blei, D.M.: Sparse stochastic inference for latent dirichlet allocation. In: *International Conference on Machine Learning*, pp. 1515–1522 (2012)
39. Mnih, A., Gregor, K.: Neural variational inference and learning in belief networks. In: *International Conference on Machine Learning*, pp. 1791–1799 (2014)
40. Naesseth, C.A., Ruiz, F.J.R., Linderman, S.W., Blei, D.M.: Reparameterization gradients through acceptance-rejection sampling algorithms. In: *International Conference on Artificial Intelligence and Statistics*, pp. 489–498 (2017)
41. Nan, F., Ding, R., Nallapati, R., Xiang, B.: Topic modeling with wasserstein autoencoders. In: *Annual Meeting of the Association for Computational Linguistics*, pp. 6345–6381 (2019)
42. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: *Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543 (2014)
43. Pergola, G., Gui, L., He, Y.: Tdam: a topic-dependent attention model for sentiment analysis. *Inform. Process. Manage.* **56**(6), 102,084 (2019)
44. Rashid, J., Shah, S.M.A., Irtaza, A.: Fuzzy topic modeling approach for text mining over short text. *Inform. Process. Manage.* **56**(6), 102,060 (2019)
45. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In: *International Conference on Machine Learning*, pp. 1278–1286 (2014)
46. Röder, M., Both, A., Hinneburg, A.: Exploring the Space of Topic Coherence Measures. In: *International Conference on Web Search and Data Mining*, pp. 399–408 (2015)
47. Srivastava, A., Sutton, C.A.: Autoencoding Variational Inference for Topic Models. In: *International Conference on Learning Representations* (2017)
48. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* **101**(476), 1566–1581 (2006)
49. Titsias, M., Lázaro-Gredilla, M.: Doubly stochastic variational bayes for non-conjugate inference. In: *International Conference on Machine Learning*, pp. 4056–4069 (2014)
50. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: *International Conference on Learning Representations* (2018)
51. Wang, C., Blei, D.M., Heckerman, D.: Continuous time dynamic topic model. In: *Uncertainty in Artificial Intelligence*, pp. 579–586 (2008)
52. Wang, R., Zhou, D., He, Y.: Atm: Adversarial-neural topic model. *Inform. Process. Manage.* **56**(6), 102,098 (2019)
53. Wang, Y., Li, X., Ouyang, J.: Layer-assisted neural topic modeling over document networks. In: *International Joint Conference on Artificial Intelligence*, pp. 3148–3154 (2021)
54. Zhang, H., Chen, B., Guo, D., Zhou, M.: WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In: *International Conference on Learning Representations* (2018)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.