Supervised Neural Topic Modeling with Label Alignment

Ruihao Chen[1], **Hegang Chen**[1], **Yuyin Lu**[1], **Yanghui Rao**[1]*, **Chunjiang Zhu**[2]*
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Department of Computer Science, University of North Carolina at Greensboro, USA
{chenrh29,chenhg25,luyy37}@mail2.sysu.edu.cn,
raoyangh@mail.sysu.edu.cn,chunjiang.zhu@uncg.edu

Abstract

Neural topic modeling is a scalable automated technique for text data mining. In various downstream tasks of topic modeling, it is preferred that the discovered topics well align with labels. However, due to the lack of guidance from labels, unsupervised neural topic models are less powerful in this situation. Existing supervised neural topic models often adopt a label-free prior to generate the latent document-topic distributions and use them to predict the labels and thus achieve label-topic alignment indirectly. Such a mechanism faces the following issues: 1) The label-free prior leads to topics blending the latent patterns of multiple labels; and 2) One is unable to intuitively identify the explicit relationships between labels and the discovered topics. To tackle these problems, we develop a novel supervised neural topic model which utilizes a chain-structured graphical model with a label-conditioned prior. Soft indicators are introduced to explicitly construct the label-topic relationships. To obtain well-organized label-topic relationships, we formalize an entropy-regularized optimal transport problem on the embedding space and model them as the transport plan. Moreover, our proposed method can be flexibly integrated with most existing unsupervised neural topic models. Experimental results on multiple datasets demonstrate that our model can greatly enhance the alignment between labels and topics while maintaining good topic quality.

1 Introduction

Topic modeling (Blei et al., 2001; Srivastava and Sutton, 2017) aims to uncover the latent semantic structure within a corpus, which has been applied in various downstream tasks, such as social event analysis (Xue et al., 2020) and opinion mining (Vamshi et al., 2018). A topic is typically described as a probability distribution over words. When conducting topic modeling, we often focus on the extent to which the discovered topics are meaningful and easy to understand, i.e., the interpretability of topics. To assess it, the topic coherence and many other evaluation have been developed (Lau et al., 2014; Röder et al., 2015). However, a high coherence score of topics is often insufficient to indicate that they are ‘‘good’’ topics (Hoyle et al., 2021, 2022). Considering that the human-defined labels describe some inherent patterns within the documents. **we expect to identify the relationships between labels and the discovered topics, namely, achieve label-topic alignment.** This will facilitate better understanding of the discovered topics, especially in the area of content analysis (Hoyle et al., 2022). For example, the movie reviews in the IMDB dataset (Maas et al., 2011) are categorized by their sentiment orientations (positive or negative), and we are interested in what a ‘‘positive/negative’’ topic is like. Unfortunately, due to the lack of guidance from manually defined labels, it is challenging for topics discovered by unsupervised topic models to be consistent with them. In response to this, several supervised topic models have been developed (Mcauliffe and Blei, 2007; Card et al., 2018). The fundamental idea behind most of them is to predict labels through document-topic distributions. However, they are faced with the following challenges: 1) A label-free prior $p(\mathbf{z})$, which treats each topic evenly, is often adopted to generate the latent topics z_i for each word (see Figure 1a). As the generated topics are not specific to the label y, this will blend the semantics of multiple categories. 2) Under the label-free prior, they are not able to construct the label-topic relationships explicitly and thus fail to provide an intuitive view of the relationships between labels and topics. While there are statistical topic models based

* Corresponding authors.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 249–263, 2025. https://doi.org/10.1162/tacl_a_00738
Action Editor: Lidong Bing. Submission batch: 7/2024; Revision batch: 11/2024; Published 3/2025.

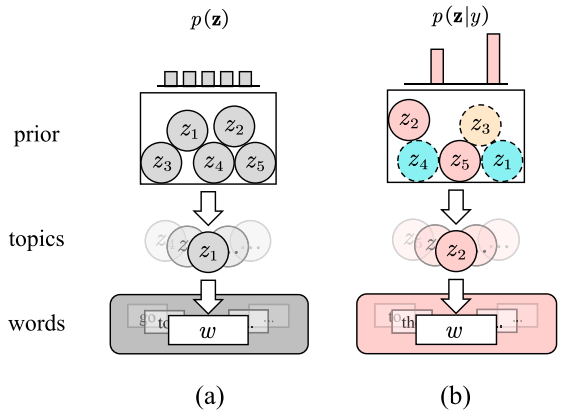


Figure 1: The generative process of documents under (a) label-free prior $p(\mathbf{z})$ and (b) label-conditioned prior $p(\mathbf{z}|y)$, where w, z, y represent words, topics, and labels, respectively, and different colors represent different labels.

on Gibbs sampling that explicitly construct the label-topic relationships (Ramage et al., 2009), they suffer from inflexible model-specific derivations and slow inference speed. They are not neural topic models (NTMs) and do not leverage the advantages of deep generative models like variational autoencoder (VAE) (Kingma and Welling, 2014; Srivastava and Sutton, 2017). Given the challenges mentioned above, it is desirable that **a label-conditioned prior is utilized**. Under such a prior $p(\mathbf{z}|y)$, label-specific topics will be generated (see Figure 1b). Motivated by this, we propose a novel supervised NTM, **Label Aligned Neural Topic Model** (**LANTM**).[1] First, within the framework of VAE, we introduce a novel chain-structured graphical model that incorporates a label-conditioned prior. Second, we propose to conduct inference under soft label-topic indicators. In order to obtain well-organized label-topic relationships, we then formalize an entropy-regularized Optimal Transport (OT) problem on the embedding space and model the label-topic relationships as the transport plan. Besides, we propose to reconstruct the label-specific pseudo-documents to allow the label-topic distributions to capture the semantics of labels. The main contributions of this paper are as follows:

- We propose a novel graphical model for supervised neural topic modeling, where latent topics are generated from a label-conditioned prior. The inference is conducted under soft indicators that explicitly describe the relationships between labels and topics.
- We formalize an entropy-regularized OT problem on the embedding space and model the label-topic relationships as the transport plan to obtain well-organized label-topic relationships. To enable the label-topic distributions to capture the label semantics, we propose to reconstruct the pseudo-documents specific to each label.
- Our proposed model can be flexibly integrated with most existing unsupervised NTMs based on VAE.
- Experimental results on multiple datasets show that our proposed model can achieve better label-topic alignment without compromising the topic quality.

[1] Our code is available at `https://github.com/Rio-Chan-0119/LANTM`.

2 Related Work

2.1 Neural Topic Model

Most statistical topic models like LDA (Blei et al., 2001) and HDP (Teh et al., 2004) treat topics as latent variables and infer model parameters using Gibbs sampling. Their most significant issues lie in the need for model-specific derivations and long training time. NTMs (Miao et al., 2016; Srivastava and Sutton, 2017; Miao et al., 2017) overcome these issues by utilizing VAE. To take a step further, ETM (Dieng et al., 2020) enhances the expressive power of topic models by mapping topics and words into the same embedding space. Based on ETM, numerous NTMs have been developed (Zhao et al., 2021; Wang et al., 2022; Wu et al., 2023). ECRTM (Wu et al., 2023) introduces embedding cluster regularization to alleviate the well-known issue of topic collapsing in NTMs.

2.2 Supervised Topic Model

A supervised topic model involves document metadata (not limited to document categories) during training. sLDA (Mcauliffe and Blei, 2007) is a classical supervised topic model where metadata is generated from the empirical topic mixture distribution of documents. Based on sLDA, DiscLDA (Lacoste-Julien et al., 2008), and MedLDA (Zhu et al., 2009) conduct training in different manners. The former attempts to maximize the

conditional likelihood of metadata, while the latter jointly trains LDA with SVM utilizing the max-margin principle. The supervised topic models mentioned above do not explicitly establish the label-topic relationships, while Labeled LDA (Ramage et al., 2009) does. Labeled LDA assumes a one-to-one correspondence between labels and topics. In addition, hLLDA (Petinot et al., 2011), HSLDA (Perotte et al., 2011), and SSHLDA (Mao et al., 2012) consider the association of hierarchical labels with latent topics.

Given the greater advantages of neural topic models compared to statistical ones, there is also a movement towards combining neural networks with supervised topic models. sNTM (Cao et al., 2015) and SLTM (Huang et al., 2018) learn topics in a non-Bayesian manner utilizing neural networks. sNTM uses a fully connected network to predict labels from document-topic distributions, while SLTM employs a Siamese network to explicitly distinguish the current document from negative sample documents during training. SCHOLAR (Card et al., 2018) simultaneously reconstructs documents and labels. Similar to most supervised topic models, SCHOLAR predicts labels using document-topic distributions, indirectly aligning the inferred topics with labels. In contrast, our proposed model assumes that topics are under the control of labels and explicitly constructs the label-topic relationships.

3 Background

Consider a document collection $\mathcal{X}$ with V unique words. Each document in $\mathcal{X}$ is typically represented by a bag-of-words (BoW) vector $\mathbf{x} \in \mathbb{R}^V$ and a latent distribution over K topics (i.e., the document-topic distribution), denoted by $\mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$. Each topic describes a specific semantic concept, and the k-th topic is regarded as a distribution over words (i.e., the topic-word distribution), denoted by $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,V}) \in \mathbb{R}^V$. State-of-the-art unsupervised NTMs (Dieng et al., 2020; Zhao et al., 2021; Wu et al., 2023) utilize topic embeddings $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_K) \in \mathbb{R}^{H \times K}$ and word embeddings $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_V) \in \mathbb{R}^{H \times V}$ in the same H-dimensional embedding space $\mathcal{E}$ to compute $\beta_{k,v}$ according to the similarity metric (like inner product) between $\mathbf{t}_k$ and $\mathbf{w}_v$.

NTMs utilize neural networks to infer the latent document-topic distribution (Srivastava and Sutton, 2017). Most existing NTMs are based on the VAE framework, where latent variables are interpreted as topics. An unsupervised VAE-based NTM often firstly assumes that the latent document-topic distribution $\mathbf{z}$ lying within a probability simplex $\Delta^K = \{\mathbf{z} \in \mathbb{R}^K_{\geq 0} | \sum_k z_k = 1\}$ is drawn from a label-free prior $p(\mathbf{z})$, then it generates the BoW vector $\mathbf{x}$ conditioned on $\mathbf{z}$, i.e., from $p(\mathbf{x}|\mathbf{z})$. In this setting, the joint distribution $p(\mathbf{x}, \mathbf{z})$ can be factorized by $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$. To approximate the intractable posterior distribution and thus calculate ELBO, the variational distribution $q(\mathbf{z}|\mathbf{x})$ is introduced in VAE.

Supervised NTMs involve the label of each training document (e.g., sentiment) in the generative process. Suppose that the label set $\mathcal{Y}$ has C labels. Consider the observations $\mathcal{O} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ in this context, where $y \in \mathcal{Y}$ denotes the assigned single label of $\mathbf{x}$. SCHOLAR (Card et al., 2018), a classic supervised NTM utilizing a forked graphical model $y \leftarrow \mathbf{z} \rightarrow \mathbf{x}$, assumes that labels are generated from topics.[2] This implies that $\mathbf{z}$ is drawn from a label-free prior $p(\mathbf{z})$ without being constrained by y in the generative process.

It is questionable whether such $\mathbf{z}$ in SCHOLAR can actually capture topic information aligned with y. This motivates us to rethink the generative process and propose a novel supervised NTM, where $\mathbf{z}$ is drawn from a label-conditioned prior $p(\mathbf{z}|y)$.

4 LANTM: Label Aligned Neural Topic Model

In this section, we propose LANTM, a novel supervised NTM for label-aligned topic discovery. We start by discussing the generative process. We then introduce soft indicators to model the variational distribution and the label-conditioned prior. In order to enforce well-organized label-topic relationships, we formalize an entropy-regularized Optimal Transport (OT) problem and model the label-topic relationships as the transport plan. Ultimately, we describe the parameter inference for LANTM, and discuss how to conduct inference in the absence of labels.

[2] In fact, supervised signals can serve as either labels or covariates in SCHOLAR. In this paper we focus on the former scenario and disregard metadata serving as covariates.

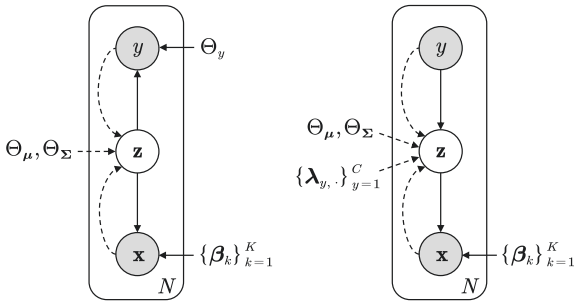


Figure 2: Graphical models of (a) SCHOLAR and (b) LANTM (ours). Solid arrows denote generative model and dashed arrows denote inference.

4.1 Generative Process

Different from the forked graphical model adopted by SCHOLAR, we introduce the label guidance by setting the graphical model of LANTM to be a *chain*, i.e., $y \rightarrow \mathbf{z} \rightarrow \mathbf{x}$. Thus, the joint distribution $p(\mathbf{x}, y, \mathbf{z})$ can be factorized as follows:

$$p(\mathbf{x}, y, \mathbf{z}) = p(y)p(\mathbf{z}|y)p(\mathbf{x}|\mathbf{z}). \tag{1}$$

Figure 2 shows the graphical models of the classic supervised NTM, i.e., SCHOLAR (Card et al., 2018), and our proposed LANTM. The motivation behind the chain-structured graphical model adopted by our LANTM is that the topics are generated under a label-conditioned prior so that they are specific to the label.

Next we deal with the objective function. We define $p(y)$ to be a discrete uniform distribution:

$$p(y) = \text{Cat}(\mathbf{1}_C/C), \tag{2}$$

where $\text{Cat}(\cdot)$ denotes the categorical distribution. $\mathbf{1}_C$ is a C-dimensional all-one vector. With this definition, we use the following loss function $\mathcal{L}(\mathbf{x}, y)$ to maximize ELBO:

$$\begin{aligned}\mathcal{L}(\mathbf{x}, y) = D_{\text{KL}}[q(\mathbf{z}|\mathbf{x}, y) \parallel p(\mathbf{z}|y)] \\ - \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)}[\log p(\mathbf{x}|\mathbf{z})],\end{aligned} \tag{3}$$

where D_{KL} denotes the KL divergence between two distributions. We can easily derive the ELBO (the derivation is given in Appendix A): $\log(\mathbf{x}, y) \geq -\mathcal{L}(\mathbf{x}, y) + \log p(y)$, where $\log p(y)$ is a constant. $\mathcal{L}(\mathbf{x}, y)$ is similar in form to the loss function of the original VAE (Kingma and Welling, 2014), as both are composed of a KL divergence term and a reconstruction term, and document reconstruction relies solely on latent variables $\mathbf{z}$. The difference between these two lies in that our supervised model utilizes a variational distribution and a prior additionally conditioned on label y, namely, $q(\mathbf{z}|\mathbf{x}, y)$ and $p(\mathbf{z}|y)$.

Intuitively, the KL divergence term in Eq. (3) drives $q(\mathbf{z}|\mathbf{x}, y)$ towards $p(\mathbf{z}|y)$ which varies for different labels. In contrast, SCHOLAR minimizes the KL divergence between $q(\mathbf{z}|\mathbf{x}, y)$ and $p(\mathbf{z})$, where $p(\mathbf{z})$ is typically an isotropic Gaussian distribution. Such a label-free prior fails to distinguish the document-topic distributions of documents with different labels, making it challenging to learn topics that align with the labels.

4.2 Inference with Soft Indicators

In this section, we present the formulations of $q(\mathbf{z}|\mathbf{x}, y)$ and $p(\mathbf{z}|y)$. How to maximally distinguish documents with different labels on the probability simplex Δ^K? A natural idea is to ensure that each topic is activated only when a specific label is observed during inference.

To this end, we introduce a soft indicator $\lambda_{y,k} \in [0, 1]$ to represent the degree to which the k-th topic is activated given the associated label y.[3] A value of $\lambda_{y,k} = 1$ means that the k-th topic is fully activated, whereas $\lambda_{y,k} = 0$ indicates the k-th topic is dummy under label y. Then, we model $q(\mathbf{z}|\mathbf{x}, y)$ and $p(\mathbf{z}|y)$ by

$$q(\mathbf{z}|\mathbf{x}, y) = \mathcal{LN}(\boldsymbol{\mu} + \ln \boldsymbol{\lambda}_{y,\cdot}, \boldsymbol{\Sigma}) \tag{4}$$

$$p(\mathbf{z}|y) = \mathcal{LN}(\boldsymbol{\mu}_0 + \ln \boldsymbol{\lambda}_{y,\cdot}, \boldsymbol{\Sigma}_0), \tag{5}$$

where $\mathcal{LN}(\cdot, \cdot)$ denotes the logistic normal distribution. $\boldsymbol{\lambda}_{y,\cdot} = (\lambda_{y,1}, \ldots, \lambda_{y,K})$ is the indicator vector for y with respect to each topic. $\boldsymbol{\mu} = f_{\boldsymbol{\mu}}(\mathbf{x}; \Theta_{\boldsymbol{\mu}})$ and $\boldsymbol{\Sigma} = f_{\boldsymbol{\Sigma}}(\mathbf{x}; \Theta_{\boldsymbol{\Sigma}})$ are the mean and diagonal covariance matrix encoded by $\mathbf{x}$, while $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are determined by Laplace approximation (Hennig et al., 2012) with hyperparameter α.

To see why it works, consider the latent normal logits $\mathbf{r} = (r_1, \ldots, r_K) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ that are independent of y. We draw the document-topic distribution $\mathbf{z}$ from $q(\mathbf{z}|\mathbf{x}, y)$ with $\mathbf{r}$ by

$$\begin{aligned}z_k &= [\text{Softmax}(\mathbf{r} + \ln \boldsymbol{\lambda}_{y,\cdot})]_k \\ &= \frac{\lambda_{y,k} \exp(r_k)}{\sum_{k'=1}^K \lambda_{y,k'} \exp(r_{k'})}.\end{aligned} \tag{6}$$

[3] With a slight abuse of notation, we use y to denote the corresponding label's index.

We can see that, after the softmax transformation, the additive term $\ln \lambda_{y,k}$ becomes the coefficient for $\exp(r_k)$. This coefficient dominates the allocation of each topic in $\mathbf{z}$, with topics not under y being assigned almost negligible weights.

4.3 Embedding-based Label-topic Relationship Modeling

We can determine the label-topic relationships in a flexible manner. For instance, for the label-topic indicator matrix $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\cdot,1}, \ldots, \boldsymbol{\lambda}_{\cdot,K}) \in [0,1]^{C\times K}$ where $\boldsymbol{\lambda}_{\cdot,k} = (\lambda_{1,k}, \ldots, \lambda_{C,k})$, we can set it as a hyperparameter, or treat it as a learnable parameter. Considering the recent advance in topic modeling that uses the similarity between topic and word embeddings to compute the topic-word distribution $\{\boldsymbol{\beta}_k\}_{k=1}^K$ (Dieng et al., 2020), we decide to utilize label embeddings $\mathbf{L} = (\mathbf{l}_1, \ldots, \mathbf{l}_C) \in \mathbb{R}^{H\times C}$ on the same embedding space $\mathcal{E}$ as topic embeddings to compute $\boldsymbol{\lambda}$.

However, unlike the topic-word distribution, we expect the topics to capture the semantics specific to each label as much as possible, thereby achieving alignment between labels and topics. Also, the emergence of super categories that dominate the majority of topics is undesirable, as it may do harm to the topic discovery of other categories. So we define $\boldsymbol{\lambda}$ as *well-organized* if

1. (*Sparsity*) $\boldsymbol{\lambda}_{\cdot,k}^T \mathbf{1}_C = 1$ and $\boldsymbol{\lambda}_{\cdot,k}$ is sparse for all $k \in [1, 2, \ldots, K]$.
2. (*Balanced Assignment*) $N_y = \boldsymbol{\lambda}_{y,\cdot}^T \mathbf{1}_K$ is relatively balanced across different y's, where N_y is the ''number'' of topics assigned to label y.

Sparsity of well-organized $\boldsymbol{\lambda}$ means that each topic is assigned to only one label, while balanced assignment aims to avoid super categories. Inspired by Transport Plan Dependency proposed in Wu et al. (2024), we formalize an entropy-regularized Optimal Transport (OT) problem (Cuturi, 2013) in Section 4.3.1 on the embedding space $\mathcal{E}$ and use the transport plan to obtain well-organized $\boldsymbol{\lambda}$.

4.3.1 Formalization of the OT Problem

Consider two probability measures: the source measure $\boldsymbol{\tau} = \sum_{y=1}^C s_y \delta_{\mathbf{l}_y}$ on label embeddings $\{\mathbf{l}_y\}_{y=1}^C \subset \mathcal{E}$ and the target measure $\boldsymbol{\nu} = \sum_{k=1}^K t_k \delta_{\mathbf{t}_k}$ on topic embeddings $\{\mathbf{t}_k\}_{k=1}^K \subset \mathcal{E}$, where δ_x denotes the Dirac delta funtion on x. Under a cost matrix $\mathbf{C} \in \mathbb{R}_{\geq 0}^{C\times K}$, we aim to find a transport plan $\mathbf{P}^* \in [0,1]^{C\times K}$ that meets the marginal constraints and minimizes the objective function which includes both the total transport cost and an entropic regularization term:

$$\mathbf{P}^* = \underset{\mathbf{P}\in[0,1]^{C\times K}}{\arg\min} \langle \mathbf{P}, \mathbf{C}\rangle - \frac{1}{\epsilon} h(\mathbf{P}) \quad \text{s.t.} \quad \mathbf{P}\mathbf{1}_K = \mathbf{s}, \mathbf{P}^T\mathbf{1}_C = \mathbf{t}, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. $1/\epsilon > 0$ is the weight of the entropic regularization term. $h(\mathbf{P}) = -\sum_{y,k} \mathbf{P}_{yk} \log \mathbf{P}_{yk}$ is the entropy of the joint distribution identified by $\mathbf{P}$. $\mathbf{s} = (s_1, \ldots, s_C)$ and $\mathbf{t} = (t_1, \ldots, t_K)$ are probability vectors corresponding to $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$, respectively. We use Sinkhorn's Algorithm (Cuturi, 2013) to compute $\mathbf{P}^*$, which is an algorithm for computing entropy-regularized transport and is executed effectively with GPUs. We provide the details of Algorithm 1 in Appendix B.

4.3.2 Modeling the Label-topic Relationship as the Transport Plan

The optimal transport plan $\mathbf{P}^*$ maintains the following properties: 1) $\mathbf{P}^*$ is sparse enough, as the original OT problem is framed as a linear program. 2) $\mathbf{P}^*$ meets the preset marginal constraints which control the assignment. These properties make $\mathbf{P}^*$ a potential candidate for well-organized $\boldsymbol{\lambda}$. To better align $\mathbf{P}^*$ with well-organized $\boldsymbol{\lambda}$, we set the probability vectors as follows:

$$\mathbf{s} = \frac{1}{|\mathcal{X}|}(|\mathcal{X}_1|, \ldots, |\mathcal{X}_C|), \tag{8}$$

$$\mathbf{t} = (1/K, \ldots, 1/K), \tag{9}$$

where $\mathcal{X}_y$ is the collection of documents with label y. Empirically, Eq. (8) requires the number of topics under each label follows the category distribution of the dataset, while Eq. (9) treats each topic evenly.

It is natural to use the distance metric $\text{dist}(\cdot, \cdot)$ on $\mathcal{E}$ as the cost: $\mathbf{C}_{y,k} = \text{dist}(\mathbf{l}_y, \mathbf{t}_k)$. The choice of $\text{dist}(\cdot, \cdot)$ depends on how topic-word distributions are computed. For instance, ETM (Dieng et al., 2020) utilizes inner product to measure the similarity between embeddings. In this situation, we define $\text{dist}(\mathbf{l}_y, \mathbf{t}_k) = \text{Softplus}(-\mathbf{l}_y^T \mathbf{t}_k)$. If the Gaussian kernel is used like ECRTM (Wu et al., 2023), then we may simply set $\text{dist}(\mathbf{l}_y, \mathbf{t}_k) = \|\mathbf{l}_y - \mathbf{t}_k\|_2^2$.

Based on the settings mentioned above, we model the label-topic relationships as the transport plan defined in Eq. (7). The well-organized $\boldsymbol{\lambda}$ is obtained by $\boldsymbol{\lambda} = K\mathbf{P}^*$, where the constant K ensures $\boldsymbol{\lambda}_{\cdot,k}^T \mathbf{1}_C = 1$ holds for all $k \in [1, 2, \ldots, K]$.

4.3.3 Pseudo-document Guided Label-topic Relationship Learning

We discuss how to get well-organized label-topic relationships as above but still remain a problem unsolved: We do not take the semantics behind the labels into consideration.

Specifically, we expect $\boldsymbol{\lambda}_{y,\cdot}$ to reliably describe the impact of each topic within label y. We define normalized $\boldsymbol{\lambda}_{y,\cdot}$, namely, $\boldsymbol{\lambda}_{y,\cdot}/N_y$, as the *label-topic distribution* for y. It is similar to the document-topic distribution. However, the document-topic distributions interact with the topic-word distributions by reconstructing BoWs and therefore capture semantics of the documents, while there is no connection between label-topic distributions and topic-word distributions. This explains the absence of the label semantics and makes $\boldsymbol{\lambda}_{y,\cdot}$ unable to reliably reflect the impact of each topic within y. Following the document-topic distributions, we propose to reconstruct the y-specific pseudo-documents, denoted as $\mathbf{x}_{\text{pseu}}^y \in \mathbb{R}^V$, to enable the label-topic distribution to capture corresponding semantics and thus become more reasonable. $\mathbf{x}_{\text{pseu}}^y$ should consider all the documents with label y. Intuitively, we let $\mathbf{x}_{\text{pseu}}^y$ be proportional to the term frequency of $\mathcal{X}_y$ (recall that $\mathcal{X}_y$ is the collection of documents with label y). Additionally, to ensure consistent scale, we let the length of $\mathbf{x}_{\text{pseu}}^y$ be $N_{\text{avg.}}$, where $N_{\text{avg.}}$ is the average length of the documents across the whole corpus. Formally, we define $\mathbf{x}_{\text{pseu}}^y$ as follows:

$$\mathbf{x}_{\text{pseu}}^y = \frac{\sum_{\mathbf{x}\in\mathcal{X}_y} \mathbf{x}}{(\sum_{\mathbf{x}\in\mathcal{X}_y} \mathbf{x})^T \mathbf{1}_V} \cdot N_{\text{avg.}}. \tag{10}$$

Then, we use y's label-topic distribution $\boldsymbol{\lambda}_{y,\cdot}/N_y$ to reconstruct $\mathbf{x}_{\text{pseu}}^y$. We propose a novel loss function:

$$\mathcal{L}_{\text{Sem.}} = -\mathbb{E}_y \left[(\mathbf{x}_{\text{pseu}}^y)^T \log \left(\boldsymbol{\beta}^T \frac{\boldsymbol{\lambda}_{y,\cdot}}{N_y} \right) \right], \tag{11}$$

to integrate the label-topic distribution with the label semantics defined by the pseudo-document.

We can regard $\mathcal{L}_{\text{Sem.}}$ as intuitive guidance for label-topic relationship learning. It provides simple knowledge about the label semantics, and makes the label-topic distribution meaningful.

4.4 Parameter Inference

In practice, we find that the computation of the OT problem is difficult to converge when optimizing $\mathcal{L}(\mathbf{x}, y)$ and $\mathcal{L}_{\text{Sem.}}$, and the decrease of Sinkhorn distance (Cuturi, 2013) will alleviate it significantly. To this end, we minimize the Sinkhorn distance using

$$\mathcal{L}_{\text{OT}} = \langle \mathbf{P}^*, \mathbf{C} \rangle . \tag{12}$$

Finally, to estimate the model parameters, we adopt the following overall loss function:

$$\begin{aligned} \mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},y)\in\mathcal{B}} \mathcal{L}(\mathbf{x}, y) \\ + \gamma_1 \mathcal{L}_{\text{OT}} + \gamma_2 \mathcal{L}_{\text{Sem.}}, \end{aligned} \tag{13}$$

where $\mathcal{B} \subset \mathcal{O}$ is the batch. γ_1 and γ_2 are the weights of $\mathcal{L}_{\text{OT}}$ and $\mathcal{L}_{\text{Sem.}}$, respectively. Algorithm 2 in Appendix C shows its training algorithm. The impact of $\mathcal{L}_{\text{OT}}$ and $\mathcal{L}_{\text{Sem.}}$ in Eq. (13) will be further studied in Section 5.4.

Note that our proposed model concentrates on inference rather than decoding. This implies that **our model can be combined with decoders of various existing NTMs**, such as ETM (Dieng et al., 2020) and ECRTM (Wu et al., 2023). For example, the decoder of ETM is parameterized by word and topic embeddings, and the topic-word distributions are computed by their inner products. We can attach ETM to LANTM by adopting the same way to compute the topic-word distributions. In our experiments, we construct our LANTM based on ETM and ECRTM, respectively. The definition of the cost matrix $\mathbf{C}$ in both cases can be found in Section 4.3.2.

4.5 Inference in the Absence of Labels

One should feed $\mathbf{x}$ and y simultaneously to LANTM to infer the document-topic distribution $\mathbf{z}$. But at the testing phase, y is absent. Considering that the ELBO of a document is maximized with its ground truth label when training, we follow Card et al. (2018) and consider all possible labels, choosing the one under which the ELBO is maximized, as the predicted label $\hat{y}$ for $\mathbf{x}$, i.e.,

$$\hat{y} = \arg\min_{y'\in\mathcal{Y}} \mathcal{L}(\mathbf{x}, y'). \tag{14}$$

When the ground truth labels are absent, we obtain the predicted labels using Eq. (14) as an estimate of the ground truth labels.

5 Experiment

5.1 Experimental Settings

5.1.1 Datasets

We conduct experiments on the following benchmark datasets: 20NewsGroup (**20NG**) (Lang, 1995), **IMDB** (Maas et al., 2011), and Yahoo! Answers (**Yahoo**) (Zhang et al., 2015). 20NG and IMDB are relatively long text datasets, with 20 and 2 labels, respectively. Yahoo is a short text dataset with 10 labels. To be consistent with the datasets used in Card et al. (2018) and Wu et al. (2023), we adopt the same pre-processing approach and set the vocabulary size to 5,000.

5.1.2 Baseline Methods

We compare our LANTM with the following baseline methods: 1) **sLDA**[4] (Mcauliffe and Blei, 2007), a classic supervised statistical topic model; 2) **SCHOLAR**[5] (Card et al., 2018), a classic supervised NTM; 3) **ETM**[6] (Dieng et al., 2020) and 4) **ECRTM**[7] (Wu et al., 2023), two state-of-the-art unsupervised NTMs. sLDA extends LDA by incorporating a response variable that corresponds to the supervised signal. It assumes that the response variable follows a Gaussian distribution, where the mean of this distribution is computed through a linear regression model based on the latent topics. Under a label-free prior, SCHOLAR learns label-relevant topics by using document-topic distributions to predict ground truth labels. ETM is a strong unsupervised NTM that computes the topic-word distributions based on the inner products between the topic and word embeddings. ECRTM is a state-of-the-art unsupervised NTM that avoids topic collapsing by embedding clustering regularization (ECR). It formalizes an OT problem between topic embeddings and word embeddings to enforce balanced topic-word assignment. As mentioned in Section 4.4, we construct our LANTM based on these two unsupervised baseline methods in the experiments, denoted by LANTM+ETM and LANTM+ECRTM, respectively. We do not consider Labeled LDA (Ramage et al., 2009) for comparison because it can be applied only when the number of topics is the same as the number of categories.

5.1.3 Settings for Our Method and Baselines

For baselines, we adopt the settings reported in the papers or implementations. We conduct 400 iterations for LANTM+ETM and LANTM+ECRTM. For all NTMs, we apply the Adam optimizer (Kingma and Ba, 2015) with a batch size of 200 and a learning rate of 0.002. For all embedding-based topic models, we set the dimension of the embedding space to 200. For λ_{ECR} in ECRTM and LANTM+ECRTM, we use the setting in Wu et al. (2023) and set λ_{ECR} to 250, 100, and 60 for 20NG, IMDB, and Yahoo, respectively. In LANTM+ETM, we let $\gamma_1 = 0.01$ and $\gamma_2 = 0.0001$. In LANTM+ECRTM, we let $\gamma_1 = 5$ and $\gamma_2 = 0.4$. In Section 5.4, we will discuss the impact of $\mathcal{L}_{\text{OT}}$ and $\mathcal{L}_{\text{Sem.}}$ using different settings for γ_1 and γ_2.

5.2 Label-topic Alignment

In this section, we examine the label-topic alignment achieved by our methods from the following aspects: 1) We visualize the label and topic embeddings and show the label-topic relationships for an intuitive view of label-topic alignment. 2) We conduct topic discovery on IMDB, where the human-defined labels do not describe the main patterns of the documents. 3) We perform clustering on the inferred document-topic distributions.

5.2.1 Visualization

We run LANTM+ECRTM on 20NG to perform visualization. Figure 3a shows the visualization of the label embeddings and topic embeddings using t-SNE (van der Maaten and Hinton, 2008), as well as top-5 words of the topics under some labels. Figure 3b is the heatmap of the label-topic indicator matrix.

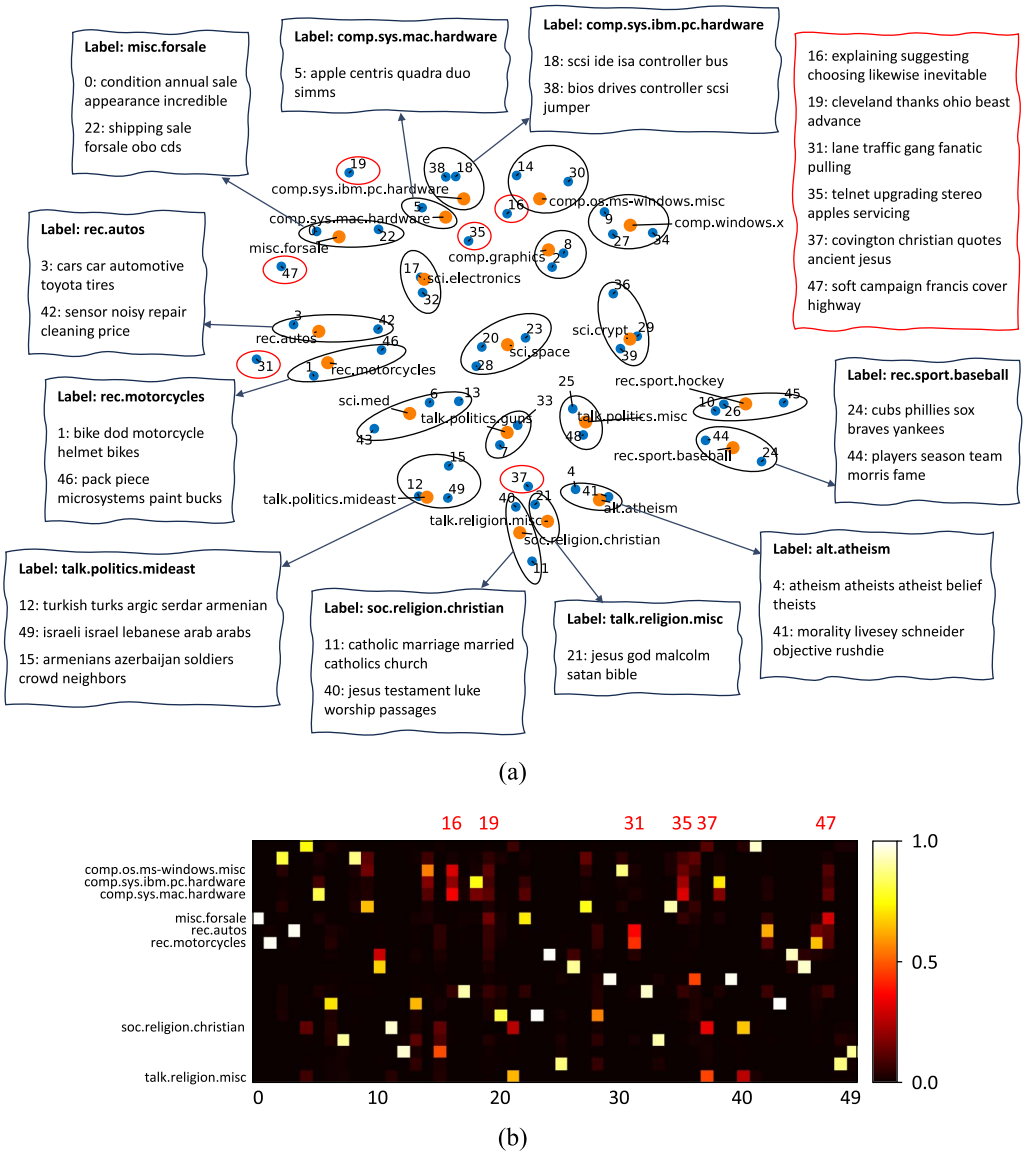
First, based on the top-5 topic words shown in Figure 3a, we can see that the label-topic indicators presented in Figure 3b reliably reflect the relationships between topics and labels. In other words, **we achieve the alignment between labels and topics**. For example, the high affiliation of topic

[4] https://github.com/chbrown/slda.

[5] https://github.com/dallascard/scholar.

[6] In order to alleviate topic collapsing (Srivastava and Sutton, 2017), we adopt an unnormalized topic-word matrix as the weight of decoder and add a batch normalization layer after decoding on the basis of the original implementation (https://github.com/adjidieng/ETM).

[7] https://github.com/BobXWu/ECRTM.



(a)

(b)

Figure 3: Visualization of the discovered topics and the label-topic relationships on 20NG by LANTM+ECRTM with $K = 50$. (a) t-SNE visualization of label embeddings (●) and topic embeddings (●). Topics under the same label are encircled together. Topics that are not explicitly assigned are circled in red, with their top-5 words presented in the red box in the upper right corner. (b) Heatmap of the label-topic indicator matrix. The indices of topics without explicit assignment are highlighted in red. The labels associated with these topics are marked on the left side.

#5 with the label `comp.sys.mac.hardware` distinguishes it from the topics under the similar label `comp.sys.ibm.pc.hardware`.[8] This brings the following benefits: 1) Using our proposed method, one can gain a more fine-grained understanding of the discovered topics from the perspective of document labels, which is beneficial for downstream tasks such as content analysis. 2) We can study the common pattern shared by different labels. One may regard a topic that is not explicitly assigned as the common pattern shared among relevant labels. For instance, topic #31 makes sense in the context of both `rec.autos` and `rec.motorcycles`. Instead, it is difficult for SCHOLAR to identify which label patterns are mixed within a topic.

Second, from the visualization of the label and topic embeddings, we can see that **modeling label-topic relationships on the embedding space leads to the refinement of the space**, where topics under the same label are close to each other while topics under different labels are far apart. In fact, we can regard the construction of embedding-based label-topic relationships

[8]Centris, Quadra, Duo are all series names by Apple.

Topic ID	Label	Polarity	Top-10 Topic Words
4	negative	0.489	waste horrible awful terrible crap worse sucked costs sucks crappy
12	positive	0.456	heartwarming tears april flawless sadness timeless compassion inspiring perfection unforgettable
26	positive	0.446	favorites fabulous bates nominated wonderful oscars flawless nomination wrenching delight
15	negative	0.367	waste flop travesty advise unlikeable choppy props checked offended awful
5	negative	0.329	hopelessly franco trashy hardcore walken boom preposterous tiresome overacting tame
37	negative	0.302	dialogue dialogues waste preview substance edgy distracting lacked poorly unlikeable
47	–	0.000	eddie murphy metal rock mario band school songs hello roll
17	–	0.000	games game saturday nicole russia friday minutes match bears night
8	–	0.006	martial chan jackie kung arts kong hong stunts jet chinese
31	–	0.012	films trailer waited spoof film powell park edward walked patience

Table 1: Some most explicitly assigned and least explicitly assigned topics discovered by LANTM+ECRTM on IMDB where two sentiment labels, negative and positive, are defined. The value of polarity in this table is calculated by $\max\{\lambda_{\text{positive},k}, \lambda_{\text{negative},k}\} - 0.5$ for the topic with ID k.

mentioned in Section 4.3 as clustering, where label embeddings serve as the centroids of topic embeddings.

Ultimately, we can observe from Figure 3b that **when modeling them as the transport plan, the label-topic relationships are well-organized**. Most of the topics are clearly assigned, and the number of topics is relatively balanced across different labels.

5.2.2 Alignment between Topics and Labels Describing Secondary Aspects

As we know, the discovered topics in topic modeling capture the co-occurrence patterns among words in the corpus. However, the human-defined labels of the documents are sometimes not aligned with such patterns. In this situation, to better understand the corpus, one would want to know which topics describe the co-occurrence patterns and which topics are label-relevant. Take the sentiment labels in IMDB dataset (positive and negative), for example. The sentiments of the audiences and descriptions of the film's content are two important patterns of the reviews in IMDB. Obviously, it is unreasonable to assign any sentiment labels to the topics that fully describe the content of a film. Thanks to the label-topic indicators, LANTM is capable of discovering label-relevant and label-irrelevant topics simultaneously. Table 1 shows some most explicitly assigned and least explicitly assigned topics by LANTM+ECRTM on IMDB. We calculate the difference between the most significant indicator's value and 0.5 as the value of polarity, namely, the degree to which a topic is assigned to a sentiment label. The most explicitly assigned topic #4 consists of negative words, while the least explicitly assigned topic #47 does not exhibit any inclination. **The results show the meaningfulness of the label-topic indicators we propose in the scenario where the labels describe secondary aspects of the corpus**.

5.2.3 Clustering-based Evaluation

It is challenging to design a quantitative metric to measure the degree to which the topics are aligned with the labels. A common practice (Hoyle et al., 2022; Zhao et al., 2021; Wu et al., 2023) is to cluster the document-topic distributions to determine the categories induced by the topic model and assess the extent to which these categories align with the ground truth labels using clustering metrics. We follow previous works and calculate the following three cluster quality metrics: Adjusted Rand Index (ARI), Purity, and Normalized Mutual Information (NMI). We use two clustering strategies on the document-topic distributions of testing documents: 1) apply **k-means** algorithm, and 2) use the most dominant topic as the assigned cluster (referred to as **top** strategy). Similar to Wang et al. (2022), we set the number of clusters to 20 in k-means algorithm. The results are presented in Table 2. Unless otherwise mentioned, we do not cite the results reported in ECRTM (Wu et al., 2023) but re-run all the baselines, so as to obtain ARI scores.

From the experimental results, we can draw the following conclusions: 1) **Our proposed methods can effectively enhance the label-topic alignment compared to unsupervised NTMs**. Due to the lack of label guidance, unsupervised NTMs cannot autonomously achieve alignment between labels and topics. In contrast, our proposed methods conduct inference with label-topic indicators and thus enhance the label-topic alignment. 2)

Dataset	Model	$K = 50$						$K = 100$					
		kmeans			top			kmeans			top		
		ARI	Purity	NMI	ARI	Purity	NMI	ARI	Purity	NMI	ARI	Purity	NMI
20NG	sLDA	†0.121	†0.457	†0.484	†0.324	†0.581	†0.490	–	–	–	–	–	–
	SCHOLAR	†0.040	†0.357	†0.391	†0.316	†0.629	†0.541	†0.011	†0.231	†0.276	†0.223	†0.637	†0.522
	ETM	*0.232	*0.492	*0.501	*0.265	*0.529	*0.429	*0.142	*0.420	*0.479	*0.186	*0.558	*0.430
	LANTM+ETM	0.333	0.584	0.560	0.312	0.624	0.507	**0.426**	**0.631**	**0.607**	0.223	0.684	0.521
	ECRTM	*0.101	*0.423	*0.466	*0.389	*0.574	*0.521	*0.049	*0.317	*0.377	*0.322	*0.559	*0.487
	LANTM+ECRTM	**0.370**	**0.595**	**0.578**	**0.415**	**0.635**	**0.573**	0.338	0.623	**0.607**	**0.444**	**0.689**	**0.599**
IMDB	sLDA	†0.033	†0.709	†0.082	†0.009	†0.676	†0.046	†0.026	†0.659	†0.063	†0.004	†0.658	†0.035
	SCHOLAR	†0.050	†0.675	†0.105	†0.017	**0.757**	0.086	†0.069	†0.666	†0.107	†0.010	**0.766**	**0.081**
	ETM	*0.020	*0.681	*0.057	*0.007	*0.656	*0.039	*0.039	*0.692	*0.076	*0.004	*0.664	*0.035
	LANTM+ETM	0.033	0.729	0.092	0.012	0.685	0.054	0.078	0.754	0.112	0.006	0.689	0.111
	ECRTM	*0.032	*0.667	*0.073	*0.013	*0.686	*0.055	*0.025	*0.656	*0.058	*0.005	*0.695	*0.046
	LANTM+ECRTM	**0.055**	**0.782**	**0.123**	**0.030**	0.742	**0.088**	**0.199**	**0.811**	**0.170**	**0.016**	0.762	0.080
Yahoo	sLDA	†0.084	†0.443	†0.312	†0.134	†0.517	†0.273	–	–	–	–	–	–
	SCHOLAR	†0.013	†0.287	†0.230	†0.113	†0.560	†0.317	†0.005	†0.199	†0.144	†0.070	†0.575	†0.313
	ETM	*0.117	*0.482	*0.310	*0.122	*0.502	*0.264	*0.068	*0.420	*0.303	*0.064	*0.496	*0.255
	LANTM+ETM	0.161	0.530	0.336	0.128	0.555	0.291	**0.222**	**0.549**	0.348	0.084	0.576	0.307
	ECRTM	*0.083	*0.455	*0.325	*0.130	*0.549	*0.299	*0.033	*0.319	*0.248	*0.103	*0.561	*0.304
	LANTM+ECRTM	**0.178**	**0.560**	**0.358**	**0.166**	**0.600**	**0.338**	0.177	0.541	**0.355**	**0.126**	**0.613**	**0.339**

Table 2: ARI, Purity, and NMI scores under 50 and 100 topics. We run all models 5 times and report the average results. † indicates the gain of LANTM+ECRTM compared to supervised baseline methods is statistically significant at 0.05 level, while $*$ indicates the gain after applying LANTM is statistically significant at 0.05 level. The best scores are in **bold**. We do not report the results of sLDA on 20NG and Yahoo when $K = 100$ because it failed to converge in 48 hours.

Our proposed methods outperform other supervised topic models on 20NG and Yahoo, and are on par with SCHOLAR on IMDB. We believe that label-irrelevant patterns dominate the IMDB dataset, for which our methods do not show a significant improvement in terms of clustering metrics on IMDB. This interferes with label prediction by Eq. (14), limiting the performance of LANTM in clustering tasks.

Although there is marginal improvement compared to SCHOLAR in terms of clustering metrics on the IMDB dataset, we argue that **one of the advantages of our proposed methods lies in its ability to accurately identify the relationships between labels and topics by label-topic indicators**, especially when label-irrelevant patterns dominate the corpus. Accurately reflecting the above ability that SCHOLAR lacks, via clustering metrics, presents a significant difficulty.

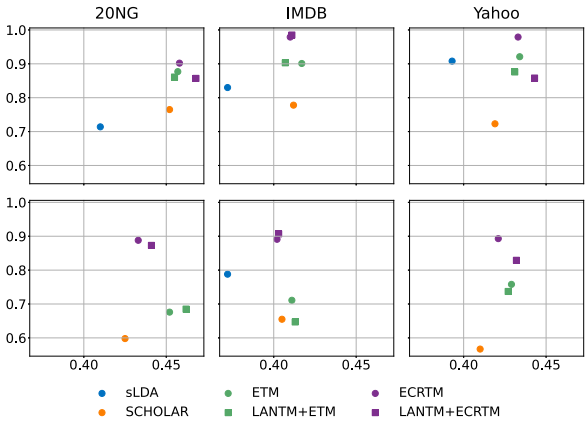


Figure 4: Topic quality of different models. The x-axis and y-axis of each subplot represent the value of TC and TD, respectively. The upper row of subplots shows the case for $K = 50$, and the lower one for $K = 100$. We run all models 5 times and report the average results. We do not report the results of sLDA on 20NG and Yahoo when $K = 100$ because it failed to converge in 48 hours.

5.3 Topic Quality

To evaluate the topic quality, we consider both Topic Coherence (TC) and Topic Diversity (TD). TC measures the coherence of the top words within a topic. We consider the top 10 words for each topic and calculate C_V (Röder et al., 2015) with Wikipedia article collection as the reference corpus. Following Dieng et al. (2020), we calculate the proportion of unique words among the top 25 words for each topic as TD. Figure 4 showcases the topic quality of the evaluated models.

The results indicate that **the topic quality is maintained after applying our proposed model**. There exists a coherence-diversity trade-off after LANTM is applied to ETM or ECRTM. It may be due to the regularization by $\mathcal{L}_{\text{Sem.}}$ on the topic embeddings. In general, LANTM+ECRTM is more stable than LANTM+ETM. Also, owing to the powerful ECR (Wu et al., 2023), **LANTM+ECRTM shows much better topic quality than sLDA and SCHOLAR**, which makes LANTM+ECRTM a better choice from the perspective of topic quality.

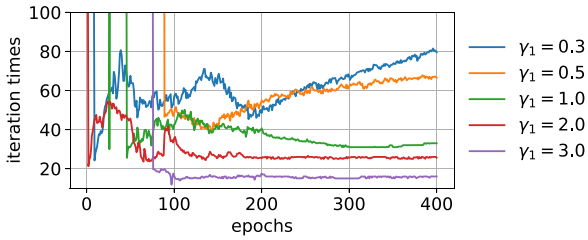


Figure 5: The iteration times of Sinkhorn's Algorithm for computing $\boldsymbol{\lambda}$ in each epoch with different γ_1 when training LANTM+ECRTM on 20NG, $K = 50$, and $\gamma_2 = 0.4$. The cases of $\gamma_1 = 0$ and $\gamma_1 = 0.1$ are not reported in the figure, as the computation has not converged within 1,000 iterations for these cases.

γ_1	**1.0**	**3.0**	**5.0**	**7.0**	**9.0**
TC	0.468	0.465	0.467	0.459	0.458
TD	0.847	0.870	0.861	0.863	0.874
k-ARI	0.366	0.361	0.373	0.340	0.342
k-Purity	0.604	0.593	0.595	0.586	0.577
k-NMI	0.575	0.569	0.578	0.571	0.559
t-ARI	0.433	0.421	0.415	0.422	0.400
t-Purity	0.631	0.627	0.635	0.638	0.592
t-NMI	0.578	0.573	0.573	0.575	0.547

Table 3: Quantitative results of LANTM+ECRTM on 20NG across different γ_1 when $K = 50$ and $\gamma_2 = 0.4$. ''k-'' and ''t-'' stand for two clustering strategies: k-means and top, respectively.

5.4 Ablation Study

In this section, we will discuss the impact of $\mathcal{L}_{\text{OT}}$ and $\mathcal{L}_{\text{Sem.}}$ by controlling γ_1 and γ_2, respectively.

As mentioned in Section 4.4, we minimize the Sinkhorn distance using $\mathcal{L}_{\text{OT}}$ for faster convergence. To further study how $\mathcal{L}_{\text{OT}}$ affects the performance of our proposed method, we fix $\gamma_2 = 0.4$ and run LANTM+ECRTM on 20NG using different γ_1. The results are presented in Figure 5 and Table 3. We can draw the following conclusions: 1) **$\mathcal{L}_{\text{OT}}$ is able to accelerate convergence when computing $\boldsymbol{\lambda}$**. As shown in Figure 5, when $\gamma_1 = 0$, Sinkhorn's Algorithm cannot even converge in 1,000 iterations, which greatly reduces computational efficiency. Neglecting the oscillation before 100 epochs, the Sinkhorn iteration times decrease as γ_1 increases. When $\gamma_1 = 3.0$, it only requires about 15 iterations, making the computation quite efficient. 2) **$\mathcal{L}_{\text{OT}}$ has a relatively small impact on topic quality and clustering performance**. This is because $\mathcal{L}_{\text{OT}}$ is neither directly related to document-topic distributions nor to topic-word distributions. We can see from Table 3 that when $\gamma_1 < 5.0$, the reported metrics vary slightly. When $\gamma_1 = 9.0$, all metrics except TD and k-ARI show a decrease. For LANTM+ECRTM, $\gamma_1 = 5.0$ is generally enough for efficient computation while avoiding a decrease in topic quality and clustering metrics.

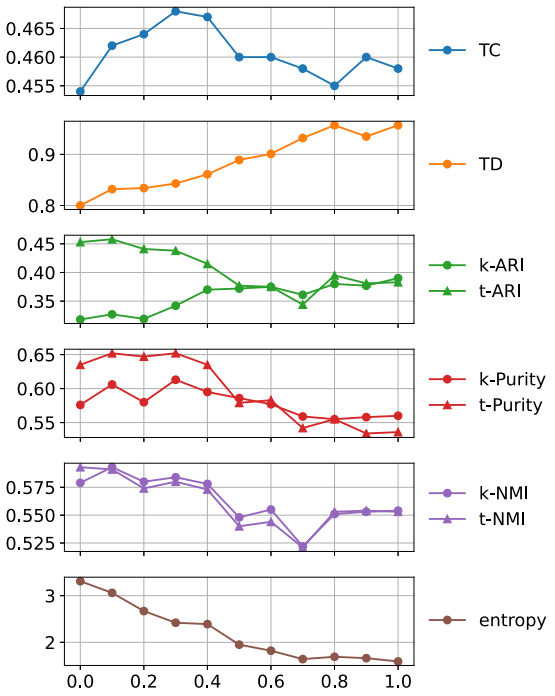
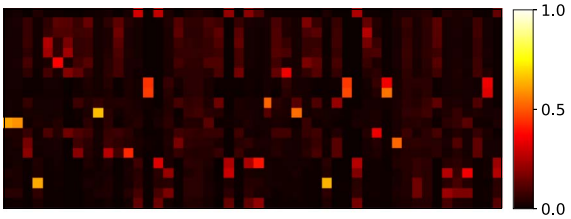
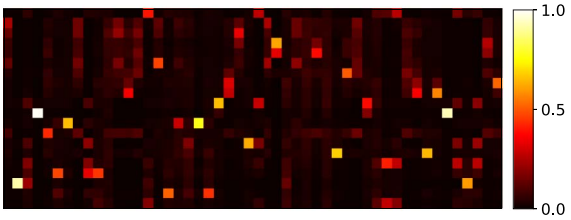


Figure 6: Quantitative results of LANTM+ECRTM on 20NG across different γ_2 (x-axis) when $K = 50$ and $\gamma_1 = 5$. ''k-'' and ''t-'' stand for two clustering strategies: k-means and top, respectively. ''Entropy'' in the last subfigure is the average entropy of $\boldsymbol{\lambda}_{\cdot,k}$, i.e., $\frac{1}{K}\sum_{k=1}^{K}\left(-\sum_{y=1}^{C}\lambda_{y,k}\log\lambda_{y,k}\right)$.

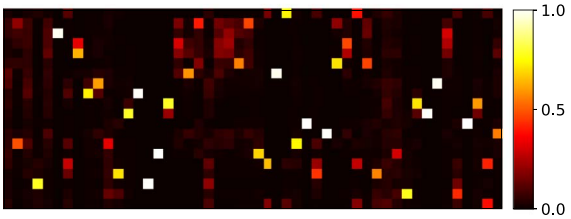
Regarding the effect of $\mathcal{L}_{\text{Sem.}}$, we control the value of γ_2 and run LANTM+ECRTM on 20NG by fixing $\gamma_1 = 5$. Figure 6 shows different evaluating metrics and entropy of $\boldsymbol{\lambda}_{\cdot,k}$, which indicate the sparsity of the label-topic relationships, with γ_2 varying from 0 to 1.0. To have a more intuitive understanding about entropy of $\boldsymbol{\lambda}_{\cdot,k}$, Figure 7 shows the heatmaps of the label-topic indicator matrix $\boldsymbol{\lambda}$ in four cases where γ_2 is set to 0, 0.1, 0.4, and 0.8, respectively. We have the following observations: 1) $\mathcal{L}_{\text{Sem.}}$ **has a huge impact on the learned structural characteristics of the label-topic relationships**. When $\mathcal{L}_{\text{Sem.}}$ is absent ($\gamma_2 = 0$), the label-topic relationships are fuzzy overall, as shown in Figure 7a. Lacking guidance of the label semantics, the topics belong to many similar labels and lose specificity



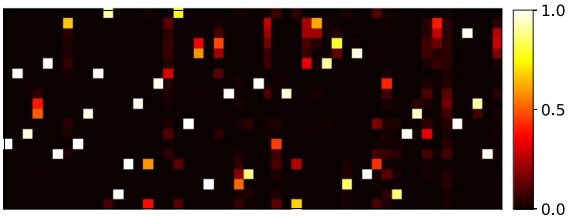
(a) $\gamma_2 = 0.0$, *entropy* $= 3.30$



(b) $\gamma_2 = 0.1$, *entropy* $= 3.01$



(c) $\gamma_2 = 0.4$, *entropy* $= 2.33$



(d) $\gamma_2 = 0.8$, *entropy* $= 1.64$

Figure 7: Heatmaps of the label-topic indicator matrix $\boldsymbol{\lambda}$ learned by LANTM+ECRTM on 20NG with different γ_2 when $K = 50$. *entropy* is the average entropy of $\boldsymbol{\lambda}_{\cdot,k}$, i.e., *entropy* $= \frac{1}{K}\sum_{k=1}^{K}\left(-\sum_{y=1}^{C}\lambda_{y,k}\log\lambda_{y,k}\right)$.

to labels. This is because under the objective of maximizing ELBO, the topics only capture patterns of word co-occurrence, which may be shared among similar labels. Meanwhile, $\mathcal{L}_{\text{Sem.}}$ is strong enough to amplify differences between labels. As γ_2 increases, the entropy of $\boldsymbol{\lambda}_{\cdot,k}$ decreases, which demonstrates well label-topic alignment. 2) **The entropy and clustering performance are not always consistent**, as shown in Figure 6. We believe that this is caused by the spurious semantics that the pseudo-documents bring. The pseudo-documents are defined to describe the overall semantics of documents with the same label, and they inevitably include some spurious word co-occurrence patterns. We believe that the issue of spurious semantics affects the accuracy of label estimation by Eq. (14) and further brings worse clustering performance. Eq. (14) depends on the maximal ELBO to predict labels. But unfortunately, spurious semantics will interfere with ducument reconstruction during training, finally cause a decrease in clustering metrics. Moreover, such spurious semantics also cause reduced coherence, as shown in Figure 6. From the discussions above, we can find that the choice of γ_2 faces a tradeoff between topic quality, clustering performance, and the learned structural characteristics.

6 Conclusion

In this paper, we propose a novel supervised NTM called LANTM, which overcomes the limitation in existing supervised NTMs brought by a label-free prior. Our LANTM utilizes a chain-structured graphical model with a label-conditioned prior. The inference is conducted under soft indicators that describe the relationships between labels and topics. To obtain well-organized label-topic relationships, we formalize an entropy-regularized OT problem and model the label-topic relationships as the transport plan. We propose to reconstruct label-specific pseudo-documents to enable the label-topic distributions to capture the label semantics. Our LANTM can be seamlessly integrated with most existing unsupervised NTMs. Experimental results show that our model can effectively enhance the label-topic alignment while maintaining topic quality. In future work, we will expand our method to model more fine-grained dependencies such as label's dependencies.

Acknowledgments

We express our profound gratitude to the action editor and reviewers for their valuable comments and suggestions. This work has been supported by the National Natural Science Foundation of China (62372483) and NSF grant CNS-2349369.

References

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*,

volume 14. MIT Press. https://doi.org/10.7551/mitpress/1120.003.0082

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). https://doi.org/10.1609/aaai.v29i1.9499

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1189

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453. https://doi.org/10.1162/tacl_a_00325

Philipp Hennig, David Stern, Ralf Herbrich, and Thore Graepel. 2012. Kernel topic models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 511–519, La Palma, Canary Islands. PMLR.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? The incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.390

Minghui Huang, Yanghui Rao, Yuwei Liu, Haoran Xie, and Fu Lee Wang. 2018. Siamese network-based supervised topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4652–4662, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1494

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings*.

Simon Lacoste-Julien, Fei Sha, and Michael Jordan. 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339. https://doi.org/10.1016/B978-1-55860-377-6.50048-7

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1056

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Xian-Ling Mao, Zhao-Yan Ming, Tat-Seng Chua, Si Li, Hongfei Yan, and Xiaoming Li. 2012. SSHLDA: A semi-supervised hierarchical topic

model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 800–809, Jeju Island, Korea. Association for Computational Linguistics.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419. PMLR.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA. PMLR.

Adler Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. 2011. Hierarchically supervised latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Yves Petinot, Kathleen McKeown, and Kapil Thadani. 2011. A hierarchical model of web summaries. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 670–675, Portland, Oregon, USA. Association for Computational Linguistics.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore. Association for Computational Linguistics. https://doi.org/10.3115/1699510.1699543

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2684822.2685324

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Krishna B. Vamshi, Ajeet Kumar Pandey, and Kumar A. P. Siva. 2018. Topic model based opinion mining and sentiment analysis. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4. https://doi.org/10.1109/ICCCI.2018.8441220

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022. Representing mixtures of word embeddings with mixtures of topic embeddings. In *International Conference on Learning Representations*.

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023. Effective neural topic modeling with embedding clustering regularization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37335–37357. PMLR.

Xiaobao Wu, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. 2024. On the affinity, rationality, and diversity of hierarchical topic modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19261–19269. https://doi.org/10.1609/aaai.v38i17.29895

Feng Xue, Richang Hong, Xiangnan He, Jianwei Wang, Shengsheng Qian, and Changsheng

Xu. 2020. Knowledge-based topic model for multi-modal social event analysis. *IEEE Transactions on Multimedia*, 22(8):2098–2110. https://doi.org/10.1109/TMM.2019.2951194

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2021. Neural topic model via optimal transport. In *International Conference on Learning Representations*.

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2009. MedLDA: Maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, page 1257–1264, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1553374.1553535

A Derivation of the ELBO

$$\log p(\mathbf{x}, y) = \log \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)} \frac{p(\mathbf{x}, y, \mathbf{z})}{q(\mathbf{z}|\mathbf{x}, y)} \right] \tag{15}$$

$$\geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)} \left[\log p(\mathbf{x}, y, \mathbf{z})\right] - \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)} \left[\log q(\mathbf{z}|\mathbf{x}, y)\right] \tag{16}$$

$$= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)} \left[\log p(\mathbf{x}|\mathbf{z})\right] + \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)} \left[\log p(\mathbf{z}|y)\right] + \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)} \left[\log p(y)\right] - \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)} \left[\log q(\mathbf{z}|\mathbf{x}, y)\right] \tag{17}$$

$$= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)} \left[\log p(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}} \left[q(\mathbf{z}|\mathbf{x}, y) \parallel p(\mathbf{z}|y)\right] + \log p(y) \tag{18}$$

B Sinkhorn's Algorithm

Algorithm 1 Sinkhorn's Algorithm

Input: probability vectors $\mathbf{s}$ and $\mathbf{t}$, the cost matrix $\mathbf{C}$, weight of the entropic regularization term $1/\epsilon$, tolerance ε;

Output: optimal transport plan $\mathbf{P}^*$;

$\mathbf{u} \leftarrow \mathbf{1}_C/C$, $\mathbf{v} \leftarrow \mathbf{1}_K/K$, $\mathbf{H} \leftarrow \exp(-\mathbf{C}/\epsilon)$;
while *err* $> \varepsilon$ and *not reach max iteration times* **do** ▷ Sinkhorn iteration
  $\mathbf{v} \leftarrow \mathbf{t}./(\mathbf{H}^T\mathbf{u})$;
  $\mathbf{u} \leftarrow \mathbf{s}./(\mathbf{H}\mathbf{v})$;
  $\tilde{\mathbf{t}} \leftarrow \mathbf{v}.*(\mathbf{H}^T\mathbf{u})$;
  err $\leftarrow \mathbf{1}_K^T \left|\mathbf{t} - \tilde{\mathbf{t}}\right|$;
end while
$\mathbf{P}^* \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{H}\mathrm{diag}(\mathbf{v})$;

C Training Algorithm

Algorithm 2 Training Algorithm for LANTM

Input: observations $\mathcal{O}$, number of epochs N_{iter};

Initialize model parameters $\Theta_{\boldsymbol{\mu}}$, $\Theta_{\boldsymbol{\Sigma}}$, $\mathbf{W}$, $\mathbf{T}$, $\mathbf{L}$;
Decide a decoder $\mathrm{Dec}(\cdot)$;
for i from 1 to N_{iter} **do**
  Compute $\mathbf{P}^*$ by Eq. (7);
  $\boldsymbol{\lambda} \leftarrow K\mathbf{P}^*$;
  for batch $\mathcal{B}$ from $\mathcal{O}$ **do**
    for $(\mathbf{x}, y)$ in $\mathcal{B}$ **do**
      $\boldsymbol{\mu} \leftarrow f_{\boldsymbol{\mu}}(\mathbf{x})$, $\boldsymbol{\Sigma} \leftarrow f_{\boldsymbol{\Sigma}}(\mathbf{x})$;
      Draw $\mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with reparameterization trick;
      $\mathbf{z} \leftarrow \mathrm{Softmax}(\mathbf{r} + \ln \boldsymbol{\lambda}_{y,\cdot})$;
      Compute $\mathcal{L}(\mathbf{x}, y)$ by Eq. (3);
    end for
    Compute $\mathcal{L}$ by Eq. (13);
    Update $\Theta_{\boldsymbol{\mu}}$, $\Theta_{\boldsymbol{\Sigma}}$, $\mathbf{W}$, $\mathbf{T}$, $\mathbf{L}$;
  end for
end for