

Neural Topic Model with Distance Awareness

Shangyu Chen (✉ shangyu.chen@monash.edu)

Monash University

He Zhao

Monash University

Viet Huynh

Monash University

Jianfei Cai

Monash University

Dinh Phung

Monash University

Research Article

Keywords: Neural topic model, Nature language Processing, Optimal Transport, Machine learning

Posted Date: February 22nd, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-2599013/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Neural Topic Model with Distance Awareness

Shangyu Chen^{1*}, He Zhao¹, Viet Huynh¹, Jianfei Cai¹ and Dinh Phung¹

^{1*}Monash University, FIT-DSAI, Exhibition Walk, Clayton, 3168, VIC, Australia.

*Corresponding author(s). E-mail(s): shangyu.chen@monash.edu;
 Contributing authors: ethan.zhao@monash.edu;
viet.huynh@monash.edu; jianfei.cai@monash.edu;
dinh.phung@monash.edu;

Abstract

Neural topic models (NTM) have shown their success in topic modelling with a wide range of applications in text analysis. NTMs based on generative models tend to have good refactoring capabilities in achieving good document representations, rather than providing a topic representation that has an understandable association with documents and words. To bridge this gap, we have proposed a neural topic model that allows simultaneously learning topics from text data as well as capturing the structural manifold in the original data space of documents. Comprehensive experiments have shown that the proposed model can improve learned topics in terms of intrinsic metrics (purity and coherence) and extrinsic downstream tasks (clustering, classification). Beyond that, the model can provide interpretable low dimensional presentations of documents.

Keywords: Neural topic model, Nature language Processing, Optimal Transport, Machine learning

1 Introduction

Leveraging unsupervised techniques to extract document topics without any categories or labels is a natural idea, and topic models have gained widespread adoption in automatic text analysis [1]. The topic is treated as the data representation of the document, in the way of Variational Autoencoder (VAE) [2]. The Neural Topic Model (NTM) is an unsupervised learning algorithm for organizing a corpus of documents

2 *Neural Topic Model with Distance Awareness*

into topics containing word groupings. NTM is used as the bag of words of the document to the encoder of the latent topics. The NTM can better overcome the problem of short text sparsity, and it also performs well in terms of consistency and diversity of topics.

The main goal of NTM is to achieve low reconstruction error, but it cares less about the interpretability of topic representations, specifically, whether the mapping between topics and text is clearly organized or the relationships that exist between texts are preserved in the subject space. Recently, an NTM based on the optimal transport (OT) framework proposed in [3] seeks to strike a better balance between obtaining good document representation and generating coherent/diverse topics. The model utilizes an encoder that outputs the topic distribution of the document by taking its word count vector as input like a standard NTM, but minimizes the OT distance between word count vector and topic distributions, which are two discrete distributions of the support for words and topics. Because the model focuses on the process of the topic distribution being as close as possible to the word distribution of a document, the geometry of the word distributions between documents is not preserved.

For topic models, such geometry is particularly important, i.e., whether the two documents that are similar on the word space are still similar in the topic space. In [3], even OT tries to find the shortest transfer path, adding OT to NTM can make it not transfer the associated elements too far, but it does not have an intuitive effect. The main idea of manifold learning [4], [5] is to map high-dimensional data to low-dimensional data, so that the low-dimensional data can reflect some essential structural features of the original high-dimensional data. The premise of manifold learning is an assumption that some high-dimensional data is actually a low-dimensional manifold structure embedded in a high-dimensional space. A “manifold” refers to a region that is connected together, and mathematically, it refers to a set of points, each of which has its neighbors. Given any point, its manifold locally looks like Euclidean space. In other words, it has the properties of Euclidean space in the local space, and can use Euclidean space for distance calculation. Therefore, it is easy to establish a dimension reduction mapping relationship locally, and then try to generalize the local relationship to the global, and then display it visually.

To model the geometry between documents, we propose a neural topic model in this paper, which is built upon a novel manifold learning framework derived from a new view of optimal transport topic modelling. Specifically, we consider two representation encodings for documents: Intuitive document representations x (can be obtained by word count vector), and topic distribution z . In the document space, the vocabulary size is very large, even if a document contains only a small part of the vocabulary, x still needs to contain information about each word in the vocabulary, so x is high-dimensional and sparse. The topic distribution z is obtained from the encoder, and z can be defined as a low-dimensional vector. In NSTM [3], the learning process of the topic model is the process of the distribution of z approaching the distribution of x . In this paper, while retaining the document reconstruction ability of VAE and the distribution approximation ability of OT, we will emphasize the structure preservation ability of topic distribution to document collections. Specifically, two

similar documents should also behave closely in topic space, which is achieved by manifold learning.

We summarize *our contributions* as follows: (i) We proposed a novel model **Distance Awareness NTM (DA-NTM)** which jointly performs deep topic modelling and manifold learning from documents; (ii) The proposed model has shown its benefits in both aspects of topic modelling and visualization via the comprehensive experiments; (iii) We demonstrated that there is a balance between the reconstruction and the independence between topics in DA-NTM.

2 Background

2.1 Neural Topic Models & Optimal Transport

Most of the existing NTMs, such as NVDM [6], are neural topic models based on standard variational autoencoders (VAEs). The target of VAE [2] is to model the true posterior distribution $p(z|x)$ of the latent variable, by solving the variational posterior distribution $q_\theta(z|x)$ (θ is a variational parameter), and continuously reduce the difference between $q_\theta(z|x)$ and $p(z|x)$ for approximation purpose.

A topic model adopts the architecture of the standard VAE described above, and each document consists of a word count vector (i.e. $x \in \mathbb{N}^V$) and a latent distribution (i.e. $z \in \mathbb{R}^K$) of K topics. Moreover, the model takes the document's bag of words representation BOW (Bag Of Word) as input. In order to maintain the data scale consistency, we sampled from the latent space and then use them as topic distributions. An NTM assumes that the topic z of the document is determined by the prior distribution $p(z)$ and that x can be generated by the conditional distribution $p_\phi(x|z)$, which is modeled by the decoder ϕ . The output goal of this model is to infer the topic given a bag of words, i.e. to compute $p(z|x)$, so we need a neural network based the encoder θ to get the distribution $q_\theta(z|x)$ for approximation $p(z|x)$. So we have an optimization objective similar to VAE.

$$\max_{\theta, \phi} (E_{q_\theta(z|x)} [\log p_\phi(x|z)] - \mathbb{KL}[q_\theta(z|x)p(z)]) \quad (1)$$

The first term is the expectation of the log-likelihood, which can be understood as the reconstruction error of the document bag of words, and the second term is the Kullback-Leibler (KL) Divergence fitting the prior $p(z)$ with $q_\theta(z|x)$. $\phi(z)$ is usually constructed by a single-layer network and $p(z)$ is usually a Gaussian distribution [7].

Here, we will discuss optimal transport(OT) restricted to discrete distributions, although it also applies to continuous distributions. We define the OT distance of two probability distributions $\mathbf{r}$ and $\mathbf{c}$ as

$$d_M(\mathbf{r}, \mathbf{c}) := \min_{\mathbf{P} \in U(\mathbf{r}, \mathbf{c})} \langle \mathbf{P}, \mathbf{M} \rangle \quad (2)$$

where $\mathbf{r} \in \Delta^{D_r}$ and $\mathbf{c} \in \Delta^{D_c}$, and Δ^D represents a $D - 1$ simplex. $\mathbf{P} \in \mathbb{R}_{>0}^{D_r \times D_c}$ is the transport plan, $U(\mathbf{r}, \mathbf{c})$ is the transport polytope of $\mathbf{r}$ and $\mathbf{c}$ which means the collection of all possible transport plan. $\mathbf{M} \in \mathbb{R}_{\geq 0}^{D_r \times D_c}$ is the cost matrix. How to define an appropriate $\mathbf{M}$ is a common challenge in specific problems.

4 Neural Topic Model with Distance Awareness

To efficiently compute OT distances, [8] introduced a regularized optimal transport distance with entropy constraints,

$$L_C^\varepsilon(a, b) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in U(\mathbf{r}, \mathbf{c})} < \mathbf{P}, \mathbf{C} > + \varepsilon \sum_{i,j} \mathbf{P}_{i,j} (\log(\mathbf{P}_{i,j} - 1)) \quad (3)$$

where $U(\mathbf{r}, \mathbf{c}) := \{\mathbf{P} \in U(\mathbf{r}, \mathbf{c}) \mid h(\mathbf{P}) \geq h(\mathbf{r}) + h(\mathbf{c}) - \alpha\}$, $h(\cdot)$ is the entropy function, and $\alpha \in [0, \infty)$. $\mathbf{C} \in \mathbb{R}^{D_r \times D_c}$ is the underlying distance matrix, whose element measures the distance between different states. ε controls the significance of the entropy regularizer.

For NTM via OT (NSTM) [3], an encoder is leveraged to generate topic z from normalized word vector $\tilde{x}$ by $z = \text{softmax}(\theta(\tilde{x}))$. Since $\tilde{x}$ and z are two distributions with different support for the same document, in order to learn the encoder, the OT distance is minimized to push z towards $\tilde{x}$, as $\min_{\theta} d_{\mathbf{M}}(\tilde{x}, z)$. Here the cost matrix $\mathbf{M}$ is specified as the following construction: $\mathbf{M}_{vk} = 1 - \cos(e_v, \mathbf{g}_k)$, where $\cos$ is the cosine similarity; $\mathbf{g}_k \in \mathbb{R}^L$ and $e_v \in \mathbb{R}^L$ are the embeddings for topic k and word v , respectively. L is the dimension of word embedding and topic embedding. The word embedding e_v is obtained by pre-training. $G \in \mathbb{R}^{L \times K}$ is used as a set of topic embedding (K is the number of documents in a batch, G is random initialised), and also as an optimization parameter, that is, the new optimization objective of OT is,

$$\min_{\theta, \mathbf{G}} d_{\mathbf{M}}(\tilde{x}, z) \quad (4)$$

Combining the OT loss with the traditional cross-entropy loss yields better performance when using either of them. According to the previously mentioned in NTM, encoder θ is simulated by a neural network and $\phi(z)$ is constructed by a single-layer network for decoder. Combining the OT distance with the expected log-likelihood, with Sinkhorn, the optimization objective becomes

$$\max_{\theta, \mathbf{G}} (\varepsilon \tilde{x}^T \log \phi(z) - L_{\mathbf{M}, \alpha}(\tilde{x}, z)) \quad (5)$$

where α is the hyperparameter for the Sinkhorn distance.[8]

2.2 Manifold Learning

In manifold learning, the observed data is actually mapped from a low-dimensional manifold to a high-dimensional space. Due to the limitations of the internal characteristics of the data, some high-dimensional data can create dimensional redundancy. If the mapping can reduce the data from the high-dimensional space to the low-dimensional space without loss of information, more essential features of the data will be output by feeding the original data.

Therefore, most manifold learning algorithms [9–23] have a similar general idea, that is: assuming that the data has a certain structural feature in high dimensions, it is expected that the structure can still be maintained after it is reduced to low dimensions.

Compared to other manifold learning algorithms, tSNE [24] measures the distance between sample data as a conditional probability. For tSNE, , suppose there are x_i and

x_j in the high-dimensional space, $p_{j|i}$ indicates that the probability that x_j is in the neighbor of x_i . $p_{j|i}$ using Gaussian distribution, the formula is as follows

$$p_{j|i} = \frac{\exp(-|x_i - x_j|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-|x_i - x_k|^2/2\sigma_i^2)} \quad (6)$$

The variance σ_i of the Gaussian distribution corresponding to the different points x_i in high-dimensional space needs to be calculated separately.

Similarly, for points in high-dimensional space mapped to the corresponding point in the low-dimensional space (i.e. y_i and y_j), its probability distribution function is defined as

$$q_{j|i} = \frac{\exp(-|y_i - y_j|^2)}{\sum_{k \neq i} \exp(-|y_i - y_k|^2)} \quad (7)$$

In order to maintain the original relative position information after mapping, we use KL (Kullback-Leibler Divergence) to measure the distance to ensure the similarity between the original high-dimensional distribution (P_i) and the mapped low-dimensional distribution (Q_i).

$$C = \sum_i \mathbb{KL}(P_i Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (8)$$

3 Proposed Model

When two representations of one dataset (such as word and topic) are manipulated together, the two representations do not have a unified metric in their respective metric spaces, which is the difficulty when we optimize NTM with OT. In Figure 1, we compute the OT distance and manifold learning loss from word embedding to topic embedding. We propose to feed the word embedding with pretrained word embedding GloVe [25]. And NSTM defines topic embedding G as one optimization parameter so that this cost matrix can be used for the target topic distribution.

In section 2.1, the loss function for NSTM is,

$$\max_{\theta, \mathbf{G}} \{ \varepsilon \tilde{x}^T \log \phi(z) - L_{M, \alpha}(\tilde{x}, z) \} \quad (9)$$

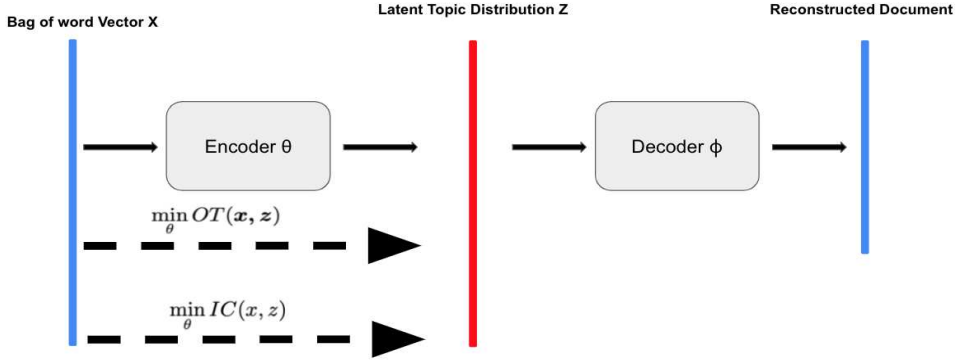
On this basis, we will add the constraint of keeping the distance, and get the loss function like

$$\max_{\theta, \mathbf{G}} (\varepsilon \tilde{x}^T \log \phi(z) - L_{M, \alpha}(\tilde{x}, z) + IC(x, z)) \quad (10)$$

$IC(x, z)$ is the constraint based on the manifold learning algorithm we selected, we name this kind constraints as isometric constraints.

Referring to the loss function of tSNE in the low-dimensional mapping space, and the isometric constrained loss that needs to be added is in the form of

$$IC(x, z) = C = \sum_i \mathbb{KL}(P_i Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (11)$$

**Fig. 1:** Model Structure

Mapping from Bag of Word Vector X to Latent Topic Distribution Z preserves neighborhood of each document x in Bag-of-word vector space

where

$$p_{j|i} = \frac{\exp(-d(x_i, x_j)/2\sigma_i^2)}{\sum_{k \neq i} \exp(-d(x_i, x_k)/2\sigma_i^2)} \quad (12)$$

The distance $d(x_i, x_j)$ between documents can be applied to various metrics (i.e. TF-IDF, WMD (word movement distance) [26] and Euclidean between word counts), but has little impact on topic representation as shown in Table 2.

For different centers x_i , the variance σ of the corresponding Gaussian distribution is also different, and each point needs to be calculated. Similarly, for the points x_i and x_j in the high-dimensional space mapped to the corresponding points z_i and z_j in the low-dimensional space, the probability distribution function is as follows. For the convenience of calculation here, we set σ for all points to $\frac{1}{\sqrt{2}}$

$$q_{j|i} = \frac{\exp(z_i - z_j^2)}{\sum_{k \neq i} \exp(-z_i - z_k^2)} \quad (13)$$

Each document is represented as a distribution over V words, this is the specific meaning of x . $\tilde{x} \in \Delta^V$ is normalised x such that $\tilde{x} := x/\text{length of document}$. Furthermore, each document is associated with a distribution of K topics: $z \in \Delta^K$, each entry in it represents the proportion of a topic in the document. Like NSTM, an encoder is leveraged to generate topic z from normalised word vector $\tilde{x}$ by $z = \text{softmax}(\theta(\tilde{x}))$ and θ is a neural network with dropout layer. And the cost matrix M represents the distance between topic k and word v , which is configured as

$$M_{vk} = 1 - \cos(\mathbf{e}_v, \mathbf{g}_k) \quad (14)$$

Now we have the loss function for isometric constrained OT as

$$\max_{\theta, \mathbf{G}} \varepsilon \tilde{x}^T \log \phi(z) - L_{M, \alpha}(\tilde{x}, z) + \varepsilon_{ic} \sum_i \mathbb{KL}(P_i Q_i) \quad (15)$$

ε_{ic} balanced the coherence and the margins between clusters of topics. This is demonstrated in the visualization of the experimental results (Figure 3b), as ε_{ic} increases, the samples that was originally crowded into a group is scattered into different clusters.

For hypothesis, we believed that in the document space, using WMD would be more beneficial, but experiments show that WMD has no advantage over TF-IDF or bag-of-words euclidean. This is due to the fact that the embedding of topics is also our training target. So the inherently uncertain distance function between topics becomes part of our optimized parameters and both the OT loss and the manifold learning loss benefit from it.

Similarly, if we use LLE instead of tSNE, we should select the k neighbors of each sample by KNN and then find the local reconstruction weight matrix of sample points to find w_{ij} that minimize

$$\sum_i x_i - \sum_j w_{ij} x_j^2 \quad (16)$$

and the isometric constraint (IC) is

$$-\sum_i z_i - \sum_j w_{ij} z_j^2 \quad (17)$$

Other Distance Awareness NTMs (DA-NTM) corresponding to specific manifold learning algorithms can also be obtained by slightly adjusting the model.

4 Related Work

The state-of-the-art NTMs including **ProdLDA**(LDA with Products of Experts) [7], **DVAE** (Dirichlet VAE) [27], **ETM** (Embedding Topic Model) [28] and **WLDA** (Wasserstein LDA) [29]. Instead of using mixture model in LDA, **prodLDA** use product of experts, then train the model using AVI. **DVAE** is a neural topic model that obtains the latent topic vector z by applying a Dirichlet prior/posterior distribution. **ETM**, on the other hand, is a topic model which applied the word embeddings and is obtained through AVI learning. **WLDA** is a topic model based on the Wasserstein AutoEncoders (WAE) framework that minimises the Wasserstein distance between topic-generated data and real data.

To our knowledge, the works that connect deep encoder with manifold learning are still very limited. The idea of leveraging visualization-based geometric methods into topic models was first proposed in **SEMAFORE** [30], aiming to combine PLSV [31] and Laplacian Eigenmaps [9]. **DWL** [32] was then proposed for optimal transmission to build topic models, and was further optimized by **OTLDA** [33]. They are all based on the traditional topic models, instead of deep topic models. Compared with their baselines, their accuracy has been improved. However, existing works do not consider preserving geometric properties.

Algorithm 1 Training algorithm for DA-NTM. $\mathbf{X} \in \mathbb{N}^{V \times B}$ and $\mathbf{Z} \in \mathbb{R}_{>0}^{K \times B}$ consists of the word count vectors and topic distributions for all the documents, respectively; $\odot$ is the element-wise multiplication.

Input: Input documents, Pretrained word embeddings $\mathbf{E}$, Pretrained document to document distance $\mathbf{d}(\mathbf{x}_i, \mathbf{x}_j)$, Topic number K , ε , α

Output: θ , $\mathbf{G}$

- 1: Compute manifold learning parameters in high dimension space (e.g. matrix P :12 for tSNE, w_{ij} for LLE), Randomly initialise θ and $\mathbf{G}$
 - 2: **while** Not converged **do**
 - 3: Sample a batch of B input documents $\mathbf{X}$
 - 4: Column-wisely normalise $\mathbf{X}$ to get $\tilde{\mathbf{X}}$
 - 5: Compute $\mathbf{M}$ with $\mathbf{G}$ and $\mathbf{E}$ by $M_{vk} = 1 - \cos(\mathbf{e}_v, \mathbf{g}_k)$
 - 6: Compute $\mathbf{Z} = \text{softmax}(\theta(\tilde{\mathbf{X}}))$
 - 7: Compute the first term of loss $\varepsilon \tilde{\mathbf{x}}^T \log \phi(z)$
 - 8: $\Psi_1 = \text{ones}(K, B)/K$, $\Psi_2 = \text{ones}(V, B)/V$ # Sinkhorn iterations #
 - 9: $\mathbf{H} = e^{-\mathbf{M}/\alpha}$
 - 10: **while** Ψ_1 changes or any other relevant stopping criterion **do**
 - 11: $\Psi_2 = \tilde{\mathbf{X}} \odot 1/(\mathbf{H}\Psi_1)$
 - 12: $\Psi_1 = \mathbf{Z} \odot 1/(\mathbf{H}^T \Psi_2)$
 - 13: **end while**
 - 14: Compute the second term of loss $d_C(\tilde{x}, z)$: $d_{\mathbf{M}, \alpha} = \text{sum}(\Psi_2^T (\mathbf{H} \odot \mathbf{M}) \Psi_1)$
 - 15: Compute manifold learning parameters in low dimension space (e.g. 13 for tSNE);
 - 16: Compute the third term of loss (e.g. 11 for tSNE, 17 for LLE);
 - 17: Compute the gradients of loss
 - 18: Update θ , $\mathbf{G}$ with the gradients
 - 19: **end while**
-

5 Experiments

5.1 Experimental Settings

5.1.1 Datasets and Baseline methods

We conduct extensive experiments on five benchmark text datasets, including 20 News Groups [34], Web Snippets (WS) [35], Tag My News [36], AG News [37], DBpedia [38].

- AG News is a collection of more than 1 million news articles. AG News has the 4 largest classes (“World”, “Sports”, “Business”, “Sci/Tech”) of AG’s Corpus. The AG News contains 30,000 training and 1,900 test samples per class.
- DBpedia is a dataset of information created in the Wikipedia project. It has 14 classes and contains 399,605 training samples and 50,060 test samples.
- The 20 News groups dataset is a collection of 18,846 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It contains 15,076 training samples and 3,770 test samples.

- Tag My News Dataset is a collection of datasets of short text fragments which are used for topic-based text classifier. It contains 26,077 training samples and 6,520 test samples.
- Web Snippets comprises four abstractive snippet datasets from ClueWeb09, Clueweb12, and DMOZ descriptions. It contains 9,867 training samples and 2,470 test samples.

We compare with the state-of-the-art NTMs, including: ProdLDA [7], Dirichlet VAE (DVAE) [27], Embedding Topic Model (ETM) [28], Wasserstein LDA (WLDA) [29] and NSTM [3].

5.1.2 Evaluation metrics

Since document topic distributions are unsupervised document representations, to assess the quality of such representations, we perform a document clustering task and report the purity and normalized mutual information (NMI), compared to document tags. Using the default train/test split of five datasets, we train the model on the training documents and infer the topic distribution z on the test documents. In the tSNE visualization mentioned below, since global information about the manifold is required, we will not split the training and testing sets.

We compute purity and NMI (denoted by top-Purity and top-NMI) using the most important topics in test documents as their cluster assignments; We apply the KMeans algorithm to z of test documents and reports the purity and NMI (denoted by km-Purity and km-NMI) of KMeans clusters. For the first strategy, the number of clusters is equal to the number of topics, while for the second strategy, we vary the number of clusters in KMeans in the range $\{20, 40, 60, 80, 100\}$; We compute AMI (Adjusted Mutual Information) and ARS (Adjusted Rand Score) between most important topics in test documents as their cluster assignments. Adjusted Rand Score (ARS) is used to measure the degree of agreement between the two distributions. The value range is $[-1, 1]$. The closer the value is to 1, the better.

Mutual Information (Adjusted Mutual Information, AMI) is also used to measure the degree of agreement between the two distributions, the larger the value, the more consistent the clustering effect is with the real situation. All the models in comparison five times with different random seeds and report the mean and standard deviation (as error bars). Since document topic distributions are unsupervised document representations, to assess the quality of such representations, we perform a document clustering task and report the purity and normalized mutual information (NMI), compared to document tags. Using the default train/test split of five datasets, we train the model on the training documents and infer the topic distribution z on the test documents. In the tSNE visualization mentioned below, since global information about the manifold is required, we will not split the training and testing sets.

Topic consistency (TC) and topic diversity (TD) are metrics that perform as indicators of topic quality. TC measures semantic consistency in the most important words in a topic with a reference corpus. We apply Normalized Pointwise Mutual Information (NPMI) which is calculated by the Palmetto package on the first 10 words of each topic to report average score over the selected topics. We vary the selected topics from 10% highest NPMI topics to 100% highest NPMI topics to get average

		top-Purity			top-NMI	
	20NG	WS	TMN	20NG	WS	TMN
ProdLDA	0.417±0.004	0.293±0.023	0.405±0.157	0.321±0.004	0.066±0.016	0.091±0.101
DVAE	0.281±0.006	0.284±0.005	0.477±0.012	0.187±0.005	0.059±0.001	0.113±0.004
ETM	0.063±0.003	0.215±0.001	0.556±0.022	0.005±0.005	0.003±0.003	0.328±0.010
WLDA	0.117±0.001	0.239±0.003	0.260±0.002	0.060±0.001	0.026±0.001	0.009±0.001
NSTM	0.477±0.011	0.451±0.009	0.637±0.010	0.415±0.012	0.201±0.004	0.334±0.004
DA-NTM	0.334±0.042	0.635±0.065	0.679±0.028	0.417±0.020	0.461±0.038	0.423±0.018

Table 1: top-Purity and top-NMI for document clustering

The best scores of each dataset are highlighted in boldface.

metric	$\varepsilon_{ic} = 1$	Euclidean between word counts	$\varepsilon_{ic} = 1$	Euclidean between TF_IDF	$\varepsilon_{ic} = 1$	word mover distance
topPurity		0.334±0.042		0.283±0.035		0.347±0.045
topNmi		0.417±0.020		0.405±0.016		0.431±0.027
topicDiversity		0.875±0.007		0.899±0.012		0.873±0.003
AMI		0.410±0.021		0.386±0.013		0.391±0.029
ARS		0.232±0.023		0.226±0.032		0.263±0.014

Table 2: different metrics to measure the distance between documents and documents on the 20NG dataset

scores. TD measures how diverse the topics are. And we define diversity for a topic as the percentage of unique words in the first 25 words.

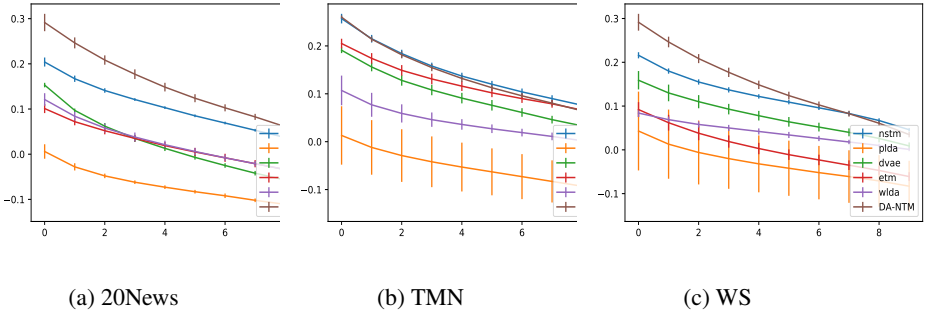
5.2 Quantitative Results

For different models in comparison five times with different random seeds, Table 1 shows the results of top-Purity/NMI with the means and stds. We have the following remarks of the results: DA-NTM outperforms the others significantly in terms of topic coherence while obtaining high topic diversity on most of the datasets. Although others may have higher top-Purity than DA-NTM in one dataset or two, they usually cannot achieve a high top-NMI at the same time. Note that our main purpose is to preserve the bag-of-words neighborhood of the text in the topic space, so it is acceptable to partially sacrifice the performance of clustering.

In Table 1 and Figure 2’s DA-NTM, $\varepsilon_{ic} = 1$ and the distance between documents and documents is Euclidean between word counts. We compared the different results of using different metrics $d(x_i, x_j)$ to measure the distance between documents and documents on the 20NG dataset in Table 2. ε_{ic} is the isometric constrained loss weight, experiments show that $\varepsilon_{ic} = 1$ is a appropriate value. When $\varepsilon_{ic} = 0$, isometric constraint does not affect, then they models are same for all document distances. It shows that different measures have no significant effect on the results and after adding isometric constraint, all other indexes except top Purity have been improved.

In Table 2, as mentioned in section 3, one can see that taking different metrics here has little effect on the performances of topic representations.

In Appendix, there are metrics in Table 2 for other datasets compared with NSTM based on $\varepsilon_{ic} = 1$ and Euclidean between word counts as document to document distance, and figures for NMI and purity for different level KMeans clusters. These tables and figures compared the performance of the NSTM and DA-NTM (with Euclidean distance between word counts and $\varepsilon_{ic} = 1$) for different datasets, they

Fig. 2: Topic Consistency (TC)

The horizontal axis indicates the proportion of selected topics according to their NPMIs

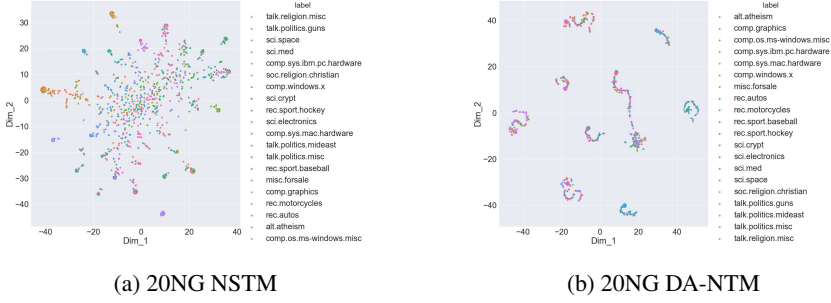
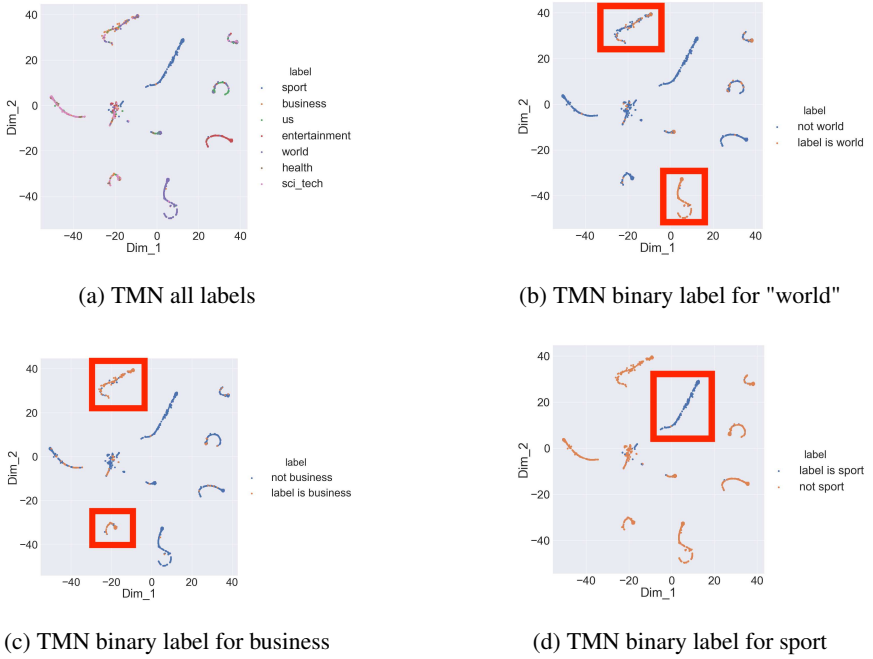
verified that adding isometric constraints has improved the performances of the model in metrics for topic quality. This matches our expectation that this model helps for scenarios with insufficient samples.

5.3 Visualizations

In order to visually verify that our model has better performance in terms of coherence, we demonstrate it by visualizing the data distribution. We visualize the output of the network for the entire data set, that is, the collection of topic distributions Z of all documents. We reduce the obtained Z to 2 dimensions by tSNE, then take the first 1,000 points (as this dataset is large, the points of taking the whole dataset is too dense), and draw a scatter plot in a 2D plane. Taking the 20NG dataset as an example in Figure 3b, we can see that when we do not add the isometric constraint, the topic distribution of dimension reduction does not show any clustering characteristics, but when the isometric constrained loss weight is not 0, the sample appears obvious clustering feature. The same color in the figure represents the same real label.

The TMN dataset is a small dataset, then the 2D tSNE dimension reduction of the entire dataset can be plotted more clearly as Figure 4a. It can be seen that different labels are obviously assigned to different clusters. In the red boxes of Figures 4b, 4c and 4d, we clearly see that a single label can be divided into some specific clusters. The label “business” and label “world” that may share more common vocabularies also share the same cluster in Figure 4(b) and 4(c). This means that there is indeed a more interpretable connection between document distribution and topic distribution under this framework, and adding isometric constraint can make up for the lack of coherence in past NTMs.

More 2D-tSNE visualized results are provided in the Appendix.

**Fig. 3:** 20NG Topic representation dimension reduction to 2D by tSNE**Fig. 4:** TMN Topic representation dimension reduction to 2D by tSNE

6 Conclusion

We propose a novel neural topic model optimized by manifold learning in this paper. Specifically, the mechanism of tSNE is integrated to constrain the topic distribution z and the topic space maintains the geometric properties of the word space. Our model can complement the shortcomings of traditional neural topic models. Extensive experiments have been conducted to demonstrate the unique advantages of the topics derived by our model, which achieves state-of-the-art performance on common metrics of document representation.

Appendix A Tables compare NSTM and DA-NTM: topPurity, topNmi, topDiversity, AMI, ARS

20NG dataset

metric	NSTM	DA-NTM	Euclidean between word counts
topPurity	<u>0.450</u> ±0.011		<u>0.334</u> ±0.042
topNmi	<u>0.403</u> ±0.004		<u>0.417</u> ±0.020
topicDiversity	<u>0.556</u> ±0.012		<u>0.875</u> ±0.007
AMI	<u>0.313</u> ±0.003		<u>0.410</u> ±0.021
ARS	<u>0.144</u> ±0.008		<u>0.232</u> ±0.023

TMN dataset

metric	NSTM	DA-NTM	Euclidean between word counts
topPurity	<u>0.600</u> ±0.011		<u>0.679</u> ±0.028
topNmi	<u>0.348</u> ±0.006		<u>0.423</u> ±0.018
topicDiversity	<u>0.921</u> ±0.006		<u>0.923</u> ±0.005
AMI	<u>0.282</u> ±0.007		<u>0.421</u> ±0.018
ARS	<u>0.218</u> ±0.018		<u>0.422</u> ±0.010

Webs dataset

metric	NSTM	DA-NTM	Euclidean between word counts
topPurity	<u>0.619</u> ±0.004		<u>0.635</u> ±0.065
topNmi	<u>0.364</u> ±0.005		<u>0.461</u> ±0.038
topicDiversity	<u>0.921</u> ±0.005		<u>0.899</u> ±0.003
AMI	<u>0.279</u> ±0.006		<u>0.456</u> ±0.039
ARS	<u>0.149</u> ±0.009		<u>0.370</u> ±0.063

AGNews dataset

metric	NSTM	DA-NTM	Euclidean between word counts
topPurity	<u>0.692</u> ±0.041		<u>0.771</u> ±0.022
topNmi	<u>0.323</u> ±0.019		<u>0.431</u> ±0.021
topicDiversity	<u>0.884</u> ±0.004		<u>0.916</u> ±0.006
AMI	<u>0.253</u> ±0.016		<u>0.430</u> ±0.021
ARS	<u>0.221</u> ±0.047		<u>0.455</u> ±0.028

Dbpedia dataset

metric	NSTM	DA-NTM	Euclidean between word counts
topPurity	<u>0.508</u> ±0.034		<u>0.447</u> ±0.060
topNmi	<u>0.457</u> ±0.014		<u>0.556</u> ±0.039
topicDiversity	<u>0.857</u> ±0.002		<u>0.878</u> ±0.006
AMI	<u>0.434</u> ±0.019		<u>0.555</u> ±0.039
ARS	<u>0.222</u> ±0.022		<u>0.369</u> ±0.046

Appendix B km-Purity and km-NMI graphs

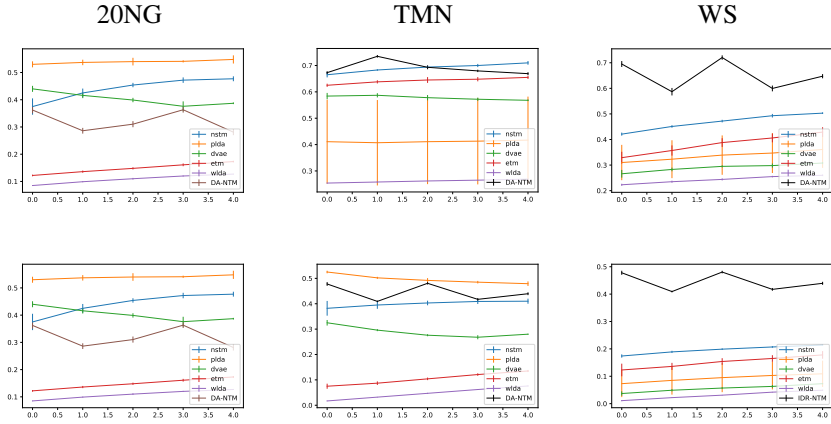


Fig. B1: The first row shows the km-Purity scores and the second row shows the corresponding km-NMI scores. In each subfigure, the horizontal axis indicates the number of KMeans clusters.

Appendix C 2D-tSNE dimension reduction for DA-NTM topic distributions:20News binary label

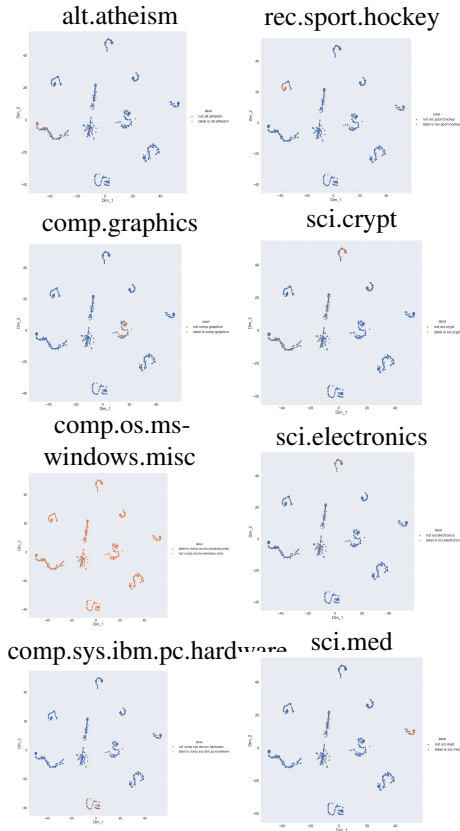


Fig. C2: 8 labels from 20News dataset binary label 2DtSNE plot with DA-NTM

Appendix D 2D-tSNE dimension reduction for all labels of 5 datasets:compare NSTM and DA-NTM

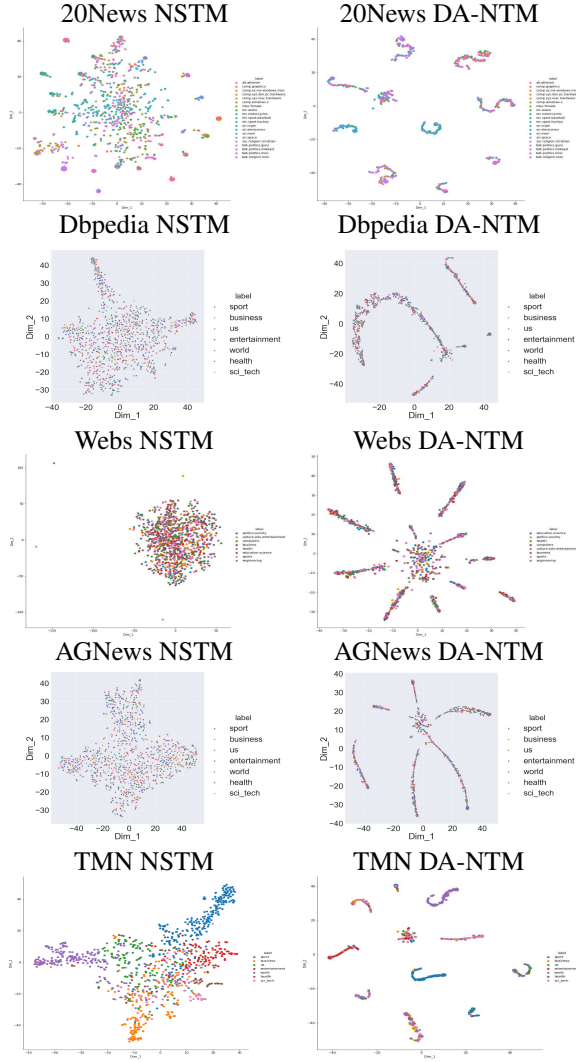


Fig. D3: 2DtSNE plot NSTM v.s. DA-NTM, all labels

References

- [1] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of machine Learning research* **3**(Jan), 993–1022 (2003)
- [2] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
- [3] Zhao, H., Phung, D., Huynh, V., Le, T., Buntine, W.: Neural topic model via optimal transport. *arXiv preprint arXiv:2008.13537* (2020)
- [4] Ma, L., Crawford, M.M., Tian, J.: Local manifold learning-based k -nearest-neighbor for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **48**(11), 4099–4109 (2010)
- [5] Lunga, D., Prasad, S., Crawford, M.M., Ersoy, O.: Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *IEEE Signal Processing Magazine* **31**(1), 55–66 (2013)
- [6] Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: *International Conference on Machine Learning*, pp. 1727–1736 (2016). PMLR
- [7] Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488* (2017)
- [8] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* **26** (2013)
- [9] Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15**(6), 1373–1396 (2003)
- [10] Canas, G., Poggio, T., Rosasco, L.: Learning manifolds with k-means and k-flats. *Advances in neural information processing systems* **25** (2012)
- [11] Carlsson, G.: Topology and data. *Bulletin of the American Mathematical Society* **46**(2), 255–308 (2009)
- [12] Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L.: Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics* **40**(2), 941–963 (2012)
- [13] Dasgupta, S., Freund, Y.: Random projection trees and low dimensional manifolds. In: *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pp. 537–546 (2008)
- [14] Donoho, D.L., Grimes, C.: Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences* **100**(10), 5591–5596 (2003)

- [15] Hastie, T., Stuetzle, W.: Principal curves. *Journal of the American Statistical Association* **84**(406), 502–516 (1989)
- [16] Kambhatla, N., Leen, T.: Fast non-linear dimension reduction. *Advances in neural information processing systems* **6** (1993)
- [17] Kégl, B., Krzyzak, A., Linder, T., Zeger, K.: Learning and design of principal curves. *IEEE transactions on pattern analysis and machine intelligence* **22**(3), 281–297 (2000)
- [18] Narayanan, H., Niyogi, P.: On the sample complexity of learning smooth cuts on a manifold. In: *COLT* (2009)
- [19] Niyogi, P., Smale, S., Weinberger, S.: Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry* **39**(1), 419–441 (2008)
- [20] Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *science* **290**(5500), 2323–2326 (2000)
- [21] Smola, A.J., Williamson, R.C., Mika, S., Schölkopf, B.: Regularized principal manifolds. In: *European Conference on Computational Learning Theory*, pp. 214–229 (1999). Springer
- [22] Tenenbaum, J.B., Silva, V.d., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *science* **290**(5500), 2319–2323 (2000)
- [23] Weinberger, K.Q., Saul, L.K.: Unsupervised learning of image manifolds by semidefinite programming. *International journal of computer vision* **70**(1), 77–90 (2006)
- [24] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* **9**(11) (2008)
- [25] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014)
- [26] Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: *International Conference on Machine Learning*, pp. 957–966 (2015). PMLR
- [27] Burkhardt, S., Kramer, S.: Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.* **20**(131), 1–27 (2019)
- [28] Dieng, A.B., Ruiz, F.J., Blei, D.M.: Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* **8**, 439–453 (2020)

- [29] Nan, F., Ding, R., Nallapati, R., Xiang, B.: Topic modeling with wasserstein autoencoders. arXiv preprint arXiv:1907.12374 (2019)
- [30] Le, T., Lauw, H.: Manifold learning for jointly modeling topic and visualization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 28 (2014)
- [31] Iwata, T., Yamada, T., Ueda, N.: Probabilistic latent semantic visualization: topic model for visualizing documents. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 363–371 (2008)
- [32] Xu, H., Wang, W., Liu, W., Carin, L.: Distilled wasserstein learning for word embedding and topic modeling. *Advances in Neural Information Processing Systems* **31** (2018)
- [33] Huynh, V., Zhao, H., Phung, D.: Otlida: A geometry-aware optimal transport approach for topic modeling. *Advances in Neural Information Processing Systems* **33**, 18573–18582 (2020)
- [34] 20 Newsgroups. <http://qwone.com/~jason/20Newsgroups/>. April 7, 2022
- [35] Phan, X.-H., Nguyen, L.-M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web, pp. 91–100 (2008)
- [36] A³ Lab - Advanced Algorithms and Applications Laboratory. <http://acube.di.unipi.it/tmn-dataset/>. April 7, 2022
- [37] <https://paperswithcode.com/dataset/ag-news>. <https://paperswithcode.com/dataset/dbpedia>. 2022
- [38] Papers with Code - DBpedia Dataset. <https://paperswithcode.com/dataset/dbpedia>. 2022